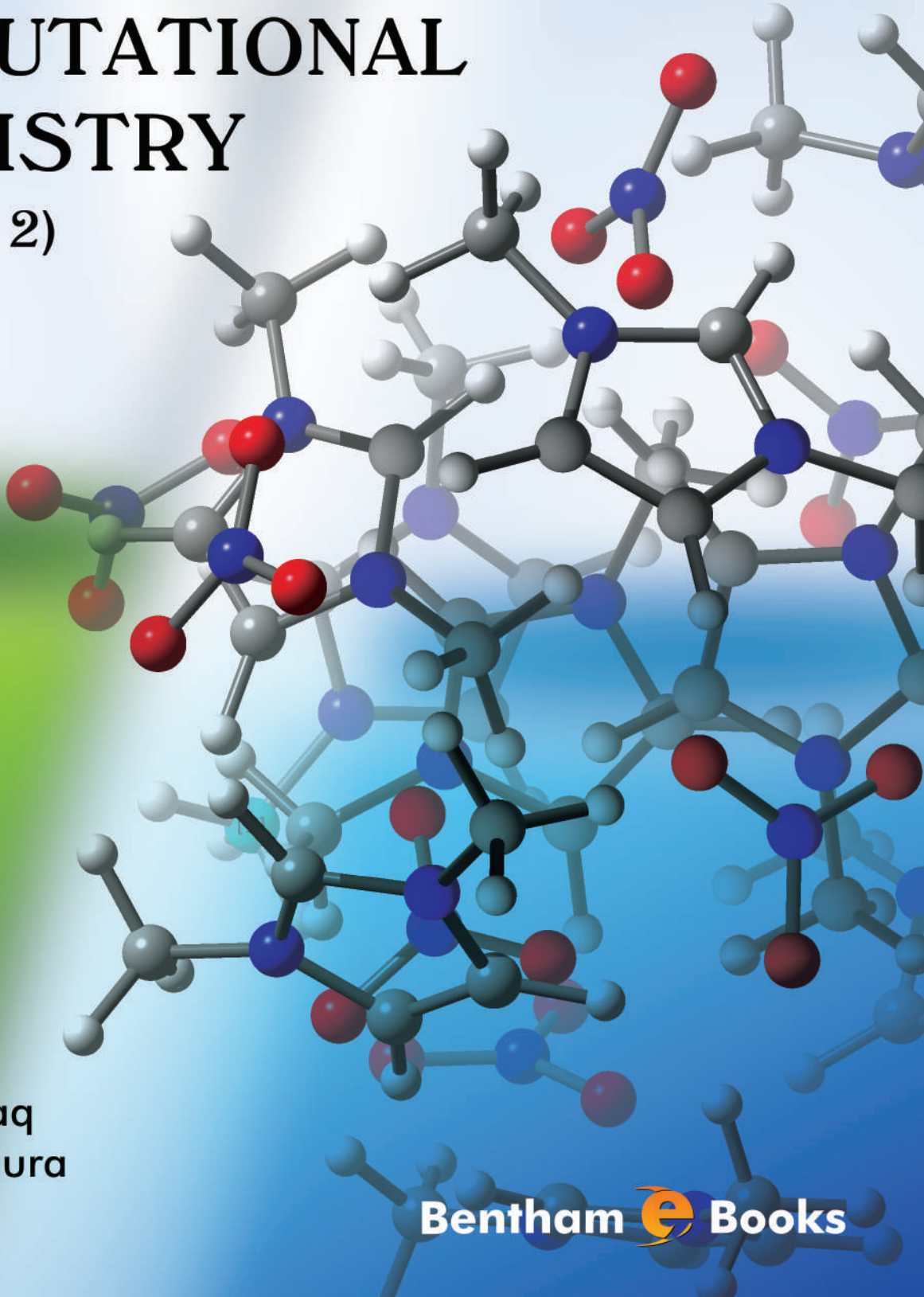
eISBN: 978-1-60805-978-2
ISBN: 978-1-60805-979-9

eISSN: 2352-9458
ISSN: 2352-944X

FRONTIERS IN COMPUTATIONAL CHEMISTRY

(VOLUME 2)



Editors:
Zaheer Ul-Haq
Jeffry D. Madura

Bentham e Books

Frontiers in Computational Chemistry
(Volume 2)

Editors

Zaheer Ul-Haq

Panjwani Center for Molecular Medicine & Drug Research
International Center for Chemical & Biological Sciences
University of Karachi
Pakistan

&

Jeffry D. Madura

Department of Chemistry & Biochemistry
Center for Computational Sciences
Duquesne University, Pittsburgh
USA

Bentham Science Publishers Ltd.
Executive Suite Y - 2
PO Box 7917, Saif Zone
Sharjah, U.A.E.
subscriptions@benthamscience.org

to use the Work. The entire liability of Bentham Science Publishers shall be limited to the amount actually paid by you for the Work.

General:

1. Any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims) will be governed by and construed in accordance with the laws of the U.A.E. as applied in the Emirate of Dubai. Each party agrees that the courts of the Emirate of Dubai shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this License Agreement or the Work (including non-contractual disputes or claims).

2. Your rights under this License Agreement will automatically terminate without notice and without the need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Ltd.
Executive Suite Y - 2
PO Box 7917, Saif Zone
Sharjah, U.A.E.
subscriptions@benthamscience.org

CONTENTS

PREFACE

Computational chemistry is a very diverse field spanning from the development and application of linear free energy relationships (*e.g.* QSAR, QSPR), to electronic structure calculations, molecular dynamics simulations, and to solving coupled differential equations (*e.g.* drug metabolism). The focus of ***Frontiers in Computational Chemistry*** is to present material on molecular modeling techniques used in drug discovery and the drug development process. Topics falling under this umbrella include computer aided molecular design, drug discovery and development, lead generation, lead optimization, database management, computer and molecular graphics, and the development of new computational methods or efficient algorithms for the simulation of chemical phenomena including analyses of biological activity. In this volume, we have collected nine different perspectives in the application of computational methods towards drug design.

Chapter 1 "The Use of Dedicated Processors to Accelerate the Identification of Novel Antibacterial Peptides" reviews the use of modern hardware advances to accelerate the identification of new antibacterial peptides. Identification of new antibiotics is of paramount importance as bacterial develop resistances to the current compounds used. The authors highlight the advantages as well as the difficulties in developing algorithms for Field Programmable Gate Arrays and Graphic Processing Units.

DNA damage by singlet oxygen is a well-known method to mitigate the presence of singlet oxygen that remains elusive. In Chapter 2 "Computational Chemistry for Photosensitizer Design and Investigation of DNA Damage" the authors review electronic structure methods to aid understanding how singlet oxygen damages DNA as well as using what they have learned to aid in the design of novel photosensitizers. They review the development of several porphyrin photosensitziers based on molecular orbital calculations.

One challenge in the QSAR field is how to judge the predictive quality of the models. The authors of Chapter 3 "How to Judge Predictive Quality of Classification and Regression Based QSAR Models?" present a review of validating QSAR models using both traditional and new validation metrics.

In Chapter 4 "Density Functional Studies of Bis-alkylating Nitrogen Mustards", the authors present a review of the application of DFT and DFRT methods on understanding the action of nitrogen mustards. Nitrogen mustards are extensively used as a chemotherapeutic agent. Identification of new nitrogen mustards is important in order to reduce their cytotoxicity and increase their effectiveness.

The authors of Chapter 5 "From Conventional Prodrugs to Prodrugs Designed By Molecular Orbital Methods" review a novel approach in the design of novel prodrugs using molecular mechanics and molecular orbital methods. In this approach, the authors review methods in which the prodrug is converted into the active drug without the enzyme.

Chapter 6 "Structural and Vibrational Investigation on a Benzoxazin Derivative with Potential Antibacterial Activity" highlights the use of DFT methods along with experimental data to

understand the properties and behavior of benzoxazin derivative. The authors present the use of scaled quantum mechanical force field methodology and Atomis in Molecules theory to explain the vibrational and bonding characteristics in benzoxazin.

In Chapter 7 "First Principles Computational Biochemistry with deMon2k" the authors present a first principles approach to investigating biochemical principles using density functional methods with the program deMon2k. Having an all-electron method to explore biochemical and pharmacological processes; is an important tool in the computational chemist's toolbox.

In Chapter 8 "Recent Advances in Computational Simulations of Lipid Bilayer based molecular systems" the authors review computational simulations of lipid bilayers. Cell membranes are a complex mixture of lipids and play a vital role in cellular function such as the control of processes that cross the cell membranes. A review of several computational methods and complex lipid mixtures is presented.

In this last chapter, "Data Quality Assurance and Statistical Analysis of High Throughput Screenings for Drug Discovery", the authors review high throughput screening (HTS) methods bringing to light the challenges to identifying novel molecules from vast and diverse databases. The authors also note the use of data from sophisticated biological assays in HTS.

Zaheer Ul-Haq
Panjwani Center for Molecular Medicine & Drug Research
International Center for Chemical & Biological Sciences
University of Karachi
Pakistan

&

Jeffry D. Madura
Department of Chemistry & Biochemistry
Center for Computational Sciences
Duquesne University, Pittsburgh
USA

List of Contributors

A. Alvarez-Ibarra Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

A. Goursot Institut Charles Gerhardt Montpellier, UMR5253 CNRS, ENSCM/UM2/UM1, 8 rue de l'Ecole Normale, 34296 Montpellier Cedex 5, France

A. M. Köster Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

A. Vela Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

Ana B. Raschi General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán. Tucumán, Argentina

B. Zuniga-Gutierrez Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

Bhabesh Chandra Deka Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India

C. Z. Gómez-Castro Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

Carlos Polanco González Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México. Cd. Universitaria, 04510 México, D.F. México and Subdirección de Epidemiología Hospitalaria y Control de Calidad de la Atención Médica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Vasco de Quiroga 15, Col. Sección XVI, 14000 D.F. México

D. R. Salahub Department of Chemistry, Centre for Molecular Simulation and Institute for Quantum Science and Technology, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

E. Laudadio Dipartimento di Scienze della Vita e dell'Ambiente (DISVA) - Università Politecnica delle Marche- *via* Brecce bianche, 60128 Ancona, Italy

Elida Romano General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán. Tucumán, Argentina

Gabriel del Rio Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Puebla, México

J. M. Vásquez-Pérez Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

Jianwei Che Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA

Kazutaka Hirakawa Department of Applied Chemistry and Biochemical Engineering, Graduate School of Engineering, Shizuoka University, Japan

Kunal Roy Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India and Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, Great Britain

L. Massaccesi Dipartimento di Scienze della Vita e dell'Ambiente (DISVA) - Università Politecnica delle Marche- *via* Brecce bianche, 60128 Ancona, Italy

María V. Castillo General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán. Tucumán, Argentina

Miguel Arias-Estrada Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Puebla, México

Nabajit Sarmah Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India

P. Calaminici Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México

Pradip Kr. Bhattacharyya Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India

R. Galeazzi Dipartimento di Scienze della Vita e dell'Ambiente (DISVA) - Università Politecnica delle Marche- *via* Brecce bianche, 60128

	Ancona, Italy
R. Grande-Aztatzi	Departamento de Física Aplicada, Centro de Investigación y de Estudios Avanzados, Unidad Mérida, Km. 6 Antigua Carretera a Progreso, A.P. 73, Cordemex, Mérida 97310, Yucatán, México
Rafik Karaman	Bioorganic Chemistry Department, Faculty of Pharmacy Al-Quds University, P.O. Box 20002, Jerusalem, Palestine
Sil*via* A. Brandán	General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán, Tucumán, Argentina
Sourab Sinha	Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India
Supratik Kar	Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
T. Mineva	Institut Charles Gerhardt Montpellier, UMR5253 CNRS, ENSCM/UM2/UM1, 8 rue de l'Ecole Normale, 34296 Montpellier Cedex 5, France
Yang Zhong	Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA
Zuojun Guo	Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA

CHAPTER 1

The Use of Dedicated Processors to Accelerate the Identification of Novel Antibacterial Peptides

Gabriel del Rio[1,*], Miguel Arias-Estrada[1] and Carlos Polanco González[2,3]

[1]Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Puebla, México; [2]Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México. Cd. Universitaria, 04510 México, D.F. México and [3]Subdirección de Epidemiología Hospitalaria y Control de Calidad de la Atención Médica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Vasco de Quiroga 15, Col. Sección XVI 14000 D.F. México

Abstract: In the past decades, the procedure to identify novel antibiotic compounds has been motivated by the heuristic discovery of the antibiotic penicillin by Fleming in 1929. Since then, researches have been isolating compounds from very wide range of living forms with the hope of repeating Fleming's story. Yet, the rate of discovery of new pharmaceutical compounds has reached a plateau in the last decade and this has promoted the use of alternative approaches to identify antibiotic compounds. One of these approaches uses the accumulated information on pharmaceutical compounds to predict new ones using high-performance computers. Such approach brings up the possibility to screen for millions of compounds in computer simulations. The better predictors though use sophisticated algorithms that take up significant amount of computer time, reducing the number of compounds to analyze and the likelihood to identify potential antibiotic compounds. At the same time, the appearance of computer processors that may be tailored to perform specific tasks by the end of the past century provided a tool to accelerate high-performance computations. The current review focuses on the use of these dedicated processor devices, particularly Field Programmable Gate Arrays and Graphic Processing Units, to identify new antibacterial peptides. For that end, we review some of the common computational methods used to identify antibacterial peptides and highlight the difficulties and advantages these algorithms present to be coded into FPGA/GPU computational devices. We discuss the potential of reaching supercomputing performance on FPGA/GPU, and the approaches for parallelism on these platforms.

Keywords: Antibacterial peptides, FPGA, GPU, high-performance computations, parallelism, QSAR, supercomputing.

***Corresponding author Gabriel Del Rio**: Department of Biochemistry and Structural Biology, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, México DF, México; E-mail: gdelrio@ifc.unam.mx

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

INTRODUCTION

The discovery of salvarsan (arsfenamine) in 1901 by Paul Erlich facilitated the treatment of syphilis; Erlich proposed the idea of "*magic bullets*" to explain the way this synthetic compound was able to kill the bacteria associated to the disease: a magic bullet (salvarsan) traversing along the body in search of its target (bacteria) without damaging any other tissue [1]. This discovery promoted the idea of synthesizing target-specific compounds as a way to find novel antibiotics. In 1929 penicillin was accidentally discovered by Flemming [2] and provided an example of an effective antibiotic whose target was unknown at the time of discovery; this in turn promoted the development of phenotype screening methods aimed to identify antibiotics by their function and not by specific target. These two strategies are still in use; for instance, in the period from 1999 to 2008, 28 out of 45 first-in-class new molecules tested by the FDA were discovered by phenotype screening methods [3]. Each of these strategies has advantages and disadvantages that have been recently reviewed [4, 5] and are out of the scope of this review. For any of these strategies, knowledge about the molecular mechanism of action of antibiotics is a desirable feature for any drug to be used and that implies knowing the target of action; however, many antibiotics approved by the FDA are poly-pharmacologic (*i.e.*, act on multiple targets) [6] and such feature troublesome the synthesis of new pharmacologic compounds based on specific targets [7] because that implies the synthesis of large molecules that tend to be non-permeable to cells and not easy to synthesize.

The recent recognition that polypharmacological compounds are among the most effective antibiotics may explain the relatively small number of new pharmaceutical compounds approved by the FDA despite the increasing amount of resources invested [8]. This has led to a shift in the last decade in the research and development strategy of the pharmaceutical industry: focus on strategic therapeutic areas and outsourcing with Universities. Among the areas gaining interest in pharmaceutical industries are new treatments for bacterial infections. This trend constitutes an impulse to explore new strategies to identify new antibacterial compounds, especially those obtained from biological means, also referred to as *biologics*. We review the results of a new promising strategy aimed to identify new antibacterial compounds combining the knowledge gained from the traditional target-based or phenotype-based strategies with computer sciences and technology.

ANTIBACTERIAL PEPTIDES

Antibacterial peptides (APs) are produced by many different organisms and have the ability to arrest cell growth (bacteriostatic) or to kill bacteria (bactericidal) [9]. Similar to other antibacterial compounds [10], APs act upon different targets in bacterial cells [11]: the cell wall, DNA replication machinery and the ribosome; furthermore, these peptides not only act on the bacterial cell, but also are able to elicit an immune response from the host as part of the innate immune response [12]. Furthermore, some APs also have antiviral, antiparasitic and antifungal activity that had led to use some APs (*e.g.*, gramicidin S and polymyxin B) to treat infections [13]; yet some features of natural APs must first be improved before they can be used as therapeutics, including: the high cost of large-scale production, stability to proteases, unspecific toxicity against eukaryotic cells, and potential development of immunological reactions [14]. Among these, the cost of production and toxicity against eukaryotic cells seem feasible to be improved simultaneously by producing linear peptides in biological systems [15] and alter the physical properties of AP to achieve selectivity towards bacteria [16]. In the current review, we analyze the strategies that have been used to identify selective antibacterial peptides mainly focus on computational approaches.

The most abundant APs are cationic antibacterial peptides or CAPs, which are relatively short (12-100 residues) and amphiphilic; despite their similar physical properties, CAPs share very little sequence similarity and fold into four main classes: amphiphilic peptides with two to four β-strands, amphipathic α-helices, loop structures, and extended structures [12]. The amphipathic α-helical peptides are linear and are suitable for large-scale production since these do not require any disulfide bond to adopt a functional structure [17].

The mechanism of action for any CAP is accepted to require the interaction of the peptide with the bacterial membrane, but the basis of their action differs according to their final target of action [18, 19] (see Fig. **1**): CAPs would approach be stabilized around membrane due to their rich composition of Arginine residues, which have the ability to interact with both lipid and water; once in contact with the membrane, CAPs may disrupt it or pass through it to find its target [20].

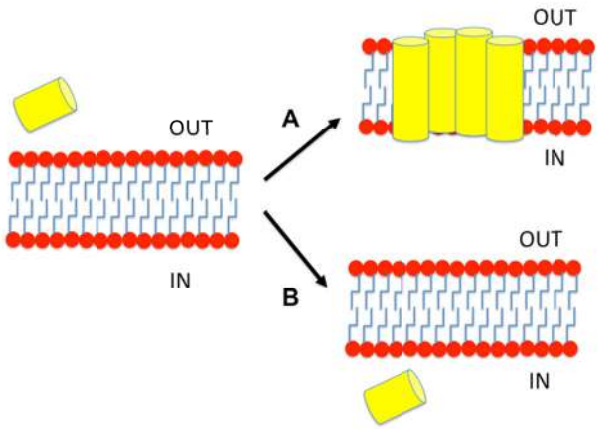


Figure 1: Mechanisms of action of CAPs. CAPs may interact with membrane and either destabilize the lipid membrane (A) or insert into the membrane in a non-disruptive manner (B). A CAP is represented in the image as a yellow cylinder and the lipid membrane as red circles with blue tails.

In any case, a desirable feature of CAPs is the ability to discriminate bacterial from mammalian cells; such peptides are referred here to as Selective CAPs, or SCAPs. Different computational approaches have been reported that aimed to identify novel antibacterial peptides [21-25].

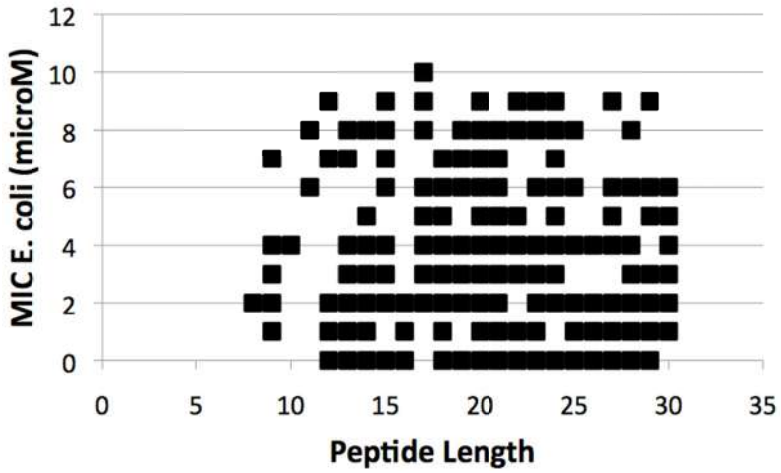


Figure 2: Size-Activiy relationship of APs. The reported activity against *Escherichia coli* by every antibacterial peptide reported in the YADAMP with 30 or less amino acids in length, with Minimum Inhibitory Concentration below or equal to 10 μM, is plotted against the length of the corresponding peptide. Data obtained from the October 5th, 2012 version of the YADAMP [26].

Considering the extremely large number of possible peptide sequences, it is expected that systematic computational screenings may be performed on small peptides (less than 30 amino acid residues). Noticeable, the most active antibacterial peptides can be found in peptides of very different lengths (see Fig. **2**; the figure shows only peptides with MIC against *Eschericia coli* < = 10 μM). Interestingly, the majority of these peptides have been isolated from the animal kingdom, which should display preferential antibiotic activity against bacteria and few if any toxicity against animal cells, that is, these may be part of the SCAPs.

In terms of multiple activities reported for CAPs, it has been noted that these share similarities with Cell Penetrating Peptides (CPPs) [27]: both families of peptides are cationic, amphipathic and their action depends on the interaction with lipid membranes. An important difference is the target membrane and the mechanism of interaction of the peptide with the membrane: some CAPs disrupt bacterial membrane's integrity while CPPs pass through mammalian membranes without affecting its functionality. Yet, it has been reported that CPPs may have antibacterial activity [28, 29] and some CAPs may pass through mammalian membranes like CPPs [30]. While it is not clear what is the biological significance of this functional redundancy, it has been proposed that the difference in the activity displayed by these two families of peptides depends on the ability to adopt a stable three-dimensional structure in solution and/or lipid membranes [20].

Furthermore, natural CAPs not only have the ability to kill bacteria but also to elicit immune response in animals [12]. So, by matching properties from peptide sequence or structure to its activity it is possible to find some computable features that may be related to the diverse activities observed in CAPs: lipid membrane disrupting activity and/or lipid membrane penetrating activity and/or recognition of intracellular bacterial targets and/or eliciting immune response. In this sense, it is possible to recognize two different types of patterns from peptide sequences: those that are conserved at the sequence level and others that are conserved at the three-dimensional structure. For instance, binding sites may be conserved at the sequence level, while peptide partitioning into lipid membranes would be a conserved physicochemical coded in the three-dimensional structure of peptides. Thus the use of computational techniques may not only assist in the identification of novel SCAPs but also in understanding the structure-function relationship of this family of peptides.

Here we review some of the computational methods used for the prediction of CAPs and their advantages and disadvantages to be coded into dedicated processors.

ALGORITHMS USED TO PREDICT ANTIBACTERIAL PEPTIDES

Two different paradigms are used in the identification of new CAPs: target-based and activity-based approaches. The first one requires knowledge about the target molecule while the second one is oriented towards finding patterns associated to the activity of the known CAPs. Alternatively, from a mathematical and computational perspective, methods that relate chemical structure with biological activity can be classified in two main groups: supervised learning and non-supervised learning [31].

Supervised Learning

This method deduces a stochastic function from a representative sample of characteristic elements from the biological pattern searched. The goal of the supervised learning is to create, from a set of examples or training data, a function capable to predict the value corresponding to any element in study, then from this data it has to generalize unseen information. Some representative algorithms in this group are: Quantitative Structure Activity Relationships, Hidden Markov model, Montecarlo method, Support Vector Machines and Fourier Transforms. Particularly, the HPC programming for these methods is the most recommended because in most of the occasions the diagram of tasks to be performed are independent or semi-independent, which allows you to dramatically reduce processing times, to the extent that will cooperatively processors. Within the supervised methods, the methods most used in the determination of biological profiles are the Quantitative Structure Activity Relationships and Hidden Markov models, which are then described in detail, due to their importance.

Quantitative Structure Activity Relationships (QSAR)

This model [32] quantifies the physicochemical properties in the element studied, characterizing its biological process. A mathematical function built this way can be used to predict the response of other chemical structures. This method gives

each chemical group parameters, such that, when modifying its chemical structure the contribution of each functional group to the pharmaceutical drug or toxic substance activity can be evaluated, determining the activity variation of a particular substance.

QSAR modeling generates predictive models correlating biological activity with physicochemical properties through statistical tools [33, 34] in QSPR models of chemicals with distinctive properties or molecular structure descriptors. QSAR are widely used in many disciplines apart from drug discovery and lead optimization [35], they are also used in toxicity prediction and regulatory decisions [36].

The success of any QSAR model depends on the accuracy of the initial data, the selection of adequate descriptors, statistical tools and most important the validation of the developed model.

Validation is the process to determine the reliability and importance of a procedure for a specific purpose [37]. To validate QSAR models we have to thoroughly consider the following aspects: the selection methods of the training set components [38]; the criteria to set the size of the training set [37] and the impact variable selection will have on the training set model to determine the prediction quality.

There are four main strategies to validate QSAR models [39]: (1) Internal validation. (2) Validation by data division in training and testing samples. (3) External validation applying the model to outside data. (4) Data randomization.

The order of complexity in a QSAR model is cO(n), it depends on c the number of variables involved. Its computational implementation frequently uses the master-slave parallelization method that reduces 90% of time compared to processing time in a mono-processor.

QSAR methods have been used to detect antibacterial peptides for quantifying contact energy between neighboring amino acids [40], quantifying based on the physicochemical properties of amino acids [41, 42], quantifying hydrophobic property parameters [43]. Applications range from QSAR models of low-

molecular weight drugs [44-51], to QSAR/QSPR models for protein and nucleic acid sequences [52, 53], protein 3D structure [44, 54], RNA secondary structures [55, 56] and of course peptides [57]. The idea has been extended to include also Quantitative Proteome-Property Relationship (QPPR) models that personalize predictions of drug cardiotoxicity [58, 59], or human prostate cancer [60] based on protein composition of blood proteomes.

Hidden Markov Model [61]

It is a statistical model where the system being modeled is presumed to be a Markov process with hidden (unobserved) states. An HMM can be considered as the simplest dynamic Bayesian network.

The order of complexity in a MM is O(n), as it acts as a dimensional contractor, therefore, it does not depend on the variables implicated but on the immediately previous state of the components evaluation. However, its computational implementation varies in efficiency. Although MC or HMM use the parallelization master-slave method to reduce 90% processing time respect to a mono-processor, HHMM computational efficiency is only between 30% and 60%.

HMM has been used for antibacterial peptide detection reproducing physicochemical properties [61], protein identification [62], data base searcher construction BLAST [63], genes sequence detection [64] or prebiotic scenario recreation [65]. The MARCH-INSIDE approach (Markov Chains Invariants for Network Simulation and Design) introduced by González-Díaz and collaborators, use the Markov Chain theory to infer QSAR/ QSPR models at different structural levels. These Markov methods use different types of transition probabilities described by atom-atom, nucleotide-nucleotide, amino acid-amino acid, or even protein-protein matrices. Two recent in-depth reviews of the field were recently published [58, 59].

Unsupervised Learning

In this method the biological model profile is built without previous knowledge of the pattern searched, thus the non-supervised learning takes the data inspected and sets it in a cluster. The group of representative algorithms from this method is called Clustering. This is a highly complex method to implement computationally, although

the description of the method is simple, its computational abstraction, demand a high level of experience in programming, that implements an HPC platform. This is mainly due to the need to update the information of all the nodes and the non-independence of tasks has carried out, a typical algorithm of this method is Clustering, which is described below.

Clustering

It is a principal task [66] of explorative data mining, and a general technique for statistical data analysis used in many fields, including machine learning, pattern recognition, and bioinformatics [67, 68].

Cluster analysis is a general task to be solved. Various algorithms that differ significantly in their concept of what forms a cluster and how to find them efficiently have been described and we will review some of them. A common concept of cluster includes groups with short distances among members, compact areas of data space and intervals or particular statistical distribution. Therefore clustering can be expressed as a multi-purpose optimization problem.

Proper clustering algorithm and parameter setting depend on the intended application of the results. Cluster analysis is not an automatic task, but a repetitive process of knowledge discovery and interactive multi-purpose optimization involving trial and error. It will often be necessary to change pre-processing and parameters until the result achieves the desired properties.

Clusters found by different algorithms vary significantly in their properties, to understand these cluster models we have to understand the differences between different algorithms: (1) Connectivity model: it builds models based on connectivity distance. (2) Centroid model: the k-means algorithm represents each cluster by a single mean vector. (3) Distribution model: clusters are modeled based on statistic distribution, such as the multivariate normal distribution used by the Expectation-maximization algorithm. (4) Density model: it defines clusters as connected compact regions in the data space.

DEDICATED PROCESSORS USED TO PREDICT ANTIBACTERIAL PEPTIDES

Predicting novel antibacterial peptides using computational techniques has many advantages but also challenges. The nature of the algorithm used and the

complexity associated with data processing can lead to long processing times often non-convenient for large peptide screenings to be performed in a reasonable amount of time; therefore options to accelerate in several orders of magnitude the computational load are desirable. From a computational architecture perspective, there are only three approaches to accelerate an algorithm: a) use a faster processor, b) optimize the algorithm reducing its complexity or finding novel ways to solve it and, c) increase the number of instructions processed per clock cycle. Option a) has already arrived to a limit where newer computers are not increasing their clock speed due to economical limitations (*i.e*. faster processors will run in frequencies close to microwave signals and would require complex design and expensive integrated circuit packages and printed circuit boards). The second option, is the choice of most of the computer science groups, trying to find out new ways to transform a data/problem domain into a less complex domain where the total amount of computer instructions is reduced, either by algorithm complexity reduction or implementation optimization, but in most cases at the expense of some kind of tradeoff, *i.e*. speed *vs*. accuracy. The third option is the parallel computing choice, that is, increasing the number of computational resources that solve simultaneously different parts of the problem or to process different parts of the whole data set simultaneously. Even with huge supercomputers available to the bioinformatics community, there are additional limitations and challenges to parallelize effectively any kind of problem in a straightforward way. We will discuss some of those challenges, co-processing platforms for acceleration, and some approaches for peptide computation acceleration in the following subsections.

Computational Complexity and Accelerating Algorithm Execution

As presented in section III, there are several approaches for peptide computational research. Not all algorithms can be straightforward parallelized, although there are general patterns and strategies to do it. In general, parallelization is guided based on how intensive is an algorithm in computations, the amount of data to be processed, and the dependencies in intermediate results (*i.e.* iterative algorithms depend on the result of the previous iteration).

In general, bioinformatics algorithms are common building blocks from algorithms already explored by computer science that have been adapted to the particularities of bioinformatics. Many techniques to parallelize the algorithms are

already in the literature and some have been adapted to bioinformatics [69]. Furthermore, the platform to execute a parallel algorithm also guides the approach to follow, programming language, fine or coarse partition, *etc*.

Some algorithms for peptide search have already been formulated in the context of a parallel implementation, for example around Markov models [70]. Since Markov models are sequential computations, the parallelization is challenging, and proposing a whole framework since the beginning is a good strategy. Other algorithms based on relaxation/iterative [71] are complicated to parallelize, so high level data partition is a possible choice. Machine Learning algorithms, like those based on neural networks [72] or evolutionary algorithms [73] have the potential for parallel implementation. Some algorithmic techniques for data post-processing, filtering, and pre or post data selection [74-76], can also be parallelized but the nature of each approach needs to be evaluated individually to find the best compromise for acceleration.

There have been efforts to identify common building blocks for bioinformatics in the past [69, 77], where several lessons can be applied to novel algorithms. Many of them deal with the sequencing and string matching problem [78, 79] and several techniques in parallelization in existing algorithms can be adapted to new platforms [80].

Parallel Computing Platforms

The main parallel computation platforms available are: supercomputers, FPGA dedicated co-processors, and the use of GPU-Graphic Processing Units, as co-processors. A brief overview is given in the following subsections with some user perspective pros and cons for peptide and in general, for bioinformatics exploration.

Supercomputing

Supercomputing is the choice for large-scale computational efforts like the human genome project. There are current large investments in academia and pharmaceutical industry for High Performance Computing (HPC) systems, based on traditional arrays of CPU nodes, or nodes with co-processing units (FPGA or

GPU, the later more popular since 3-5 years ago due to the price/performance and power consumption issues). Multi-node systems has been the mainstream in HPC since decades, with more computational power attained with new CPU nodes available in each generation, but the limitation is how to program and effectively split data/computation in the most efficient way across a multi-node system composed of thousands of nodes. In a first approach, the use of supercomputers, and in general any cluster of processors could seem straightforward since the software already running in a single node could be executed in multiple nodes. A first complication arises in the implementation/design on how to partition data among nodes and manage the data partition and results gathering. Things complicate further if the algorithm requires access the full data set or a subset that is spread among several nodes, creating bottlenecks in data transfer that can limit the actual global performance. Furthermore, some algorithms require iterate based on results generated during processing, and sharing and communicating those results to other nodes can complicate data management issues. Several languages and software platforms are available to facilitate this, but the user must understand the tradeoffs to be able to write efficient programs to do the job. Some bioinformatics programs like BLAST [81], that performs multiple sequence comparisons, are already ported to multiple nodes/supercomputing efforts where teams of several programmers can architect a parallel software, but exploring new algorithms on a supercomputer context could be complex for most users. Another effort is the proprietary platform from CMD Bioscience [82] for peptide discovery that executes on HPC resources, for commercial research.

Field Programmable Gate Arrays - FPGA

Field Programmable Gate Arrays are special integrated circuits where digital architectures can be configured to cope with user needs by defining specific computing architectures. FPGAs are a mature technology used mainly for low to medium size production of custom digital electronics for a wide range of applications. Two of the largest FPGA manufacturers are Xilinx and Altera. Fig. **3** shows the internal architecture of an FPGA. The CLB - Configurable Logic Blocks, are the actual building units for logic design. The FPGA contains hundreds to thousands CLBs inside. The CLBs are basic combinational and sequential modules that are user configurable. CLBs are connected among them

using special connectivity lines that are also programmable. The logic and connectivity reconfiguration gives the FPGA the flexibility to implement all kind of digital architectures so a dedicated processor can be built using these blocks.

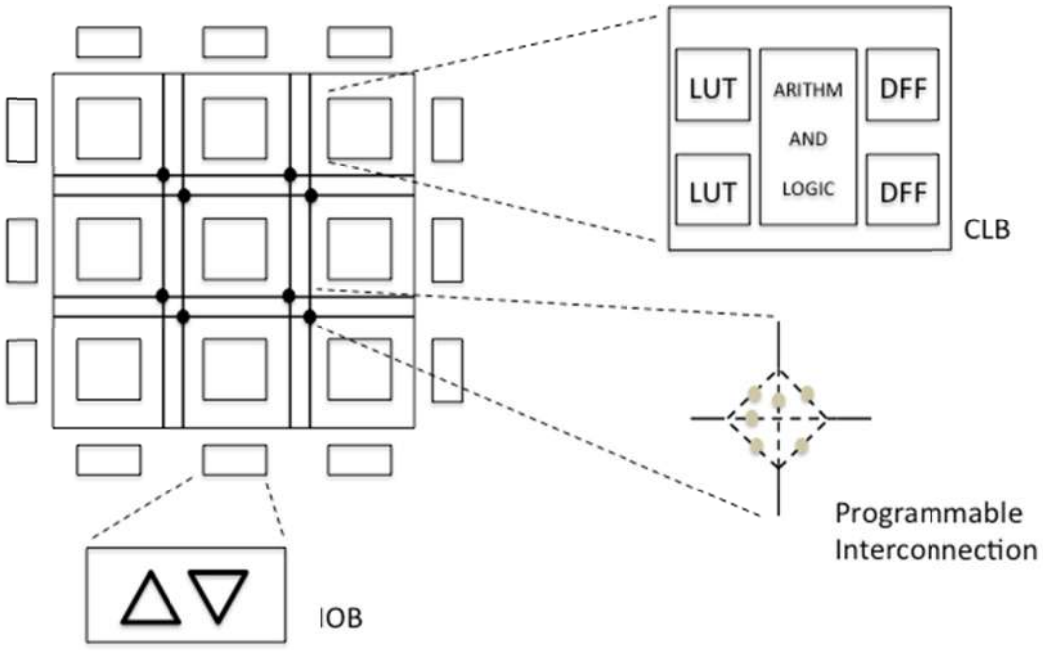


Figure 3: General structure of an FPGA. The Field Programmable Gate Array is a matrix of Configurable Logic Blocks (CLB) that can implement basic logic/arithmetic functions, interconnected by configurable buses for local and long distance interconnection. Each CLB could implement 4 to 8 bit logic. Usually the CLB integrates 2 to 4 LUTs (Look up Table), Add/Subtract/Multiply logic and D Flip-flops. The Input/output (IOB) modules are also configurable to be any direction, so full flexibility of the device is possible to implement dedicated digital architectures.

The digital architecture in an FPGA can be defined using a Hardware Description Language (HDL) and then synthesized into the FPGA building blocks. The task of "programming" an FPGA is actually closer to the design process of an Integrated Circuit than to programming, although the use of a HDL to describe the architecture borrows some techniques and methodologies from Software Engineering to speed up development and reuse components.

Modern FPGA devices can have enough resources to build custom parallel computing architectures, and can easily reach around 10 million equivalent digital

gates, enough to accelerate 10x to 100x compared to simple processor implementation, or 100x to 1000x acceleration if custom processors are designed for a particular algorithm. FPGAs is the platform of choice for Reconfigurable Computing, an area of computer science focused on proposing custom architectures for a particular algorithm, in contrast to traditional computer implementation where the algorithm is limited to the fixed architecture of a CPU. A custom architecture can accelerate performance since the inherent limitation of a general purpose CPU is eliminated with a reduced instruction unit that is optimized for the particular algorithm to be executed, reducing the hardware resources required, but limiting the flexibility once the architecture is implemented in the FPGA. The reconfigurability of the FPGA allows relatively fast "programming", for debugging and fast prototyping, providing an interesting balance between flexibility, programmability and hardware acceleration. A limitation is the number of custom processors that can fit in an FPGA, so acceleration can be gained with parallelism but limited to the actual number of units that can be implemented in an FPGA device or set of FPGA devices in a platform, as well as other architectural issues when moving large amount of data among processors and devices [83].

In order to propose a particular architecture to accelerate an algorithm, the user has to master architecture design and algorithm design/understanding, so a solution can require large amounts of efforts (months to years) to achieve a fully functional and optimal solution. Some efforts in component reuse, parametric core generation and special HDL that can be used for fast prototyping have focused on encapsulating the complexity to allow faster development time, but they have not become widespread. For example, the platform created by the company Mitrionics allows for bioinformatics development around FPGAs but the knowledge is kept inside the company or can be used on particular applications were they train people to use the platform.

Other languages, like Handel-C, were created to accelerate development by keeping a standard C syntax and exposing explicitly parallelism to the programmer, hiding the design of state machines in the hardware. A recent alternative is the use of OpenCL to synthesize FPGA based architectures that can be parallelized automatically due to the OpenCL nature. Other high level

languages and synthesis approaches, where the complexity is encapsulated into semi-automated synthesis tools to generate building blocks for bioinformatics and FPGA are reviewed in [84].

The potential of FPGAs for bioinformatics algorithm acceleration have been explored with several algorithms. For example the basic Smith-Waterman algorithm, that is the basis of many bioinformatics algorithms, had shown more than 160x acceleration [85] showing the potential of the technology.

Peptide research has been benefited with FPGA acceleration, in particular in peptide sequencing [86] and peptide spectrometry identification [82, 87] that deals with large number of comparisons. Some researches have concentrated on the parallelization issues, while others on the optimum architecture implementation on an FPGA [86], that gives an idea of the complexity of the design with FPGAs since the developer needs to master parallel algorithms and architectural design.

Other FPGA based supercomputing platforms, not limited to bioinformatics, are presented in [88]. Bioinformatics algorithm acceleration can be pursued with FPGA platforms, but as in the case of Mitrionics approach, it requires investing in time and effort to build an efficient workflow for a particular accelerated application. For instance, TimeLogic Inc. has developed a suite of bioinformatics applications, Decypher, embedded in FPGAs cards to offer mature algorithms already accelerated in proprietary modules.

Graphic Processing Units - GPU

Another alternative for implementing parallel algorithms are the Graphic Processing Units. Traditionally the demands on high performance graphics in personal computers have driven the development of the GPUs, in a parallel path to the development of the microprocessors. GPUs have evolved of a graphics coprocessor to a full computational workhorse with dedicated high speed bus to the CPU with highly parallel (from 128 to 1000s) processors to deal with the graphic computations of 3D visualization applications, in particular games. Graphic programming also have evolved with specialized languages like Open GL, that in the mid-2000s were used to program custom parallel programs but with limitations (like double precision arithmetic) since the mathematical units in the GPU were simplified only for

graphics. Other efforts migrated to facilitate programming on the GPUs with languages like OpenCL (a superset of Open GL but generic of all kind of parallel computations) and CUDA-C (a language for multithread programming suited initially for Nvidia GPUs). In particular, the 2007-2009 GPUs began to support CUDA-C and OpenCL as generic programming languages so the GPU can be used effectively as a parallel computation engine attached to the CPU. Furthermore, the last generation of GPUs from Nvidia [89] with Tesla and Kepler architectures are designed specifically as parallel engines, so a "supercomputer" can be configured in a desktop computer with 2 to 4 GPU boards reaching 1-2 TeraFlops of processing power (see Fig. **4**).

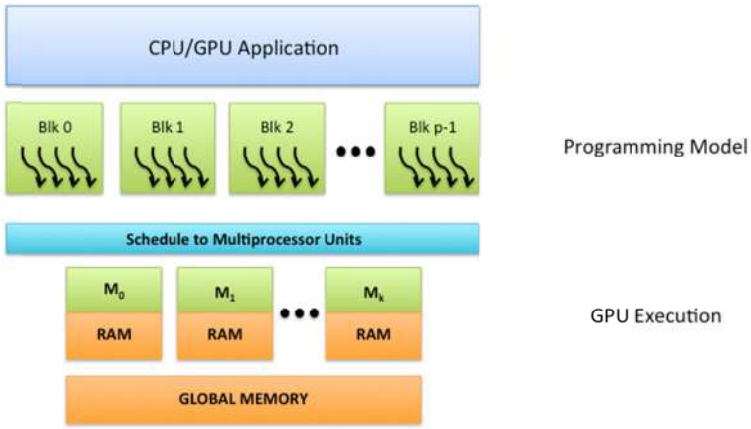


Figure 4: GPU structure abstraction. The Graphic Processing Unit can be seen as a parallel computer with a multiple thread programming model. The threads are executed in block units. The GPU executes the block instances in multiprocessor units with local RAM and access to a Global Memory to communicate data with the CPU. The programming model allows for generic parallelization of hundred of thousands of threads (organized as multiple blocks of threads), and the GPU executing engine automates the real parallel execution in the multiprocessor array available.

The current generation of GPU boards integrate 500-2000 core units, so a well parallelized algorithm can be accelerated hundreds of times, delivering orders of magnitude acceleration compared to a standard multi-core processors. On the other hand, the programming model differs from a multi-core processor or a multiple node cluster or supercomputer. Since the GPU is more like a multi-threaded architecture where multiple (thousands) of threads are executing in parallel in the internal GPU cores. The programming model requires a different

way of formulating the algorithm to take advantage of the whole potential: on one hand, the threads have to be light, *i.e*. not very long, but they have the flexibility of being able to use floating point and single or double precision arithmetic; and on the other hand, the programmer has to reformulate carefully the data partition of the data set to take in consideration the limitations in the local memory of the processing cores, and reduce the number of read/write cycles to the main memory as well as data transfers to the host processor, since those are the bottlenecks to achieve peak performance.

Depending on the nature of the algorithm to be accelerated there is a variety of techniques to deal with the optimum problem partition, parallelization and data gathering. A review of those techniques can be found in [69, 90], and applied to GPU programming in [91, 92]. Common bioinformatics algorithms have been already ported to GPU, for example [80] presents the BLAST algorithm accelerated on a GPU.

Accelerating Peptide Computation

Previous work related with Peptide Computation acceleration was carried out initially in the spectrometry area to identify peptides, where data obtained from spectrometer needs to be compared to a an existing database and where the comparison time increases quadratically with the amount of data. Some implementations on FPGA architecture for the comparison and the labeling of peptides have been described [87, 93], and [94] follows a similar approach but with a HPC.

On the other hand, large scale efforts to build a platform for peptide research and discovery is proposed by the company Mitrionics for commercial purposes, so the low level details are not given but their approach is based on high level synthesis of building blocks for bioinformatics.

Following the GPU approach, [95, 96] formulates under a parallel SIMD structure their algorithms so they can be implemented in GPU.

Most of the research on peptide discovery does not deal with the acceleration requirements and they focus on GPU or cluster/supercomputing implementation

to reach the performance required. Another approach considered since the beginning the acceleration requirements is discussed in next subsection.

Our Approach

In our research, we proposed a design to accelerate the prediction of physicochemical properties based on extended algorithms from del Rio [16], that has been reported in [97]. As a first approximation, the architecture to accelerate the algorithms was developed using Handel-C, and validated on a Xilinx FPGA device. The architecture was replicated four times in a medium size FPGA and performance was evaluated. Current FPGAs allow more instantiations of the same architecture in a single device and there are platforms to build processing systems with several FPGAs as co-processing units of a HPC. Therefore an extrapolation of the performance that can be achieved is possible.

In 2011-2012 we did additional explorations implementing the same set of peptide algorithms in a modest size GPU with good results, reported in [98]. We have observed that the computing time required for different lengths of peptide and different GPU models provides an average of 64 Million peptides evaluated in around an hour (3400 to 4300 seconds depending on GPU model) (our unpublished data). Global memory is not relevant for our implementation due to the independent nature of the algorithms that execute in their own private portion of the sequence, and is used only for data transfer between the CPU and GPU. These results can be extrapolated to the new Kepler architectures and multiple GPU on a single node, reducing in an order of magnitude the time. An extended implementation in a HPC cluster with GPUs, could reach another two order of magnitude time reduction or process more data in the same time. In other words, the timing reported could be extrapolated to an additional time reduction of 1000x, or the possibility of processing 1000x more data in the same time with current technology using an HPC cluster.

Concerning FPGA and GPU performance, we have observed that GPU reaches faster processing time and is better suited to deal with large amount of data (our unpublished data). The FPGA implementation, although effective, it would be complicated to scale to the same amount of parallelism of the GPU since all the

data communication and coordination has to be scaled up too, therefore GPU would be the technology of choice to continue exploring CAP in our group.

Discussion

GPU based computing and the new trend for GPU co-processing in HPC, is an opportunity for fast software implementation, exploration and scalability. The trend will continue at least for the following 10 years, while FPGA based computing could be relegated to other applications where robust algorithms will be embedded in specific processing boxes for low cost/high performance computing in small laboratories, or for portable applications. From our experience, GPU offers the best tradeoff to explore parallel algorithms for peptide search, in terms of flexibility, ease of programming, performance and cost. Even if additional expertise is required to develop an efficient parallel algorithm in GPU, the skills required to develop such expertise are shorter compared to FPGA programming. Also, platforms like the Amazon AWS (aws.amazon.com) where nodes with GPU co-processing boards can be rented on demand, allow accessibility for HPC resources on constrain budgets and the advantage of testing in few nodes before scaling to thousands of nodes to accelerate processing. An example of Amazon AWS for bioinformatics is reported in [99].

Extrapolating our experience with CAP search with length 11 peptides, using a 400 computers cluster with state of the art GPUs (2012 - Kepler boards), could compute all sequences in less than 2 hours, so exploring more complex algorithms to refine the selection or compute longer sequences would be feasible in hours to days computing time.

PROSPECTIVE ANALYSIS

All different methods mentioned to detect and/or predict antibacterial peptides vary in their mathematical-computational degree of complexity. The methods developed in the last six decades are the first efforts to consolidate Bioinformatics and Biomedicine, and it seems this tendency will continue during the next decades. It is possible to envision that these disciplines will be combined with robotics to generate intelligent “nanorobots” that will “learn” from exhaustive transversal analysis about networks and databases.

From this first effort we have learnt that consolidation will unfold -as in many other disciplines of basic science-, with acknowledgment and regular study of the phenomena. Peptide detection and prediction to develop new pharmaceutical drugs will also follow the same path.

We know computational limitations are not an obstacle for any discipline, the real obstacle lies in the efficiency of the methods used, directly linked to the regularity of the event, its fractality and catastrophic bifurcation points [100, 101].

Fractality is a pattern in the dynamics of the phenomenon studied to determine the points in time where regularity changes and gives way to a new pattern. It is not obvious at first sight but it underlies in every structure and it is used from subatomic particle tracking to Universe accelerating expansion.

How can we understand the construction of 10^{-6} meter sized "intelligent" nanorobots or robotic units that do not recognize regular rhythms in microorganism behavior and their interaction in a watery-lipid medium?

Certainly mathematical-computational methods to detect and predict antibacterial peptides will have to be featured by:

i. Having a greater mathematical orientation to recognize structures with O(nn) degree of complexity and being able to identify in massive database, irregularities in peptide sequences of diverse length.

ii. Being a 100% parallelized to run in GPU and FPGA clusters and grids, where processing average speed for peptide sequences of variable length (< 25aa) is given in tebibytes per second (tebibytes = 2^{40} bytes), and be perfectly capable to differentiate information to avoid unnecessary storage but to learn from it.

iii. Including membrane and watery lipid interaction from peptide lineal sequence reading.

Next generation of prediction methods will include broader interdisciplinary teams, the support of linguists and semiologists will be as necessary as the support

of Dynamic system specialists is now for the way databases are designed to keep information related to microorganisms and there will be also a substantial improvement in the variety of formats to store and access data.

Furthermore, recognizing the poly-pharmacological nature of APs, it is relevant to generate specialized databases of these peptides to study the structure-function relationship of APs. For instance, a database for SCAPs may be useful to understand the nature of the determinants of peptides capable to preferentially target bacteria instead of mammalian membranes. Alternatively, using every known AP may reveal the common features associated to every AP, that is, the association of these peptides with lipid membranes. This knowledge has to impact on the nature and aim of current databases that specialize on antibacterial peptides.

ACKNOWLEDGEMENTS

We acknowledge the technical support provided by Dra. Maria Teresa Lara Ortiz from the Instituto de Fisiologia Celular, UNAM. This work was support in part by grant IN205911 to GDR.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

REFERENCES

[1] Zaffiri, L.; Gardner, J.; Toledo-Pereyra, L. H. *J. Invest. Surg.* **2012**, *25*, 67-77.
[2] Kardos, N.; Demain, A. L. *Appl. Microbiol. Biotechnol.* **2011**, *92*, 677-687.
[3] Swinney, D. C.; Anthony, J. *Nat. Rev. Drug Discov.* **2011**, *10*, 507-519.
[4] Williams, M. *Biochem. Pharmacol.* **2005**, *70*, 1707-1716.
[5] Flordellis, C. S.; Manolis, A. S.; Paris, H.; Karabinis, A. *Curr. Top. Med. Chem.* **2006**, *6*, 1791-1798.
[6] Brötz-Oesterhelt, H.; Brunner, N. A. *Curr. Opin. Pharmacol.* **2008**, *8*, 564-573.
[7] Hopkins, A. L.; Mason, J. S.; Overington, J. P. *Curr. Opin. Struct. Biol.* **2006**, *16*, 127-136.
[8] Arrowsmith, J. *Nat. Rev. Drug Discov.* **2012**, *11*, 17-18.
[9] Jenssen, H.; Hamill, P.; Hancock, R. E. W. *Clin. Microbiol. Rev.* **2006**, *19*, 491-511.
[10] Falconer, S. B.; Czarny, T. L.; Brown, E. D. *Nat. Chem. Biol.* **2011**, *7*, 415-423.
[11] Brogden, K. A. *Nat. Rev. Microbiol.* **2005**, *3*, 238-250.
[12] Hancock, R. E.; Diamond, G. *Trends Microbiol.* **2000**, *8*, 402-410.
[13] Hancock, R. E. *Lancet. Infect. Dis.* **2001**, *1*, 156-164.
[14] Baltzer, S. A.; Brown, M. H. *J. Mol. Microbiol. Biotechnol.* **2011**, *20*, 228-235.

[15] Li, L.; Wang, J.-X.; Zhao, X.-F.; Kang, C.-J.; Liu, N.; Xiang, J.-H.; Li, F.-H.; Sueda, S.; Kondo, H. *Protein Expr. Purif.* **2005**, *39*, 144-151.
[16] Del Rio, G.; Castro-Obregon, S.; Rao, R.; Ellerby, H. M.; Bredesen, D. E. *FEBS Lett.* **2001**, *494*, 213-219.
[17] Verdon, J.; Girardin, N.; Marchand, A.; Héchard, Y.; Berjeaud, J.-M. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 5401-5412.
[18] Marcos, J. F.; Muñoz, A.; Pérez-Payá, E.; Misra, S.; López-García, B. *Annu. Rev. Phytopathol.* **2008**, *46*, 273-301.
[19] Lohner, K.; Blondelle, S. E. *Comb. Chem. High Throughput Screen.* **2005**, *8*, 241-256.
[20] Su, Y.; Li, S.; Hong, M. *Amino Acids* **2013**, *44*, 821-833.
[21] Hilpert, K.; Fjell, C. D.; Cherkasov, A. *Methods Mol. Biol.* **2008**, *494*, 127-159.
[22] Loose, C.; Jensen, K.; Rigoutsos, I.; Stephanopoulos, G. *Nature* **2006**, *443*, 867-869.
[23] Zhu, W. L.; Lan, H.; Park, I.-S.; Kim, J. Il; Jin, H. Z.; Hahm, K.-S.; Shin, S. Y. *Biochem. Biophys. Res. Commun.* **2006**, *349*, 769-774.
[24] Juretić, D.; Vukičević, D.; Petrov, D.; Novković, M.; Bojović, V.; Lučić, B.; Ilić, N.; Tossi, A. *Eur. Biophys. J.* **2011**, *40*, 371-385.
[25] Dings, R. P. M.; Mayo, K. H. *Acc. Chem. Res.* **2007**, *40*, 1057-1065.
[26] Piotto, S. P.; Sessa, L.; Concilio, S.; Iannelli, P. *Int. J. Antimicrob. Agents* **2012**, *39*, 346-351.
[27] Henriques, S. T.; Melo, M. N.; Castanho, M. A. R. B. *Biochem. J.* **2006**, *399*, 1-7.
[28] Nan, Y. H.; Park, I.-S.; Hahm, K.-S.; Shin, S. Y. *J. Pept. Sci.* **2011**, *17*, 812-817.
[29] Li, L.; Shi, Y.; Su, G.; Le, G. *Int. J. Antimicrob. Agents* **2012**, *40*, 337-343.
[30] Matsuzaki, K.; Murase, O.; Fujii, N.; Miyajima, K. *Biochemistry* **1995**, *34*, 6521-6526.
[31] Hinton, G. E.; Osindero, S.; Teh, Y.-W. *Neural Comput.* **2006**, *18*, 1527-1554.
[32] Yang, G.-F.; Huang, X. *Curr. Pharm. Des.* **2006**, *12*, 4601-4611.
[33] Polanco, C.; Samaniego, J. L.; Buhse, T.; Mosqueira, F. G.; Negron-Mendoza, A.; Ramos-Bernal, S.; Castanon-Gonzalez, J. A. *Int. J. Pept.* **2012**, *2012*, 1-2.
[34] Rodríguez Plaza, J. G.; Villalón Rojas, A.; Herrera, S.; Garza-Ramos, G.; Torres Larios, A.; Amero, C.; Zarraga Granados, G.; Gutiérrez Aguilar, M.; Lara Ortiz, M. T.; Polanco Gonzalez, C.; Uribe Carvajal, S.; Coria, R.; Peña Díaz, A.; Bredesen, D. E.; Castro-Obregon, S.; Del Rio, G. *PLoS One* **2013**, *8*.
[35] Dearden, J. C. *J. Comput. Aided. Mol. Des. 17*, 119-127.
[36] Tong, W.; Hong, H.; Xie, Q.; Shi, L.; Fang, H.; Perkins, R. *Curr. Comput. Aided-Drug Des.* **2005**, *1*, 195-205.
[37] Roy, P. P.; Leonard, J. T.; Roy, K. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 31-42.
[38] Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2006**, *25*, 235-251.
[39] Wold, S.; Eriksson, L.; Clementi, S. *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH: Weinheim, Germany, 1995.
[40] Jenssen, H.; Fjell, C. D.; Cherkasov, A.; Hancock, R. E. W. *J. Pept. Sci.* **2008**, *14*, 110-114.
[41] Wang, X.; Zang, M.; Xiao, G. *Stat. Med.* **2013**, *32*, 2292-2307.
[42] Wang, Y.; Ding, Y.; Wen, H.; Lin, Y.; Hu, Y.; Zhang, Y.; Xia, Q.; Lin, Z. *Comb. Chem. High Throughput Screen.* **2012**, *15*, 347-353.
[43] Shu, M.; Jiang, Y.; Yang, L.; Wu, Y.; Mei, H.; Li, Z. *Protein Pept. Lett.* **2009**, *16*, 143-149.
[44] González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990-1995.
[45] Cruz-Monteagudo, M.; González-Díaz, H.; Borges, F.; Dominguez, E. R.; Cordeiro, M. N. D. S. *Chem. Res. Toxicol.* **2008**, *21*, 619-632.

[46] Cruz-Monteagudo, M.; González-Díaz, H.; Agüero-Chapín, G.; Santana, L.; Borges, F.; Domínguez, E. R.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1909-1923.
[47] Cruz-Monteagudo, M.; Munteanu, C. R.; Borges, F.; Cordeiro, M. N. D. S.; Uriarte, E.; Chou, K.-C.; González-Díaz, H. *Polymer (Guildf).* **2008**, *49*, 5575-5587.
[48] Cruz-Monteagudo, M.; Munteanu, C. R.; Borges, F.; Cordeiro, M. N. D. S.; Uriarte, E.; González-Díaz, H. *Bioorg. Med. Chem.* **2008**, *16*, 9684-9693.
[49] González-Díaz, H.; Prado-Prado, F. J. *J. Comput. Chem.* **2008**, *29*, 656-667.
[50] Prado-Prado, F. J.; Martinez de la Vega, O.; Uriarte, E.; Ubeira, F. M.; Chou, K.-C.; González-Díaz, H. *Bioorg. Med. Chem.* **2009**, *17*, 569-575.
[51] Vilar, S.; González-Díaz, H.; Santana, L.; Uriarte, E. *J. Theor. Biol.* **2009**, *261*, 449-458.
[52] Agüero-Chapín, G.; Antunes, A.; Ubeira, F. M.; Chou, K.-C.; González-Díaz, H. *J. Chem. Inf. Model.* **2008**, *48*, 2265-2277.
[53] Agüero-Chapín, G.; Gonzalez-Díaz, H.; de la Riva, G.; Rodríguez, E.; Sanchez-Rodríguez, A.; Podda, G.; Vazquez-Padrón, R. I. *J. Chem. Inf. Model.* **2008**, *48*, 434-448.
[54] Gonzalez-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. *J. Proteome Res.* **2007**, *6*, 904-908.
[55] González-Díaz, H.; Pérez-Bello, A.; Uriarte, E. *Polymer (Guildf).* **2005**, *46*, 6461-6473.
[56] González-Díaz, H.; Agüero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. *J. Comput. Chem.* **2007**, *28*, 1049-1056.
[57] De Armas, R. R.; Díaz, H. G.; Molina, R.; Uriarte, E. *Biopolymers* **2005**, *77*, 247-256.
[58] González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. *Proteomics* **2008**, *8*, 750-778.
[59] González-Díaz, H.; Prado-Prado, F.; Ubeira, F. M. *Curr. Top. Med. Chem.* **2008**, *8*, 1676-1690.
[60] Ferino, G.; González-Díaz, H.; Delogu, G.; Podda, G.; Uriarte, E. *Biochem. Biophys. Res. Commun.* **2008**, *372*, 320-325.
[61] Polanco, C.; Samaniego, J. L. *Acta Biochim. Pol.* **2009**, *56*, 167-176.
[62] Likic, V. A.; Dolezal, P.; Celik, N.; Dagley, M.; Lithgow, T. *Methods Mol. Biol.* **2010**, *619*, 271-284.
[63] Neuwald, A. F.; Poleksic, A. *Nucleic Acids Res.* **2000**, *28*, 3570-3580.
[64] Ghosh, T. S.; Gajjalla, P.; Mohammed, M. H.; Mande, S. S. *Genomics* **2012**, *99*, 195-201.
[65] Melvin, I.; Weston, J.; Noble, W. S.; Leslie, C. *PLoS Comput. Biol.* **2011**, *7*, e1001047.
[66] Bai, L.; Liang, J.; Dang, C.; Cao, F. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1509-1522.
[67] Das, S.; Idicula, S. M. *Adv. Exp. Med. Biol.* **2010**, *680*, 181-188.
[68] Mooney, C.; Haslam, N. J.; Pollastri, G.; Shields, D. C. *PLoS One* **2012**, *7*, e45012.
[69] Zomaya, A. Y. Parallel Computing for Bioinformatics and Computational Biology: Models, Enabling Technologies, and Case Studies http://www.wiley.com/WileyCDA /WileyTitle/productCd-0471718483.html (accessed Dec 9, 2014).
[70] Braga-Neto, U.; Dougherty, E. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*; IEEE, **2011**; pp. 1057-1059.
[71] Andreotti, S.; Klau, G. W.; Reinert, K. **2011**.
[72] Kushwaha, S. K.; Shakya, M. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*; IEEE, **2009**; pp. 98-101.
[73] Belda, I.; Llorà, X.; Giralt, E. *Soft Comput.* **2005**, *10*, 295-304.
[74] Gallia, J.; Lavrich, K.; McGough, R.; Tan-Wilson, A.; Madden, P. H. In *2012 IEEE 2nd International Conference on Computational Advances in Bio and medical Sciences (ICCABS)*; IEEE, **2012**; pp. 1-6.

[75] Rwebangira, M.; Burge, L.; Southerland, W. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*; IEEE, **2011**; pp. 251-254.
[76] Shi, J.; Chen, B.; Wu, F.-X. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*; IEEE, **2011**; pp. 191-196.
[77] Luethy, R.; Hoover, C. *Drug Discov. Today BIOSILICO* **2004**, *2*, 12-17.
[78] Hasan, L.; Al-Ars, Z.; Vassiliadis, S. In *2007 International Conference on Design & Technology of Integrated Systems in Nanoscale Era*; IEEE, **2007**; pp. 92-97.
[79] Chandramohan, P.; Geist, A.; Samatova, N. In *19th IEEE International Parallel and Distributed Processing Symposium*; IEEE, 2005; p. 72b - 72b.
[80] Vouzis, P. D.; Sahinidis, N. V. *Bioinformatics* **2011**, *27*, 182-188.
[81] Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. *BMC Bioinformatics* **2009**, *10*, 421.
[82] Weber, R.; Peterson, G. D.; Hettich, R. In *2012 Symposium on Application Accelerators in High Performance Computing*; IEEE, **2012**; pp. 102-110.
[83] Lin, F.; Schröder, H.; Schmidt, B. *J. VLSI Signal Process. Syst. Signal Image. Video Technol.* **2007**, *48*, 185-188.
[84] Abbas, N. Acceleration of a bioinformatics application using high-level synthesis, **2012**.
[85] Li, I. T. S.; Shum, W.; Truong, K. *BMC Bioinformatics* **2007**, *8*, 185.
[86] Vidanagamachchi, S. M.; Dewasurendra, S. D.; Ragel, R. G.; Niranjan, M. In *2011 6th International Conference on Industrial and Information Systems*; IEEE, **2011**; pp. 140-145.
[87] Pascoe, C.; Box, D.; Lam, H.; George, A. In *2012 Symposium on Application Accelerators in High Performance Computing*; IEEE, **2012**; pp. 111-120.
[88] Awad, M. In *2009 International Conference on Field Programmable Logic and Applications*; IEEE, **2009**; pp. 564-568.
[89] Nickolls, J.; Dally, W. J. *IEEE Micro* **2010**, *30*, 56-69.
[90] Dematté, L.; Prandi, D. *Brief. Bioinform.* **2010**, *11*, 323-333.
[91] Hwu, W. W. **2011**.
[92] Hwu, W. W. **2011**.
[93] Bogdan, I.; Coca, D.; Rivers, J.; Beynon, R. J. *Bioinformatics* **2007**, *23*, 724-731.
[94] Coca, D.; Bogdan, I.; Beynon, R. J. *Methods Mol. Biol.* **2010**, *604*, 163-185.
[95] Zhang, J.; McQuillan, I.; Wu, F.-X. *Proteomics* **2011**, *11*, 3779-3785.
[96] Zhang, J.; McQuillan, I.; Wu, F. In *2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*; IEEE, **2011**; pp. 208-213.
[97] Polanco González, C.; Nuño Maganda, M. A.; Arias-Estrada, M.; del Rio, G. *PLoS One* **2011**, *6*, e21399.
[98] Arias-Estrada, M.; Garcia-Ordaz, D.; Nuno-Maganda, M.; Polanco, C.; Rio, G. Del. In *International Supercomputing Conference*; **2012**.
[99] Halligan, B. D.; Geiger, J. F.; Vallejos, A. K.; Greene, A. S.; Twigger, S. N. *J. Proteome Res.* **2009**, *8*, 3148-3153.
[100] Hsu, Jong-Ping (University of Massachusetts Dartmouth, U.; Hsu, Leonardo (University of Minnesota, U. *A Broader View of Relativity (World Scientific)*; 2nd ed.; World Scientific, **2006**.
[101] Wise, M. N. *Growing Explanations: Historical Perspectives on Recent Science*; Wise, M. N., Ed.; Duke University Press, **2004**; p. 360.

CHAPTER 2

Computational Chemistry for Photosensitizer Design and Investigation of DNA Damage

Kazutaka Hirakawa*

Department of Applied Chemistry and Biochemical Engineering, Graduate School of Engineering, Shizuoka University, Japan

Abstract: Computational chemistry can be used for the prediction of photochemical reactivity and the design of photosensitizers for cancer phototherapy. For example, the activity of a photosensitizer for DNA damage can be estimated from the calculation of the HOMO energy of the molecules. In general, DNA damage is mediated by the following two processes: 1) photo-induced electron transfer from the DNA base to the photoexcited photosensitizer and 2) base modification by singlet oxygen generation through photo-energy transfer from the photosensitizer to oxygen. The DNA-damaging activity of the photosensitizer through electron transfer is closely related to the HOMO energy level of the molecule. It has been demonstrated that the extent of DNA damage photosensitized by xanthone analogues is proportional to the energy gap between the HOMO level of the photosensitizer and that of guanine. In addition, computational chemistry can be used to investigate the mechanism of the chemopreventive effect on phototoxicity. Furthermore, the molecular orbital calculation is useful to design a photosensitizer in which the activity of singlet oxygen generation is controlled by DNA recognition. Singlet oxygen is an important reactive oxygen species to attack cancer. The control of singlet oxygen generation by DNA is necessary to achieve the tailor-made cancer photo-therapy. Several porphyrin photosensitizers have been designed on the basis of the molecular orbital calculation to control the activity of singlet oxygen generation.

Keywords: *Ab initio* molecular orbital calculation, density functional treatment (DFT), DNA damage, electron transfer, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), molecular mechanics calculation, photosensitizer, porphyrin, redox potential, singlet oxygen ($^1O_2$), Zerner's intermediate neglect of differential overlap (ZINDO) procedure.

INTRODUCTION

Computational chemistry is an important tool to design medical drugs. This

***Correspoding author Kazutaka Hirakawa:** Department of Applied Chemistry and Biochemical Engineering, Graduate School of Engineering, Shizuoka University, Japan; Tel/Fax: +81-53-478-1287;
E-mail: tkhirak@ipc.shizuoka.ac.jp

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

method can be applied for the prediction of photochemical reactivity and the design of photosensitizers for cancer phototherapy. To examine the photo-carcinogenicity of drugs, complicated experiments using cells, various chemicals, and expensive apparatuses are required. Furthermore, many samples are inversely consumed by these experiments. Since computational study does not require such samples or apparatuses, this method can reduce the cost and magnitude of the task. In addition, computational study can support experimental study. In the case of drug design, the prediction of a drug's characteristics through computational study is important chart. Although not all computational studies are simple, several studies can be carried out following a relatively simple method. For example, the activity of a photosensitizer for DNA damage can be simply estimated from the calculation of the highest occupied molecular orbital (HOMO) energy of the molecules. In this chapter, several examples of computational chemistry for studies in the photobiological field are introduced; the molecular design of a photosensitizer is discussed as well. In addition, related applications for the photochemistry of porphyrins are also presented.

Photosensitized Reaction and UVA Carcinogenesis

Exposure to solar ultra-violet (UV) radiation is undoubtedly linked to skin carcinogenesis [1]. It has been well demonstrated in early studies that UVB (280 ~ 320 nm) radiation, which constitutes about 5% of the solar UV radiation that reaches the surface of the earth, directly activates the DNA molecule to generate dipyrimidine photoproducts such as cyclobutane pyrimidine dimers and pyrimidine (6-4) pyrimidone photoadducts, resulting in mutations and carcinogenesis. However, many studies have provided sufficient evidence that UVA radiation (320 ~ 400 nm), which accounts for the major portion of the solar UV radiation, is also mutagenic and carcinogenic, although it is unlikely that UVA directly activates DNA bases to produce dipyrimidine photoproducts [2, 3]. It is, therefore, generally recognized that solar UVA carcinogenesis involves a mechanism by which UVA radiation indirectly induces DNA damage through photosensitized reactions mediated by intracellular chromophores. Accordingly, a variety of cellular compounds have been considered to be potential endogenous photosensitizers. In addition, certain drugs may act as exogenous photosensitizers.

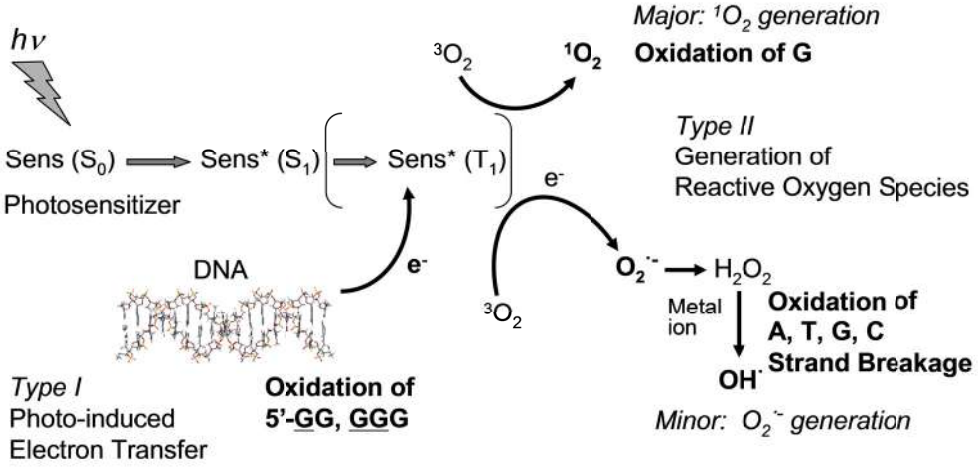


Figure 1: Mechanisms of the photosensitized DNA damage.

Photosensitized DNA Damage

In general, the excited photosensitizer induces damage of cellular components, including DNA, *via* the generation of reactive oxygen species such as singlet oxygen ($^1O_2$) or the oxidative electron transfer mechanism, leading to cell death or carcinogenesis. DNA damage *via* photo-induced electron transfer is called the Type I mechanism, whereas the reactive oxygen mediated damage is referred to as the Type II mechanism (Fig. **1**). The Type I mechanism does not require oxygen for the induction of DNA damage, whereas the Type II mechanism proceeds only in the presence of oxygen. The Type I mechanism involves electron transfer from the DNA base to the excited photosensitizer. This mechanism is dependent on the energy of the molecular orbital (MO) of the photosensitizer (Sens), its excited state (Sens*), and its nucleobase, namely, the oxidation potential of the DNA base and the reduction potential of the excited photosensitizer. Guanine has the lowest oxidation potential among the four DNA bases, that is, guanine is most likely to be oxidized. The guanine radical cation formed through the electron transfer reacts with water or molecular oxygen, leading to the formation of the oxidized product of guanine. These photoproducts cause mutation and cancer. In the Type I mechanism, one electron is transferred from guanine to photoexcited sensitizers, resulting in damage at the consecutive guanine sequence in double-stranded DNA [4, 5]. Various endogenous and exogenous photosensitizers mediate DNA photodamage *via* this mechanism. The major Type II mechanism involves the

generation of $^1O_2$ from photoexcited sensitizers, resulting in damage to guanines without preference for consecutive guanines. In the minor Type II mechanism, a superoxide anion radical ($O_2^{\bullet-}$) is generated, and DNA damage is then induced by reactive species generated through the interaction of hydrogen peroxide (H_2O_2), which is formed from the dismutation of $O_2^{\bullet-}$, with metal ions. Computational chemistry can be used to evaluate the activity of photosensitizers.

Photodynamic Therapy

Photodynamic therapy (PDT), which is a less invasive treatment for cancer, employs a photosensitizer and visible light to produce oxidative stress in cells and ablate cancerous tumors [6-8]. PDT is also used for treating some non-malignant conditions that are generally characterized by the overgrowth of unwanted or abnormal cells. The human tissue has relatively high transparency for visible light, especially red light, and visible light hardly demonstrates any side-effects. Because $^1O_2$ can be easily generated by visible light, $^1O_2$ is considered as an important reactive species of PDT. Critical sites of the generated $^1O_2$ include mitochondria and lipid membranes [6-9]. Moreover, DNA is also an important target biomolecule of photosensitized reactions [10-13]. Computational study can be used for the design of a photosensitizer of PDT.

RELATIONSHIP BETWEEN THE DNA-DAMAGING ABILITIES OF PHOTOSENSITIZERS AND THEIR HOMO ENERGIES

Computational chemistry is important tool to evaluate the activity of photosensitizer. The oxidative activity of photosensitizer depends on the HOMO level. Lower HOMO energy level is advantageous for the oxidative electron transfer (Fig. **2**). The calculation of HOMO energy may be applied to predict the activity of various photosensitizers. Thus, the mechanism of DNA photodamage induced by xanthone (XAN) analogues (Fig. **3**), exogenous photosensitizers and the relationship between the DNA-damaging abilities and their HOMO energies were investigated. Derivatives of XAN and its analogues, thioxanthone (TXAN) and acridone (ACR), have been isolated from various plants [14-18] and used as antitumor drugs [19]. The mechanism of DNA damage induced by UVA irradiation in the presence of XAN analogues was examined using ^{32}P-labeled

DNA fragments obtained from the *p53* tumor suppressor gene. The HOMO energies of XAN analogues were estimated from *ab initio* MO calculation.

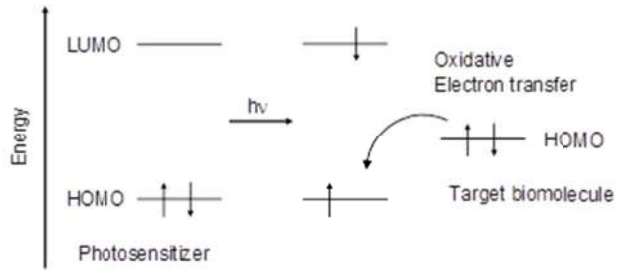


Figure 2: HOMO energy level and the oxidative electron transfer from the target biomolecule.

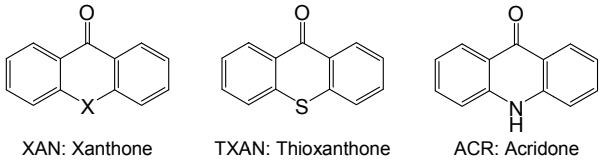


Figure 3: Structures of xanthone analogues.

Calculation of HOMO Energies of Xanthone Analogues

HOMO energies of XAN, TXAN and ACR were estimated from *ab initio* MO calculation at Hartree-Fock 6-31G* level. Structures of these molecules were optimized by calculation of equilibrium geometry at Hartree-Fock 6-31G* level. These calculations were performed using Spartan '02 for Windows (Wavefunction Inc., CA, USA). Calculated HOMO energies of XAN, TXAN and ACR were 8.57, 8.07 and 7.75 eV, respectively. These values are larger than that of 5'-G of GG in double-stranded DNA (6.73 eV) [20], suggesting that photoexcited XAN analogues can oxidize 5'-G of GG through electron transfer from an energetic viewpoint.

Experimental Results of DNA Damage Photosensitized by Xanthone Analogues

DNA damage induced by UVA irradiation in the presence of XAN analogues was examined using ^{32}P-labeled DNA fragments by an electrophoresis. The damaged sites were determined by direct comparison of the positions of the oligonucleotides with those produced by the chemical reactions of the Maxam-

Gilbert procedure [21]. It was also measured that the content of 8-oxo-7,8-dihydro-2'-deoxyguanine (8-oxo-G) [22], an oxidative product of 2'-deoxyguanosine (dG), formed by photoexcited XAN analogues with an electrochemical detector coupled to high-performance liquid chromatography. The present study has demonstrated that photoexcited XAN analogues as well as riboflavin, a Type I photosensitizer [23], generate piperidine-labile products specifically at 5'-G of GG sequence and both guanines of 5'-AGGA sequence in double-stranded DNA. Effects of scavengers of reactive oxygen and D_2O on DNA damage suggested that the contribution of reactive oxygen (Type II mechanism) to the DNA photodamage is negligibly small. Therefore, these results can be reasonably explained by assuming that nucleobase oxidation is induced by photoexcited XAN analogues mainly through electron transfer (Type I mechanism) (Fig. **4**). Guanine is most easily oxidized among the four DNA bases because the oxidation potential of guanine is lower than that of the other DNA bases [24-26]. MO calculations have revealed that 5'-G in GG sequence in double-stranded DNA significantly lowers the HOMO energy [20, 27]. Therefore, the cation radical on the 5'-G in GG sequence arises from either the initial electron abstraction of this guanine by photoexcited XAN analogues or through charge migration from a relatively distant one-electron oxidized nucleobase [24, 28-32]. The formed guanine cation radicals may react with water molecules to form the C-8 OH adduct radical [24, 28]. This radical may be converted by a reducing process into 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FapyGua) residue, a piperidine-labile product [24, 28]. On the other hand, competitive oxidation, which may be achieved by molecular oxygen, gives rise to 8-oxo-G [24, 28, 33]. The formation of 8-oxo-G causes DNA misreplication that may lead to mutations such as G-C → T-A transversion [34, 35]. Although the 8-oxo-G site is not efficiently cleaved under piperidine treatment [36], 8-oxo-G can be converted into piperidine-labile products (*e.g.* imidazolone, oxazolone) through further oxidation [24, 37, 38]. Imidazolone and oxazolone might be also produced through deprotonation of guanine cation radical followed by reaction with molecular oxygen [24, 28, 39, 40]. The present study regarding photon fluence dependence of DNA photodamage suggests that 8-oxo-G oxidation into piperidine-labile products can occur in an irradiation dose-dependent manner. It has been reported that imidazolone and oxazolone are major oxidation products of

guanine by the Type I mechanism [39, 40]. Imidazolone forms a stable base pair with G comparable with the Watson-Crick G-C base pair [37, 38] and may cause G-C → CG transversion [41-43]. These transversions can partly explain the mutation induced by UVA as previously reported [2].

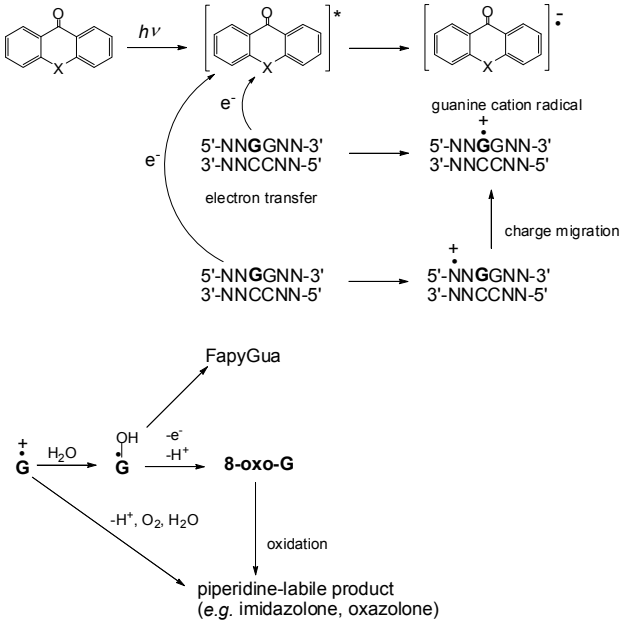


Figure 4: Possible mechanism of base oxidation at 5' site of GG sequence in double-stranded DNA induced by UVA-irradiated xanthone analogues.

Relationship between the Calculated HOMO Energy and DNA Damage

The extents of DNA damage increased depending on the HOMO energies of XAN, TXAN and ACR. Fig. **5** shows the plots of the quantum yields of DNA damage against the gaps of HOMO energies (ΔE) between the photosensitizers and 5'-G of GG. The quantum yield of 8-oxo-G formation ($\Phi_{8\text{-oxo-G}}$) was estimated using the photon fluence and molar absorption coefficients of XAN analogues. Similarly, the relative quantum yield of the piperidine-labile product (Φ_P) was estimated from the results of the electrophoresis. The logarithm plots indicate that the DNA-damaging abilities of these photosensitizers almost increased exponentially with ΔE (Fig. **5A**). The plots have shown that Φ_P is almost proportional to $\Phi_{8\text{-oxo-G}}$. The electron transfer reaction should also be affected by an interaction between DNA and photosensitizer. The absorption and

fluorescence spectra of XAN analogues were not significantly changed by addition of DNA (data not shown), suggesting that noncovalent bonding to DNA can be negligible. Therefore, the quantum yield of the electron transfer reaction may be determined by the collision frequency between photoexcited XAN analogues and DNA in addition to ΔE. The collision frequency can be determined by the diffusion control rate coefficient (k_{dif}), concentration of nucleobase ([base]), triplet quantum yield (Φ_{isc}) and lifetime of excited triplet state (τ_0^T) of photosensitizer. Because k_{dif} and [base] can be taken as constants, the collision frequency varies depending on $\Phi_{isc} \times \tau_0^T$. The values of Φ_{isc} and τ_0^T of XAN analogues were previously reported [44]. The values of $\Phi_{8\text{-oxo-G}}$ and Φ_P were divided by $\Phi_{isc} \times \tau_0^T$ and plotted against ΔE (Fig. **5B**). These plots also showed an increase of extent of DNA damage exponentially, almost depending on ΔE. Strictly, the DNA-damaging ability of the photosensitizer should be determined by not only ΔE but also other factors containing the free energy of the electron transfer reaction and an interaction between DNA and photosensitizer. However, these results suggest that the DNA-damaging abilities of these XAN analogues practically depend on their HOMO energies mainly.

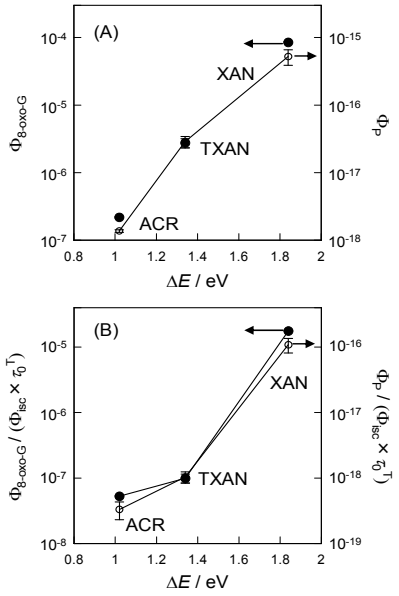


Figure 5: Relationship between the DNA-damaging abilities of xanthone analogues and their HOMO energies. The quantum yield of 8-oxo-G formation ($\Phi_{8\text{-oxo-G}}$) and the relative quantum yield of piperidine-labile product (Φ_P) by UVA (wavelength: 365 nm) irradiation in the presence xanthone analogues (A) and those values divided by ($\Phi_{isc} \times \tau_0^T$) (B) are plotted against ΔE.

Summary of the Relationship between DNA-damaging Activity of Photosensitizer and the HOMO Energy

In summary, this study demonstrated that photoexcited XAN analogues, exogenous photosensitizers, mediate poly-G-specific DNA oxidation through electron transfer. DNA oxidation through photoinduced electron transfer may play an important role in photocarcinogenesis mediated by photosensitizers. This study has also shown that the DNA-damaging abilities of XAN analogues increase exponentially with their HOMO energies. It is concluded that the DNA-damaging ability of derivatives of XAN analogues can be roughly estimated from their HOMO energies.

COMPUTATIONAL EVALUATION OF CHEMOPREVENTIVE ACTION ON PHOTOSENSITIZED DNA DAMAGE

Photosensitized damage to biomacromolecules, such as DNA and protein, participates in phototoxicity and photogenotoxicity of drugs and solar-UV carcinogenesis [1]. Various endogenous molecules and natural products act as photosensitizers [23, 24, 45-49]. In addition, a side effect of PDT [6] is also caused by photosensitized biomacromolecular damage. An antioxidant can scavenge reactive oxygen species generated through photosensitized reaction and protect against cancer occurrence [50]. For example, β-carotene is an efficient scavenger of $^1O_2$. However, β-carotene, vitamin A, and vitamin E generate reactive oxygen species through the oxidation process, leading to oxidative DNA damage [51, 52]. Indeed, an excess amount of these antioxidants elevates cancer incidence [53-55]. A physical sunscreen does not show a side effect, but cannot effectively protect the phototoxicity induced by visible-light. It has been demonstrated that the photoexcited pteridine moiety of folic acid is effectively quenched by the aminobenzoyl moiety through intramolecular electron transfer, resulting in the inhibition of DNA photodamage by the photoexcited folic acid [48]. This result leads us to the idea that an effective quencher can be used as a chemopreventive agent for photodamage of biomacromolecules. In this study, the action of XAN derivatives (bellidifolin (BEL), gentiacaulein (GEN), norswertianin (NOR), and swerchirin (SWE)) (Fig. **6**) on photosensitized DNA

damage was demonstrated. These XAN derivatives were isolated from various plants [56-60]. The preventive effects of XAN derivatives on DNA damage by photoexcited riboflavin were examined. The mechanism of the preventive effect on DNA photodamage was investigated by *ab initio* MO calculations [61].

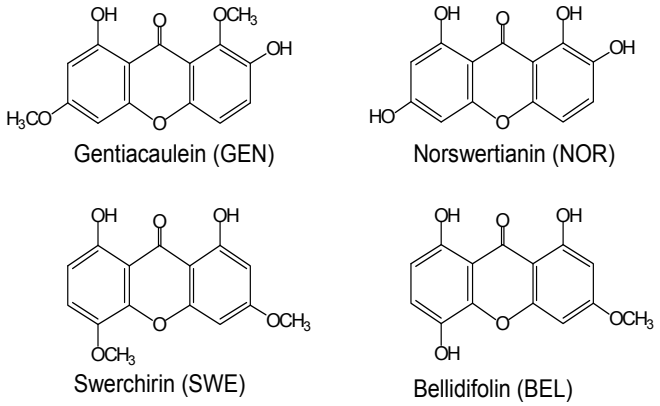


Figure 6: Structures of xanthone derivatives.

Calculation of Excited Triplet State Energies and Ionization Potentials of Xanthone Derivatives and Riboflavin

The excited triplet state energies of XAN derivatives and riboflavin were estimated from the density functional treatment (DFT) at the B3LYP/6-31G* level. The ionization potentials (*IP*s) of XAN derivatives and riboflavin were estimated from *ab initio* MO calculation at the Hartree-Fock 6-31G* level. The structures of these molecules were optimized by the calculation of equilibrium geometry at the Hartree-Fock 6-31G* level. These calculations were performed utilizing Spartan'02 for Windows (Wavefunction Inc.). The energies of an excited triplet state (E^T) of XAN derivatives were higher than those of riboflavin (Table **1**), suggesting that XAN derivatives cannot quench the excited riboflavin through excitation energy transfer. The calculated *IP*s of BEL, GEN, NOR, and SWE were larger than those of riboflavin (Table **1**), suggesting that the electron transfer from XAN derivatives to photoexcited riboflavin is possible. These results have shown that the excited triplet state of riboflavin can be quenched through electron transfer from XAN derivatives and subsequent reverse electron transfer.

Table 1: Calculated triplet energy and *IP*s of xanthone derivatives and riboflavin

Compounds	E^T / kcal mol^{-1}	*IP* / eV
BEL	68.40	8.11
GEN	70.27	8.14
NOR	66.98	8.14
SWE	70.74	8.20
XAN	79.52	8.57
Riboflavin	57.00	8.80

E^T: Energies of excited triplet state of compounds calculated from *ab initio* MO method.

Experimental Results of Preventive Effect of Xanthone Analogues on DNA Damage Photosensitized by Riboflavin

XAN derivatives, BEL, GEN, NOR, and SWE, inhibited DNA damage induced by photoexcited riboflavin. Photoexcited riboflavin oxidizes specifically at the underlined G of 5'-GG and 5'-GGG sequences in double-stranded DNA through photoinduced electron transfer (Type I mechanism). The underlined G in the 5'-GG and 5'-GGG sequences acts as a hole-trap [20, 27] and is finally oxidized through hole-transfer [62]. These photosensitizers generate 8-oxo-G and piperidine-labile products, such as imidazolone and oxazolone, at consecutive G residues [23, 63]. The piperidine-labile products can be generated *via* further oxidation of 8-oxo-G. These photoproducts cause mutation and/or cancer [34, 35, 37, 38, 41, 42]. A qualitative evaluation by the electrophoresis showed that GEN and NOR act as a protector for piperidine-labile DNA photodamage, whereas the preventive action of BEL and SWE is very weak. The preventive action of these XAN derivatives was evaluated quantitatively by the inhibitory effect of 8-oxo-G formation by photoexcited riboflavin (Fig. **7**). The preventive action of XAN derivatives increased in the following order: GEN > NOR >> BEL > SWE. Especially, 5 μM GEN completely inhibited 8-oxo-G formation by 50 μM riboflavin. These findings have shown that GEN can act as most effective chemopreventive agent for photosensitized DNA damage among the four XAN derivatives.

Mechanism of the Preventive Effect

Spectroscopic studies have shown that these compounds cannot act as a physical sunscreen. The fluorescence intensity of riboflavin was less affected by addition of

XAN derivatives. These findings exclude the interaction between the ground state of riboflavin and XAN derivatives and the quenching of the excited singlet state of riboflavin. The possible mechanism of the prevention of the DNA photodamage is the quenching of the excited triplet state of riboflavin (Fig. **8**). MO calculations have shown that, although the quenching through energy transfer is impossible, the mechanism through electron transfer is possible. The electron transfer from XAN derivatives to the excited triplet state of riboflavin generates a radical ion pair, and reverse electron transfer regenerates a ground state of riboflavin and XAN derivative. The excitation energy of the photosensitizer is dispersed as thermal energy through this quenching mechanism. The τ_0^T of riboflavin is 22 μs in aqueous media [64]. The rate constant of quenching reaction of triplet excited state of riboflavin is required to be the magnitude at least comparable to the decay rate constant ($1/\tau_0^T = 4.5 \times 10^4$ s^{-1}). The value of this rate constant is close to that of the diffusion control reaction rate constant (k_{dif}) in this experimental condition. The value of k_{dif} can be estimated from following equation:

$$k_{dif} = 8RT[Q]/3\eta \quad (\mathbf{1})$$

where R, T, [Q], and η are gas constant, absolute temperature, concentration of XAN derivatives, and viscosity of water (8.91×10^{-4} kg m^{-1} s^{-1}), respectively. Therefore, the excited riboflavin should be quenched by XAN derivatives in the diffusion control process.

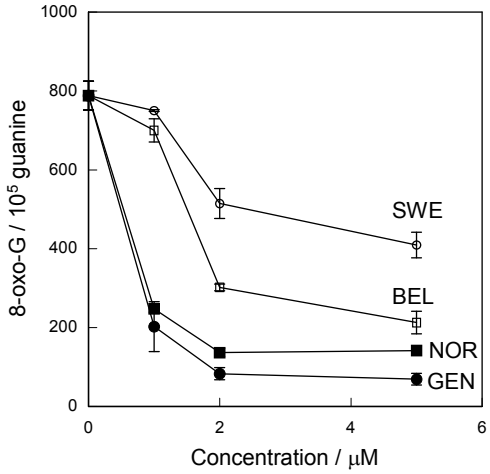


Figure 7: Effects of xanthone derivatives on the formation of 8-oxo-G by photoexcited riboflavin.

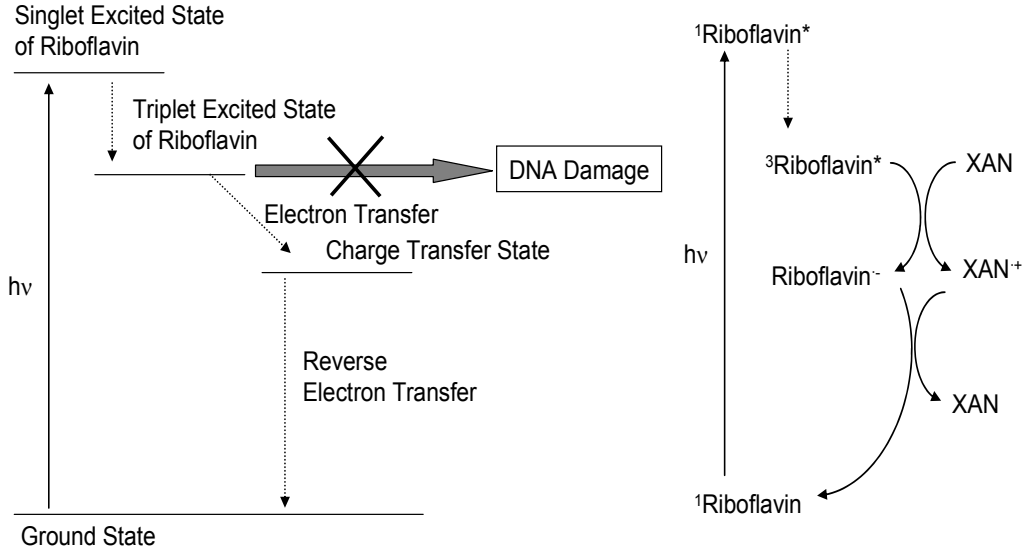


Figure 8: Mechanism of chemopreventive effect of xanthone derivatives on DNA photodamage by riboflavin.

Summary of the Mechanism of Chemoprevention for Phototoxicity

In summary, this study demonstrated that XAN derivatives prevent DNA damage by photoexcited Type I photosensitizers by quenching of the excited triplet state of a photosensitizer. This chemopreventive mechanism is not based on antioxidation or the effect of sunscreen. This preventive mechanism may be used for the novel chemoprevention of phototoxicity, photogenotoxicity, and solar carcinogenesis. An antioxidant, such as β-carotene, can scavenge reactive oxygen species generated through photosensitized reaction and protect against cancer occurrence [50]. However, β-carotene, vitamin A, and vitamin E generate reactive oxygen species through the oxidation process, leading to oxidative DNA damage [51, 52]. Indeed, an excess amount of these antioxidants elevates cancer incidence [53-55]. Although this study does not exclude the possibilities of actions as antioxidant or sunscreen by XAN derivatives, these results have shown that XAN derivatives act as effective quencher and protect photosensitized DNA damage. The quenching mechanism of an excited photosensitizer does not lead to the formation of a secondary reactive species. Furthermore, this quencher can protect from the phototoxicity induced by visible-light, which is difficult to shade with a physical sunscreen.

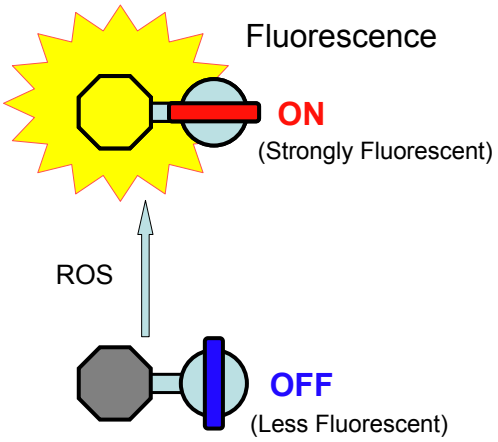


Figure 9: Scheme of the formation of strongly fluorescent pteridines through ROS-mediated decomposition of folic acid analogues.

COMPUTATIONAL INVESITIGATION OF FLUORESCENCE PROBE FOR REACTIVE OXYGEN DETECTION

Next topic is the evaluation of the possibility of molecules for fluorescence probe to detect reactive oxygen species (ROS) (Fig. **9**). Molecular probes and molecular biosensors for ROS are important in environmental and bioanalytical sciences [65-68]. The modification of biomacromolecules upon exposure to ROS, including H_2O_2, $O_2^{\bullet-}$, •OH, and $^1O_2$, is the likely initial event involved in the induction of the mutagenic and lethal effects of various oxidative stress agents [69-73]. Therefore, the activity of ROS generation by various compounds is closely related to their toxicity, carcinogenicity, or both. H_2O_2 is a long-lived ROS and plays an important role in biomacromolecular damage induced by various chemical compounds [71-73]. As mentioned above, $^1O_2$ is also an important ROS and can mediate the oxidative degradation of many molecules [65-68]. The detection of $^1O_2$ and determination of the quantum yield of $^1O_2$ generation (Φ_Δ) by a photosensitizer are necessary to evaluate the PDT activity or phototoxic risk of photosensitizers. The measurement of the near-infrared emission through a deactivation process of $^1O_2$ is sensitive and one of the most important methods to detect $^1O_2$ [73-75], but this method requires the use of costly apparatuses. Another sensitive method for $^1O_2$ detection is fluorometry using a molecular probe, which is the less fluorescent precursor of the fluorescent molecule [76-79]. Fluorometry is very sensitive and

useful method for other ROS including H_2O_2. In general, the sensitive fluorometry of ROS requires costly fluorescent probe compounds. In this section, the fluorometry of H_2O_2 and $^1O_2$ (photosensitized $^1O_2$ generation) using folic acid and its commercially available less fluorescent analogue, *N*-[4-[[(2,4-diamino-6-pteridinyl)methyl]methylamino]benzoyl]-*L*-glutamic acid (methotrexate; MTX, Fig. **10**) [80], was investigated. UV light or a photosensitized reaction causes the oxidation of folic acid, leading to the formation of pterine-6-carboxylic acid (PCA) and aminobenzoyl-*L*-glutamic acid (ABG) [48, 81-84]. The fluorescence intensity of PCA is strong, although folic acid itself scarcely fluoresces [48, 85]. This fluorescence enhancement mechanism is similar to that of fluorescence probes to detect ROS and reactive intermediates of peroxide. Such a character of folic acid led us to hypothesize that the fluorescence analysis of decomposed folic acid could be used as an indicator of oxidative stress. The determination method of Φ_Δ using these molecules is also reported [85].

Calculation of the Molecular Orbital Energies of Folic Acid and MTX

The MO energy level of folic acid and MTX was estimated from the *ab initio* MO calculation at the Hartree-Fock/6-31G* level. The structures of these molecules were also optimized by this calculation. The order of the energy levels of MO in the ground state was calculated by this method to estimate the electron transfer direction in the photoexcited state. This calculation was performed utilizing Spartan'06.

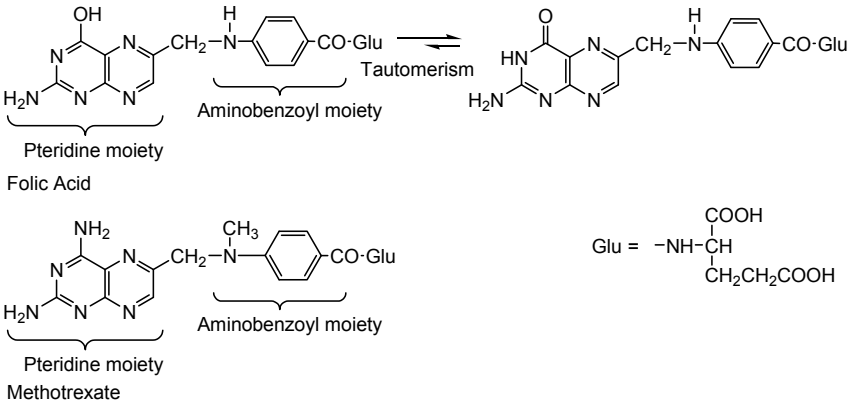


Figure 10: Structures of folic acid and methotrexate.

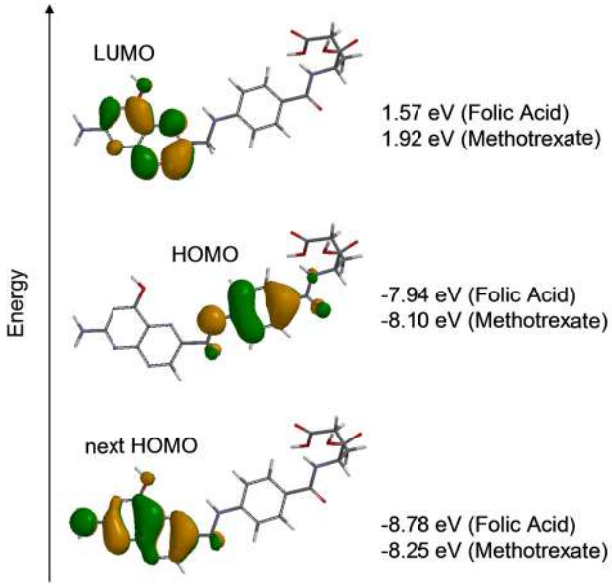


Figure 11: Calculated MO energy levels of folic acid analogues.

Strategy for the Fluorometry of Reactive Oxygen Species Using Folic Acid and Methotrexate

The calculated energy level of the MO showed that the aminobenzoyl moieties of folic acid and MTX are the electron-donating groups (Fig. **11**), suggesting that the fluorescence from their pteridine moieties can be quenched through intra-molecular electron transfer from their aminobenzoyl moieties [48, 80, 81]. It has been reported that the C^9-N^{10} bonds connecting the pteridine and aminobenzoyl moieties of folic acid [48, 80-85] and its analogues [80, 85-89] are unstable against oxidative stress agents. Reactive oxygen-mediated cleavage of this C^9-N^{10} bond or oxidation of the electron-donating moiety should recover the fluorescence intensity of their pteridine moieties (Fig. **12**).

Experimental Results of Fluorometry of Hydrogen Peroxide Using Folic Acid Analogues

The absorption and fluorescence spectra of folic acid and related compounds are presented in Fig. **13**. In general, pteridine compounds absorb the ultra-violet A region and show fluorescence in the visible-light region. Although folic acid itself

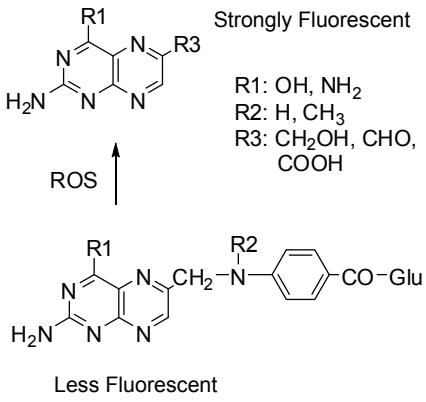


Figure 12: Strategy for the ROS detection using folic acid analogues.

scarcely fluoresces, PCA shows strong fluorescence in the visible-light region. The fluorescence quenching of folic acid is due to the aminobenzoyl moiety [48]. Folic acid was decomposed by H_2O_2 in the presence of Cu(II) and increased the fluorescence intensity in a dose-dependent manner. At pH 7.6, folic acid exhibited optimum stability [90], and the sensitivity was independent on pH in the range 6.0 ~ 8.0. The limit of detection (LOD; at S/N=3) for H_2O_2 was 0.5 μM under this experimental condition. The fluorescence intensity of folic acid was not increased by H_2O_2 without the metal ion, indicating that H_2O_2 itself does not decompose folic acid. Two scavengers of •OH, ethanol and mannitol, showed an inhibitory effect on the decomposition of folic acid but could not completely inhibit the decomposition. H_2O_2 and Cu(II) are considered to generate other reactive species, including copper-peroxo intermediates, which are less reactive than •OH [70]. Relevantly, a reactive species generated from H_2O_2 and Cu(II) cannot be completely quenched by typical •OH scavengers but effectively damages biomacromolecules, such as DNA [70, 91, 92].

Folic acid was slightly decomposed by H_2O_2 plus Fe(II). The decomposition of folic acid mediated by Fe(II) was completely inhibited by •OH scavengers, suggesting the involvement of •OH generated through the Fenton reaction. Since the lifetime of •OH is very short [93, 94], •OH cannot effectively decompose folic acid. Other metal ions, such as Ca(II), Mg(II), Fe(III), Co(II), Ni(II), Ag(I), Pd(II), and Au(III), did not mediate the decomposition of folic acid by H_2O_2. The

selectivity of metal ions is due to the activation of H_2O_2. The effect of fluorescence quenching by these heavy metal ions is negligible. Superoxide from KO_2 and sodium hypochlorite plus H_2O_2, a $^1O_2$ source, did not induce the fluorescence enhancement of folic acid. These findings demonstrate that folic acid can be selectively used for the fluorometry of H_2O_2 in the presence of Cu(II). The concentration of H_2O_2 ($[H_2O_2]$) can be determined by a calibration curve method.

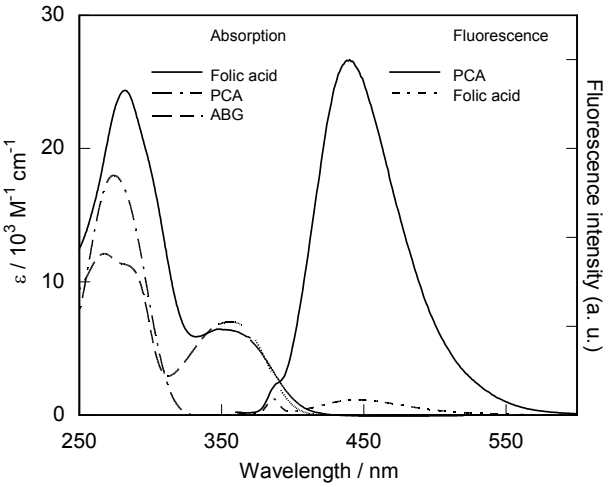


Figure 13: Absorption and fluorescence spectra of folic acid and related compounds. PCA: Pterin-6-carboxylc acid. ABG: Aminobenzoyl-*L*-glutamic acid.

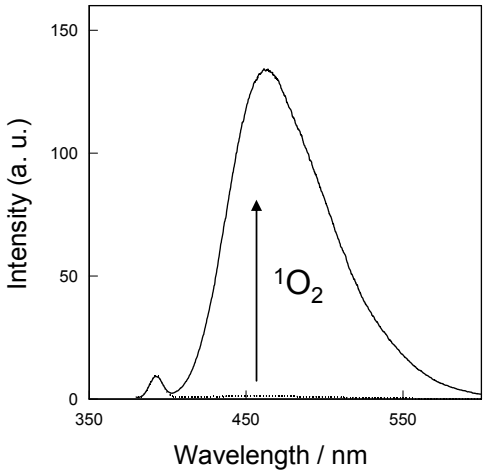


Figure 14: Fluorescence enhancement of methotrexate by $^1O_2$ generated through photosensitized reaction of methylene blue.

Experimental Results of Fluorometry of Singlet Oxygen Generated by Photosensitized Reaction Using Folic Acid Analogues

Singlet oxygen generated *via* photosensitized reaction decomposes less-fluorescent folic acid and MTX, leading to the formation of a strongly fluorescent pteridine compound in D_2O (Fig. **14**). This method can be applied to determine the Φ_Δ of photosensitizers absorbing visible light. One of the most important applications of the photosensitized reaction is PDT, in which a water-soluble photosensitizer absorbing visible light is necessary. The determination of Φ_Δ is required to evaluate the activity of the PDT photosensitizer. In addition, the value of Φ_Δ is important for the risk evaluation of phototoxic materials and various other studies about the photochemical oxidation process. In many cases, the determination of Φ_Δ of the water-soluble photosensitizer was carried out in D_2O, because the detection of $^1O_2$ is easier due to its longer lifetime than that of H_2O. The present method can be simply performed for these purposes without an expensive apparatus and reagent.

Summary of the Examination of Reactive Oxygen Probe

Reactive species generated from H_2O_2 and Cu(II) or $^1O_2$ generated through photosensitized reaction oxidized folic acid and MTX, leading to the cleavage of the C^9-N^{10} bond and the formation of strongly fluorescent pteridine. These findings demonstrate that folic acid analogues could be used for the fluorometry of ROS. The reactive species from H_2O_2 and Cu(II), such as copper-peroxo intermediates [70, 91, 92], are much more reactive than H_2O_2 and are considered to play important roles in the damage to biomacromolecules. Since the fluorescence quantum yield of pteridine, such as PCA, is markedly larger than that of folic acid analogues [48, 95], this analysis is sensitive to H_2O_2 detection. This method can be applied to the detection of small amounts of H_2O_2 generated from a carcinogenic compound. In addition, the Φ_Δ of the water-soluble photosensitizer could be determined using the folic analogues.

COMPUTATIONAL STUDY OF THE PHOTOCHEMICAL PROPERTY OF BERBERINE AND PALMATINE

Photosensitized generation of $^1O_2$ contributes to phototoxicity and photocarcinogenesis [70, 96-98]. Furthermore, this process is important in the

medical application of photosensitized reactions such as PDT [6-8]. Critical sites of the generated $^1O_2$ include mitochondria and lipid membranes [6-8, 65, 99]. Moreover, DNA is also an important target biomolecule of photosensitized reactions [63, 70, 72]. Since the administered photosensitizers necessarily interact with cellular components, the photosensitized reaction occurs in a microenvironment consisting of biomolecules [100]. Therefore, the interaction between biomolecules such as DNA and photosensitizers plays an important role in the photosensitized reaction and may be applied to the control of the activity of PDT photosensitizers [101]. Berberine and palmatine (Fig. **15**) are the alkaloid constituents of *Goldenseal* (*Hydrastis canadensis* L.) [102], and display cytotoxic activities against various human cancer cell lines [103]. Their phototoxicity and DNA-photodamaging abilities have been also reported [71, 104, 105]. These alkaloids bind to the DNA [106-109] and form the fluorescent intermolecular complexes [71, 110] with DNA. It has been reported that berberine binds preferentially to AT-rich minor groove [108], and the binding property of palmatine consistent with a mixed-mode DNA binding model in which a portion of the ligand molecule intercalates into the duplex, while the nonintercalated portion protrudes into the minor groove [107]. The DNA-binding interaction changes their photochemical property and markedly enhances the fluorescence intensity of these alkaloids. Their chemical property is useful in designing an experimental system to clarify the environmental effects of DNA, one of the most important biomaterials, on a photosensitized reaction. Moreover, the microenvironmental effect of the DNA strand should be one of the key factors in controlling the activity of photosensitizers, of which the target biomolecule is DNA. In this study, the photosensitized $^1O_2$-generation activity and the photochemical property of these alkaloids were examined.

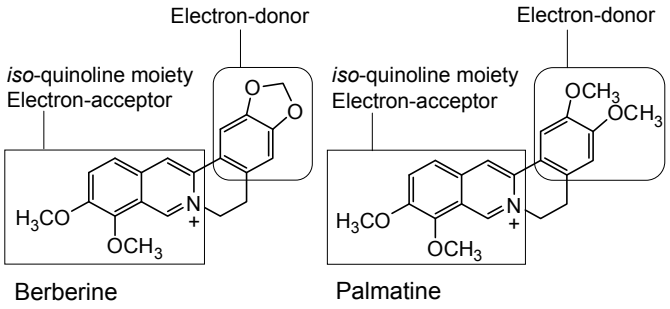


Figure 15: Structures of berberine and palmatine.

Calculations of Intermolecular Complex between DNA and the Alkaloids

The equilibrium geometry of an intermolecular complex between double-stranded DNA and berberine or palmatine was obtained by molecular mechanics calculation utilizing the Spartan 04'. The geometry of 20-mer of double-stranded DNA was constructed using Spartan 04'. The absorption transitions of these alkaloids binding to DNA were calculated by the semi-empirical Zerner's intermediate neglect of differential overlap (ZINDO) procedure utilizing the CAChe WorkSystem Pro 6.0 (Fujitsu Ltd. 2003, Tokyo, Japan). The energy of their orbital was estimated by DFT (B3LYP/6-31G*) calculation utilizing the Spartan 04'.

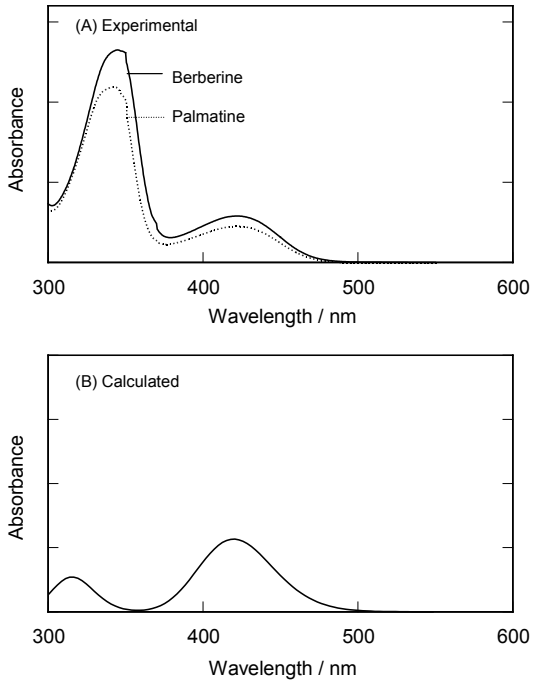


Figure 16: Experimental and calculated absorption spectra of berberine. The sample contained 50 μM berberine or palmatine in sodium phosphate buffer (pH 7.6) (A). The calculated spectrum of the isoquinoline compound (B).

Experimental and Computational Study of the Absorption Spectra of Berberine and Palmatine under an Interaction with DNA

The absorption maximum of these photosensitizers showed a blue shift depending on the solvent polarity, showing that this absorption is due to the n-π* transition.

The ZINDO calculation indicated that this absorption band is due to the excitation of the isoquinoline moiety of berberine and palmatine (Fig. **16**). The absorption spectra of DNA-binding photosensitizers were estimated from the subtraction of the absorption spectra of free photosensitizers from those of DNA containing photosensitizers (Fig. **17**). The binding ratio was calculated from the binding constants and the DNA concentration. The obtained spectra of these alkaloids showed a large red shift through the complex formation with DNA in an aqueous solution. The long wavelength absorption maxima (berberine: 449 nm, palmatine: 445 nm) were similar to those in dichloromethane. This large red shift could not be explained by the reduced polarity effect of DNA microenvironment, in which the polarity of surroundings is almost the same as that of ethanol [111]. ZINDO calculation of absorption spectra of the alkaloids interacting with phosphate anion of DNA showed a large red shift (Fig. **18**), suggesting that the spectral shifts by the DNA binding was due to the electrostatic interaction. Moreover, the ZINDO calculation showed that the structure change of these photosensitizers by the binding to DNA scarcely affected their spectra.

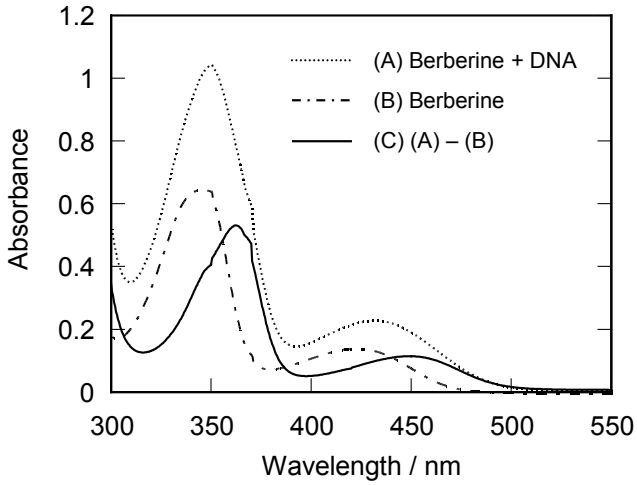


Figure 17: Absorption spectra of berberine and the complexes with DNA. The sample contained 50 μM berberine (A) in the presence of 200 bp-μM calf thymus DNA in sodium phosphate buffer (pH 7.6). The absorption spectrum of free berberine (B) in the above sample was calculated from the binding constant and concentrations of berberine and DNA. The absorption spectrum of the DNA-berberine complex was estimated from the subtraction of the spectrum of free berberine from that of DNA containing sample (C).

Computational Study of Direction of the Intramolecular Electron Transfer in the Photoexcited Berberine and Palmatine

Berberine and palmatine consist of isoquinoline moiety and dimethoxy benzene moiety. The absorption spectrum of the isoquinoline moiety was calculated from the ZINDO method (Fig. **16**). The comparison between the observed and calculated absorption spectra showed that the S_1 state of berberine and palmatine is excitation of the isoquinoline moiety. The DFT calculation showed that the HOMO of these alkaloids is localized on the dimethoxy benzene moiety. These calculation results suggest that S_1 of berberine and palmatine can be quenched through intramolecular electron transfer from their dimethoxybenzene moiety and the short photoexcited lifetimes of these alkaloids are due to this effect.

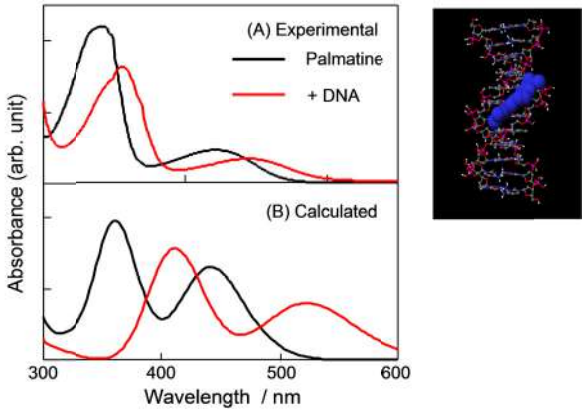


Figure 18: Experimental and calculated absorption spectra of palmatine. The sample contained 50 μM palmatine with or without 200 bp-μM calf thymus DNA in sodium phosphate buffer (pH 7.6) (A). Calculated absorption spectra of palmatine with or without DNA (B).

Control of Photosensitized Singlet Oxygen Generating Activity of Berberine and Palmatine through Interaction with DNA

Berberine and palmatine bound to DNA by an electrostatic interaction, leading to the formation of a fluorescent inter-molecular complex [71, 75, 110]. The thermodynamic study showed that the DNA binding is enthalpy driven. The activity of photosensitizers to generate $^1O_2$ is sensitive to their surroundings. The photosensitized $^1O_2$ generation from berberine and palmatine was markedly enhanced by the DNA-binding. The typical emission from $^1O_2$ at *ca*. 1270 nm was

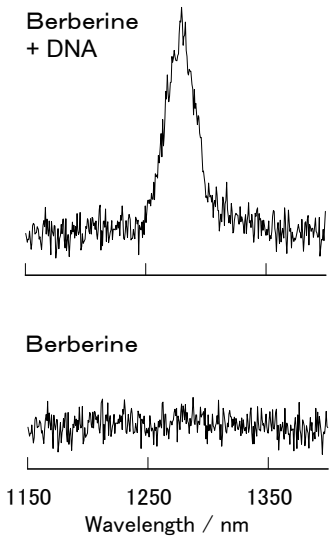


Figure 19: Near-infrared emission spectra from photo-irradiated berberine.

observed during photo-irradiation of the DNA-photosensitizer complexes. In the absence of DNA, the photo-irradiated berberines showed no emission around 1270 nm in aqueous solution. These findings have demonstrated that photoexcited berberines can generate $^1O_2$ only when the DNA-photosensitizer complex is formed (Fig. **19**). This study showed that the microenvironment of the DNA strand activates the $^1O_2$ generation of the photosensitizer. The S_1 state of these alkaloids immediately returns to the ground state *via* a nonradiative mechanism [104, 105]. This rapid deactivation should be the reason for the decreased $^1O_2$ generation by photosensitized berberine and palmatine in an aqueous solution (Fig. **20**). The measurement of the fluorescence decay demonstrated that the DNA-binding interaction stabilizes the photoexcited states of berberine and palmatine, resulting in the enhancement of their S_1 lifetimes. The ZINDO calculation showed that the S_1 state is produced by an excitation of their isoquinoline moieties. The DFT calculation showed that their HOMOs are localized on their dimethoxybenzene moieties. Therefore, the S_1 of berberine and palmatine can be quenched through intramolecular electron transfer from their dimethoxybenzene moieties. An inhibition of the intramolecular electron transfer can increase the lifetime of the photoexcited states of these photosensitizers. The spectral measurements suggest that the photochemical property of these

photosensitizers is mainly affected by the electrostatic interaction with anionic polymer DNA. The electrostatic interaction between the positive charge of the alkaloids and the negative charge of the phosphate group of DNA should increase the energy levels of the isoquinoline moieties of berberine and palmatine. The large red shift of their spectra of these alkaloids by the DNA binding is possibly due to the electrostatic interaction. Therefore, the levels of the CT states are raised through this interaction, leading to the inhibition of the intramolecular electron transfer and a prolongation of their photoexcited states. Consequently, the intersystem crossing yields of these alkaloids should be increased by the DNA-binding, resulting in the enhancement of the energy transfer to $^3O_2$. In the processes of intersystem crossing and $^1O_2$ generation, magnetic interactions including spin-orbit coupling is important [112, 113].

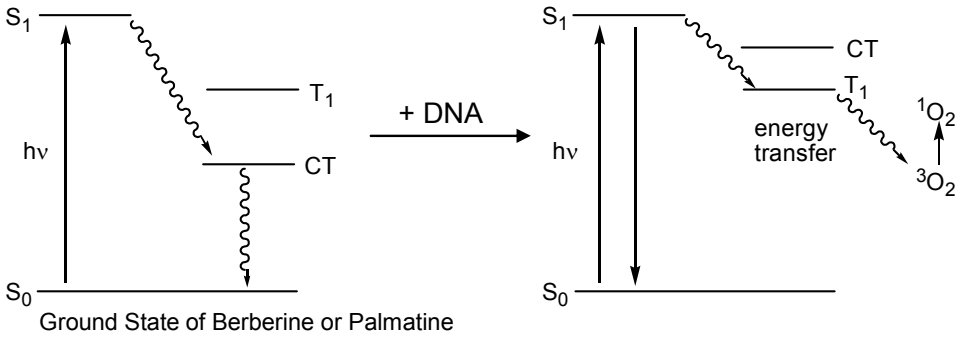


Figure 20: Mechanism of activity control of berberine and palmatine through interaction with DNA.

In general, other ROS, for example, $O_2^{\bullet-}$, H_2O_2, and •OH, can also contribute to DNA damage [69-73]. Further, the Type I mechanism is also important process of DNA photodamage [4, 5]. Photoexcited these alkaloids might generate $O_2^{\bullet-}$, H_2O_2, and •OH through following reactions [114, 115]:

$$^3\text{Sens}^* + O_2 \rightarrow \text{Sens}^{\bullet+} + O_2^{\bullet-} \quad (2)$$

$$O_2^{\bullet-} + H^+ \rightarrow HO_2 \quad (3)$$

$$HO_2 + H^+ + e^- \rightarrow H_2O_2 \quad (4)$$

$$H_2O_2 + H^+ + e^- \rightarrow H_2O + \bullet OH \quad (5)$$

where ^{3}Sens* is the triplet excited state of these alkaloids and Sens$^{\bullet+}$ is the radical cation of these alkaloids. The electron in the equations (4) and (5) can be provided by $O_2^{\bullet-}$ or endogenous metal ions, such as Fe(II) and Cu(I). In addition, DNA photodamage by berberine and palmatine through the Type I mechanism may be possible from the energetically point of view [116]. However, the patterns of DNA damage photosensitized by berberine and palmaine were quite similar to that by the Type II mechanism [71]. Photoexcited berberine and palmaine oxidized DNA at every guanine residue and other bases were not damaged. $O_2^{\bullet-}$ and H_2O_2 themselves cannot induce DNA damage without metal ions, and •OH damages DNA at every base [4, 5]. H_2O_2 induces the oxidation of thymine, cytosine, and guanine in the presence of copper ion [91, 92]. In the case of the Type I mechanism, consecutive guanines, such as underlined G of 5'-GG and 5'-GGG, are selectively damaged [4, 5, 63]. These results suggest that $^1O_2$ is the predominant reactive species responsible for DNA photodamage by these alkaloids.

Conclusion of the Controlled Generation of Singlet Oxygen by Berberine and Palmatine

This study demonstrated that berberine and palmatine bind to DNA, and their activity in the photosensitized $^1O_2$ generation is markedly enhanced. This study showed that the electrostatic interaction with a DNA strand can change the $^1O_2$ generation activity of photosensitizers. Singlet oxygen is the major oxidative and damaging species formed during the Type II process of photosensitization and plays an important role in the PDT process. It has been reported that $^1O_2$ is able to induce the oxidation of cellular DNA. $^1O_2$ can diffuse in a very short distance during its lifetime, which is much shorter in the cell (0.01 ~ 0.2 μs) than in simple aqueous solutions (2 ~ 4 μs) [65, 117]. Therefore, the contact of a photosensitizer with biomacromolecules, such as DNA, is very important. In addition, control of key therapeutic parameters, including $^1O_2$ generation level, is also important and optimized synthetic procedures of PDT photosensitizers have been developed [101, 118]. This mechanism through the interaction with DNA microenvironment may be applicable to the activity control of the PDT photosensitizers. Computational study played the important role in the speculation of the mechanism of the controlled generation of $^1O_2$ by berberine and palmatine.

MOLECULAR DESIGN OF PORPHYRIN PHOTOSENSITIZERS FOR CONTROL OF SINGLET OXYGEN GENERATION THROUGH INTERACTION WITH DNA

Computational chemistry is useful to design the photosensitizer for PDT. As an anti-cancer agent, DNA is one of the most important target biomacromolecules, and DNA-targeting drugs have been extensively studied [119, 120]. An important mechanism of PDT is the oxidation of biomacromolecules by $^1O_2$, which is generated through energy transfer from the excited photosensitizer to molecular oxygen. A DNA-selective photosensitizer should be developed to improve the treatment effect [101, 121-123]. The control of $^1O_2$ generation by a specific DNA sequence using photosensitizer/quencher/oligonucleotides systems has been studied [101, 121, 122]. The demonstrated principle is selectively placing the $^1O_2$ photosensitizer close to a molecule that can be quench the excited state of the photosensitizer by using a positioning system that can then be manipulated to change the distance between photosensitizer and the quencher. Furthermore, the pH regulated $^1O_2$ photosensitizer/quencher/DNA i-motif system was reported [123]. As mentioned above, berberine and palmatine, which can easily bind to DNA through electrostatic interaction and generate $^1O_2$ only when the DNA-photosensitizer complex is formed [71, 75]. The interaction changes their redox potentials and suppresses the quenching by intramolecular electron transfer, resulting in the elongation of the lifetime of the photoexcited state, making the energy transfer to molecular oxygen possible [75]. Berberine and palmatine can act as a DNA-targeting photosensitizer, and guanines are specifically oxidized through $^1O_2$ generation. However, these photosensitizers cannot absorb long-wavelength light, which is advantageous for PDT. Thus, on the basis of this controlling mechanism of $^1O_2$ generation, a porphyrinoid photosensitizer, which is important for clinical use because of its high absorptivity for the red region (> 630 nm) was designed and synthesized [124].

Molecular Orbital Calculation for Design of Photosensitizer

The equilibrium geometry of porphyrin and its MO energy were estimated from the *ab initio* MO calculation at the Hartree-Fock/6-31G* level utilizing the Spartan 08'. Fig. **21** shows the molecular structure and the MO of the synthesized

photosensitizer, An-TPyP, in which the anthracene moiety directly connects to the porphyrin ring as an electron donor. The MO calculation was performed at the Hartree-Fock 6-31G* level to predict the equilibrium geometry and the photophysical property of the porphyrinoid photosensitizer. The optimized structure of An-TPyP indicated the steric rotational hindrance of the anthracene moiety around the *meso*-position of the porphyrin, which keeps the two π-electronic systems nearly orthogonal to each other (Fig. **22**). The MO calculation showed that the HOMO of the protonated An-TPyP locates on the anthracene moiety (Fig. **21**), suggesting that the S_1 excitation of the porphyrin ring corresponds to the electron transition from the next HOMO to the LUMO. Consequently, the photoexcited state of the protonated form of An-TPyP should be deactivated *via* intramolecular electron transfer from the anthracene moiety to the porphyrin moiety, forming a CT state (Fig. **23**). The electrostatic interaction with anionic DNA and the hydrophobic environment of the DNA strand should raise the CT state energy, leading to the recovery of the photochemical activity, as in the case reported previously [71, 75]. From the molecular mechanics calculation and the CPK model, An-TPyP is considered to bind to the DNA major groove (Fig. **24**). A similar binding form of another cationic porphyrin has been reported [125].

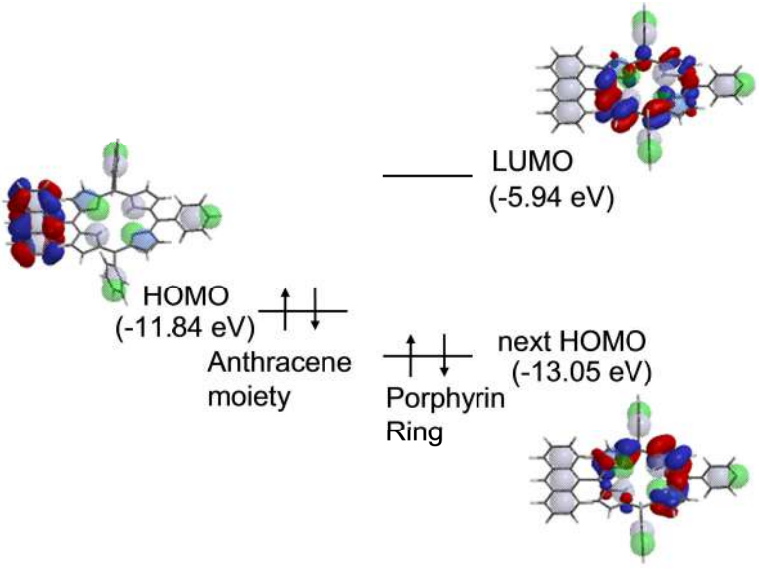


Figure 21: Designed and synthesized porphyrin.

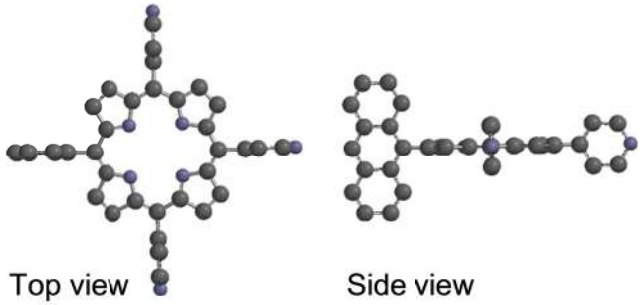


Figure 22: Calculated structure of An-TPyP.

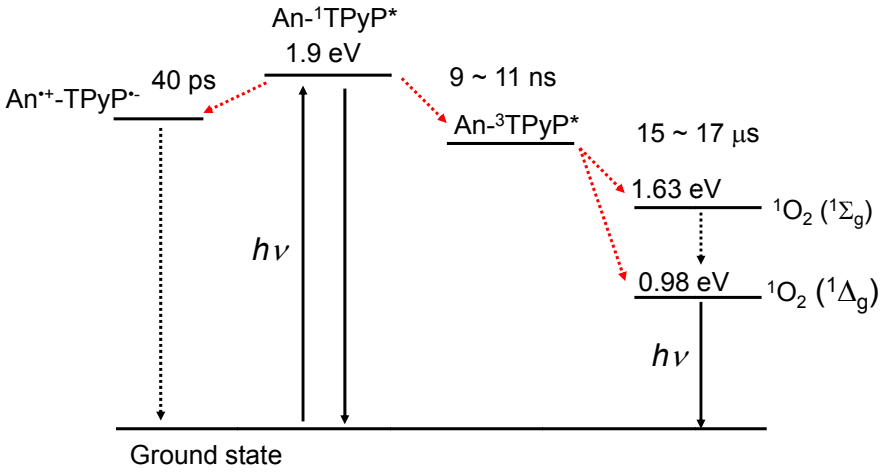


Figure 23: Mechanism of the activity control of An-TPyP under an interaction with DNA.

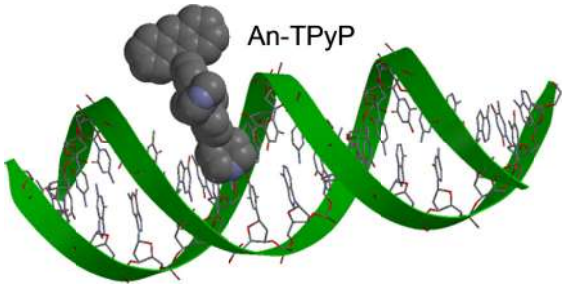


Figure 24: Calculated structure of An-TPyP binding to DNA.

Experimental Demonstration

The protonated An-TPyP showed almost no fluorescence in an aqueous solution, and the fluorescence lifetime (τ_f) was markedly short. It is notable, however, that, in

the presence of DNA, the fluorescence intensity increased and the lifetime became long. The extremely short lifetime for free An-TPyP stems from an effective quenching of the photoexcited state by the intramolecular electron transfer from the anthracene moiety to the porphyrin moiety to form the CT state. Owing to the positive charge in the protonated porphyrin moiety, An-TPyP can interact with anionic DNA. The interaction between An-TPyP and DNA was confirmed by the change in the UV-Vis absorption spectra. The electron-accepting ability of the porphyrin moiety should be decreased by the electrostatic interaction with anionic DNA, as in the case of previously reported photosensitizers [71, 75]. Furthermore, the hydrophobic environment of DNA is also unfavorable for the CT state. Thus, the intramolecular electron transfer process was suppressed, and the τ_f of the porphyrin ring elongated. To evaluate the $^1O_2$ generation activity of the photosensitizer, the near-infrared emission was measured. The near-infrared emission at around 1,270 nm, which is assigned to the radiative deactivation of $^1O_2$ to its ground state, was clearly observed during the photoexcitation of An-TPyP in the presence of DNA, whereas very weak emission was observed in the absence of DNA. The apparent value of Φ_f reached a plateau at 50 μM-bp, where all An-TPyP molecules interact with DNA. The of Φ_Δ with An-TPyP was estimated to be 0.75 and 0.52 for DNA-binding An-TPyP cases of AT-only sequence and guanine-containing sequence, respectively, in comparison with the $^1O_2$ emission intensity for methylene blue in D_2O (Φ_Δ= 0.52) [126].

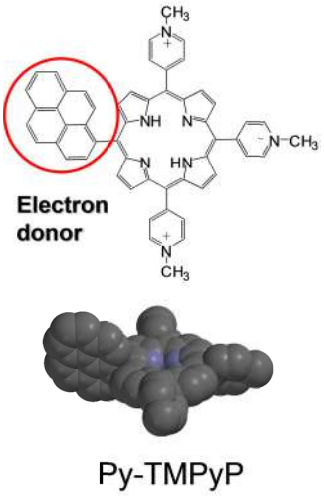


Figure 25: Design and synthesis of Py-TMPyP.

Design and Synthesis of Water-soluble Porphyrin for the Control of the Singlet Oxygen Generation by DNA

As predicted by the computational study, the singlet excited state of An-TPyP is effectively quenched through intramolecular electron transfer from the anthracene moiety [124]. The interaction with DNA suppresses the intramolecular electron transfer, resulting in the elongation of the lifetime of the photoexcited state. This elongation enhances the intersystem crossing and makes the photoenergy transfer to molecular oxygen possible. The activity control of porphyrinoid photosensitizers for PDT through an interaction with DNA should provide a possible increase in the selectivity for targeting DNA. Although An-TPyP can act at pH 2 ~ 3, which is not a normal physiological pH, this study demonstrated the activity control of an electron-donor connecting porphyrin by DNA. The next step is the design of porphyrinoid photosensitizer, whose activity can be controlled through interaction with DNA in water of physiological pH. Thus, the electron donor-connecting porphyrin, *meso*-(1-pyrenyl)-tris(*N*-methyl-*p*-pyridinio) porphyrin (Py-TMPyP, Fig. **25**) [127], was designed and synthesized. The MO calculation was performed at the Hartree-Fock 6-31G* level to predict the photophysical property of the porphyrinoid photosensitizer. This calculation showed that the photoexcited state of Py-TMPyP can be deactivated *via* intramolecular electron transfer from the pyrene moiety to the porphyrin moiety, forming a CT state. The interaction with DNA predicts a raise in the CT state energy, leading to the recovery of the photochemical activity.

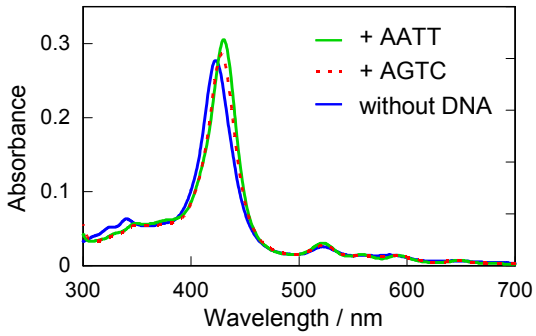


Figure 26: Absorption spectra of Py-TMPyP with or witout DNA.

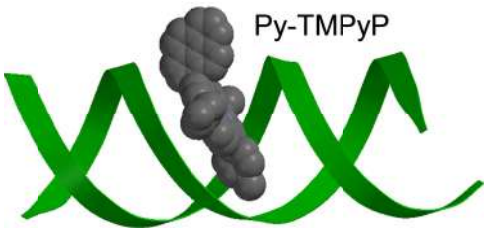


Figure 27: Calculated structure of Py-TMPyP binding to DNA.

Experimental Evaluation of the Activity of Synthesized Porphyrin

As the DNA, the synthesized 16-mer oligonucleotides (AATT: $d(AAAATTTTAAAATTTT)_2$ and AGTC: $d(AAGCTTTGCAAAGCTT)_2$ were used. The UV-Vis absorption spectrum of Py-TMPyP was red-shifted by the addition of DNA (Fig. **26**), indicating the binding interaction of Py-TMPyP into the DNA strand. From the molecular mechanics calculation and the CPK model, Py-TMPyP is considered to bind to the DNA major groove (Fig. **27**). The fluorescence spectra of Py-TMPyP are shown in Fig. **28**. Py-TMPyP showed almost no fluorescence in an aqueous solution. The extremely weak fluorescence for free Py-TMPyP stems from an effective quenching of the photoexcited state by the intramolecular electron transfer from the pyrene moiety to the porphyrin moiety to form the CT state. Note that, in the presence of DNA, the fluorescence intensity increased. The apparent values of Φ_f are 0.12 and 0.10 in the presence of 50 μM-base pair (μM-bp) AATT and AGTC, respectively. In this experimental condition, 98 and 97% of Py-TMPyP is binding with AATT and AGTC, respectively. The electron-accepting ability of the porphyrin moiety should be decreased by the electrostatic interaction with anionic DNA. Thus, the intramolecular electron transfer was suppressed, resulting in the enhancement of the fluorescence becouse of the elongation of the singlet excited state. With the GC-containing sequence of DNA, the Φ_f value of Py-TMPyP was slightly smaller than that of the AT-only sequence. Since guanine has the lowest oxidation potential in the nucleobases, an electron transfer from guanine to the photoexcited porphyrin ring possibly decreases the Φ_f of Py-TMPyP. To evaluate the $^1O_2$ generating activity of the photosensitizer, the near-infrared emission spectrum was measured. The typical near-infrared emission spectrum around at 1,270 nm, which is assigned to the deactivation of $^1O_2$ to its ground state, was clearly observed during the photoexcitation of Py-TMPyP with DNA, whereas the emission was not observed without DNA (Fig. **29**). The near-

infrared emission was effectively diminished by sodium azide, a physical quencher of $^1O_2$. The Φ_Δ was estimated from the comparison of the emission intensity by Py-TMPyP-DNA and that of methylene blue [126]. The apparent values of Φ_Δ by Py-TMPyP-DNA were 0.051 and 0.038 in the presence of 50 μM-bp AATT and AGTC, respectively. Thus, the photosensitized $^1O_2$ generation by Py-TMPyP became possible through the interaction with DNA. Because $^1O_2$ generation occurs near DNA, the generated $^1O_2$ should interact with the DNA strand. AT sequences quench $^1O_2$ through mainly a physical mechanism with the rate constant of 4.1×10^5 $M^{-1}s^{-1}$ [128], whereas guanine can quench $^1O_2$ through a chemical process (guanine oxidation) with a higher rate constant (1.7×10^7 $M^{-1}s^{-1}$) [129]. Therefore, the actual quantum yield of $^1O_2$ generation may be higher than the estimated values.

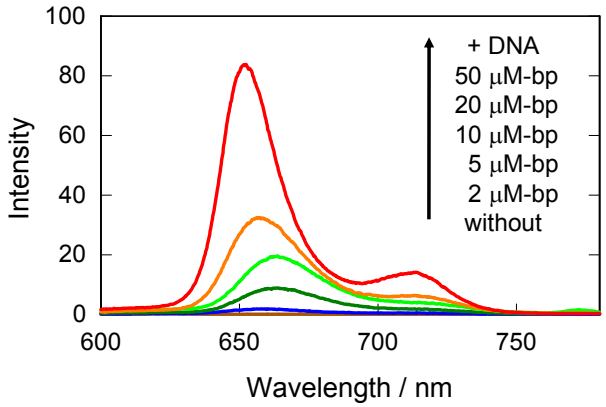


Figure 28: Fluorescence spectral change of Py-TMPyP under an interaction with DNA.

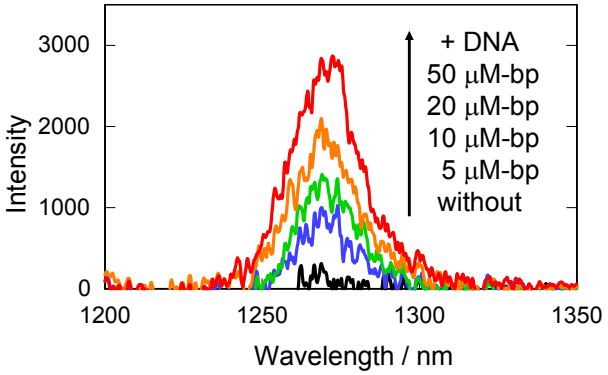


Figure 29: Near-infrared emission of singlet oxygen generated by the photosensitized reaction of Py-TMPyP under an interaction with DNA.

Conclusion of the Controlled Generation of Singlet Oxygen by DNA-binding Porphyrin Photosensitizer

In conclusion, the singlet excited states of Py-TMPyP and protonated An-TPyP are effectively quenched through intramolecular electron transfer from the pyrene or anthracene moieties (Figs. **23** and **30**). The interaction with DNA suppresses the intramolecular electron transfer, resulting in an increase in the fluorescence intensity. This suppression of the electron transfer quenching enhances the intersystem crossing and makes the photo-energy transfer to molecular oxygen possible. The activity control of porphyrin photosensitizers for PDT through interaction with DNA should provide a possible increase in the selectivity for targeting DNA. This study demonstrated the activity control of $^1O_2$ generation of a water-soluble porphyrin, Py-TMPyP, by DNA at a normal physiological pH. Computational chemistry was important tool to design the porphyrin photosensitizers.

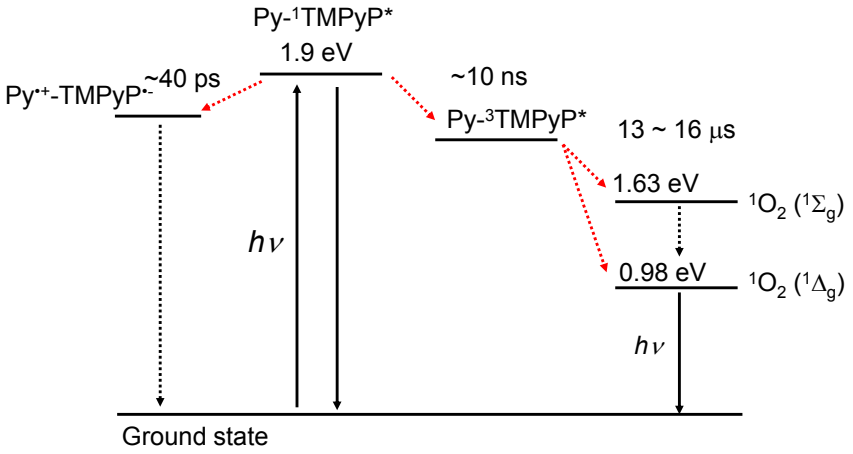


Figure 30: Proposed mechanism of the activity control of Py-TMPyP under an interaction with DNA.

STRUCTURE CALCULATION OF PHOSPHORUS(V) PORPHYRIN DERIVATIVES

The phosphorus(V) porphyrin (P(V) porphyrin) is a strong electron acceptor by the central cationic phosphorus atom. The derivatives of this porphyrin are used in the photochemical reaction. The structure of the porphyrin ring is strongly

affected by the coordination bond with the central phosphorus atom. The computational study of this structure is useful to investigate the characteristics. For example, the structure of dihydroxoP(V) tetraphenylporphyrin was determined by the MO calculation at the Hartree-Fock 6-31G* level (Fig. **31**). This porphyrin photosensitized DNA damage through $^1O_2$ generation and electron transfer [72]. This porphyrin is relatively planer, suggesting that intercalation to the DNA strand may be possible. The binding interaction with DNA is important for photosensitized damage because the damaging mechanisms of electron transfer and $^1O_2$ generation are sensitive to the distance between the photosensitizer and the nucleobase. Another example is the structure calculation of hydroxy(1-pyrenebutoxy)P(V) tetrakis(*p*-butoxyphenyl)porphyrin chloride (Fig. **32**) [130]. The energy transfer to the porphyrin competes with the electron transfer to the porphyrin in the photo-excited state of the pyrene of this porphyrin. The donor-acceptor distance is very important for these photophysical processes. The distance between the pyrene and porphyrin was easily predicted by the MO calculation. Computational study is useful in photochemical study.

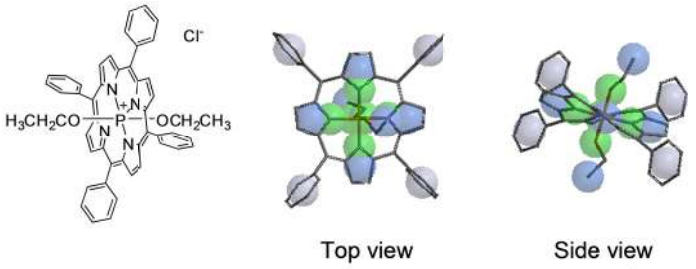


Figure 31: Example of calculated structure of phosphorus(V) tetraphenyl porphyrin.

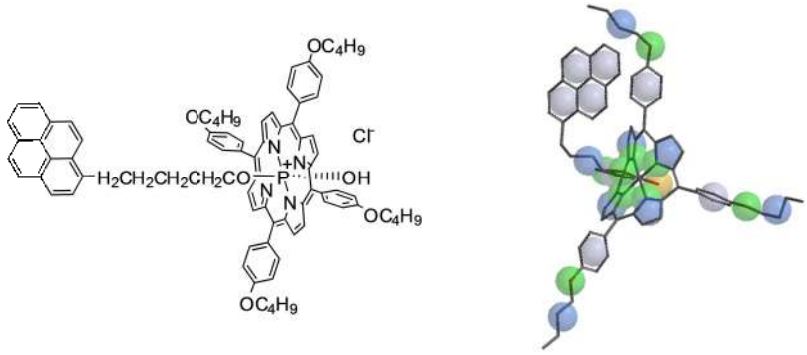


Figure 32: Structure of hydroxy(1-pyrenebutoxy)phosphorus(V) tetrakis(*p*-butoxyphenyl) porphyrin.

CONCLUDING REMARKS

A relatively simple calculation of HOMO of the photosensitizer was used to evaluate the DNA-damaging activity. Computational chemistry may be applied to predict the risk of the phototoxicity and photo-carcinogenicity of various compounds. On the other hand, the investigation of the photochemical reaction can be supported by the calculation of the molecular structure and energy. Therefore, the cost of experiments would be reduced by the use of computational study. Furthermore, computational chemistry is an important tool for the molecular design of drugs. In this chapter, examples for the design of a photosensitizer for PDT were introduced. Computational chemistry can be used in the fields of photochemistry, photobiology, and photomedicine.

ACKNOWLEDGEMENTS

The author thanks to Prof. Toru Hirano and Prof. Shigetoshi Okazaki (Hamamatsu University School of Medicine), Prof. Yoshio Nosaka (Nagaoka University of Technology), Prof. Shosuke Kawanishi (Suzuka University of Medical Science), Shinji Oikawa (Mie University), Prof. Yoshinobu Nishimura and Prof. Tatsuo Arai (University of Tsukuba), Prof. Hiroshi Segawa (The University of Tokyo), and Ms. Mari Harada (Shizuoka University) for their collaborations.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

DISCLOSURE

The presented works were partially supported by MEXT KAKENHI Grant Number 17029029 and JSPS KAKENHI Grant Number 20750131, 23750186 (Japanese Government).

REFERENCES

[1] International Agency for Research on Cancer (IARC) Working Group, Solar and ultraviolet radiation, in: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, IARC, Lyon, **1992**, Vol. 55, pp. 43-281.

[2] Drobetsky, E.A.; Turcotte, J.; Chateauneuf, A. A role for ultraviolet A in solar mutagenesis. *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 2350-2354.

[3] Besaratinia, A.; Synold, T.W.; Chen, H.H.; Chang, C.; Xi, B.; Riggs, A.D.; Pfeifer, G.P. DNA lesions induced by UV A1 and B radiation in human cells: comparative analyses in the overall genome and in the p53 tumor suppressor gene. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 10058-10063.

[4] Kawanishi, S.; Hiraku, Y.; Oikawa, S. Sequence-specific DNA damage induced by UVA radiation in the presence of endogenous and exogenous photosensitizers. *Curr. Probl. Dermatol.*, **2001**, *29*, 74-82.

[5] Hiraku, Y.; Ito, K.; Hirakawa, K.; Kawanishi, S. Photosensitized DNA damage and its protection *via* a novel mechanism. *Photochem. Photobiol.*, **2007**, *83*, 205-212.

[6] Dolmans, D.E.J.G.J.; Fukumura, D.; Jain, R.K. Photodynamic therapy for cancer. *Nat. Rev. Cancer*, **2003**, *3*, 380-387.

[7] Ackroyd, R.; Kelty, C.; Brown, N.; Reed, M. The history of photodetection and photodynamic therapy. *Photochem. Photobiol.*, **2001**, *74*, 656-669.

[8] Moan, J.; Peng, Q. An outline of the hundred-year history of PDT. *Anticancer Res.*, **2003**, *23*, 3591-3600.

[9] Ji, Z.; Yang, G.; Vasovic, V.; Cunderlikova, B.; Suo, Z.; Nesland, J.M.; Peng, Q. Subcellular localization pattern of protoporphyrin IX is an important determinant for its photodynamic efficiency of human carcinoma and normal cell lines. *J. Photochem. Photobiol. B: Biol.*, **2006**, *84*, 213-224.

[10] Ravanat, J.-L.; Sauvaigo, S.; Caillat, S.; Martinez, G.R.; Medeiros, M.H.; Di Mascio, P.; Favier, A.; Cadet, J. Singlet oxygen-mediated damage to cellular DNA determined by the commet assay associated with DNA repair enzymes. *Biol. Chem.*, **2004**, *385*, 17-20.

[11] Cadet, J.; Ravanat, J.-L.; Martinez, G.R.; Medeiros, M.H.; Di Mascio, P. Singlet oxygen oxidation of isolated and cellular DNA: product formation and mechanistic insights. *Photochem. Photobiol.*, **2006**, *82*, 219-225.

[12] Kawai, K.; Osakada, Y.; Fujitsuka, M.; Majima, T. Effects of reaction rate of radical anion of a photosensitizer with molecular oxygen on the photosensitized DNA damage. *Chem. Commun.*, **2006**, 3918-3920.

[13] Kawai, K.; Osakada, Y.; Fujitsuka, M.; Majima, T. Hole transfer in DNA and photosensitized DNA damage: importance of adenine oxidation. *J. Phys. Chem. B*, **2007**, *111*, 2322-2326.

[14] Ito, C.; Mishima, Y.; Litaudon, M.; Cosson, J.P.; Furukawa, H. Xanthone and dihydroisocoumarin from Montrouziera sphaeroidea. *Phytochemistry*, **2000**, *53*, 1043-1046.

[15] Ito, C.; Itoigawa, M.; Mishina, Y.; Filho, V.C.; Mukainaka, T.; Tokuda, H.; Nishino, H.; Furukawa, H. Chemical constituents of *Calophyllum brasiliensis*: structure elucidation of seven new xanthones and their cancer chemopreventive activity. *J. Nat Prod.*, **2002**, *65*, 267-272.

[16] Yoshimi, N.; Matsunaga, K.; Katayama, M.; Yamada, Y.; Kuno, T.; Qiao, Z.; Hara, A.; Yamahara, J.; Mori, H. The inhibitory effects of mangiferin, a naturally occurring glucosylxanthone, in bowel carcinogenesis of male F344 rats. *Cancer Lett.*, **2001**, *163*, 163-170.

[17] Weniger, B.; Um, B.H.; Valentin, A.; Estrada, A.; Lobstein, A.; Anton, R.; Maille, M.; Sauvain, M. Bioactive acridone alkaloids from *Swinglea glutinosa*. *J. Nat. Prod.*, **2001**, *64*, 1221-1223.

[18] Wu, T.S.; Chen, C.M. Acridone alkaloids from the root bark of *Severinia buxifolia* in Hainan. *Chem. Pharm. Bull.*, **2000**, *48*, 85-90.

[19] Stevenson, J.P.; DeMaria, D.; Reilly, D.; Purvis, J.D.; Graham, M.A.; Lockwood, G.; Drozd, M.; O'Dwyer, P.J. Phase I/pharmacokinetic trial of the novel thioxanthone SR233377 (WIN33377) on a 5-day schedule. *Cancer Chemother. Pharmacol.*, **1999**, *44*, 228-234.

[20] Sugiyama, H.; Saito, I. Theoretical studies of GG-specific photocleavage of DNA *via* electron transfer: significant lowering of ionization potential and 5'-localization of HOMO of stacked GG bases in B-form DNA. *J. Am. Chem. Soc.*, **1996**, *118*, 7063-7068.

[21] Maxam, A.M.; Gilbert, W. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.*, **1980**, *65*, 499-560.

[22] Nair, U.J.; Floyd, R.A.; Nair, J.; Bussachini, V.; Friesen, M.; Bartsch, H. Formation of reactive oxygen species and of 8-hydroxydeoxyguanosine in DNA *in vitro* with betel quid ingredients. *Chem. Biol. Interact.*, **1987**, *63*, 157-169.

[23] Ito, K.; Inoue, S.; Yamamoto, K.; Kawanishi, S. 8-Hydroxydeoxyguanosine formation at the 5' site of 5'-GG-3' sequences in double-stranded DNA by UV radiation with riboflavin. *J. Biol. Chem.*, **1993**, *268*, 13221-13227.

[24] Burrows, C.J.; Muller, J.G. Oxidative nucleobase modifications leading to strand scission. *Chem. Rev.*, **1998**, *98*, 1109-1151.

[25] Lewis, F.D.; Wu, Y. Dynamics of superexchange photoinduced electron transfer in duplex DNA. *J. Photochem. Photobiol. C: Photochem. Rev.*, **2001**, *2*, 1-16.

[26] Steenken, S.; Jovanovic, S. How easily oxidizable is DNA? One-electron reduction potentials of adenosine and guanosine radicals in aqueous solution. *J. Am. Chem. Soc.*, **1997**, *119*, 617-618.

[27] Yoshioka, Y.; Kitagawa, Y.; Takano, Y.; Yamaguchi, K.; Nakamura, T.; Saito, I. Experimental and theoretical studies on the selectivity of GGG triplets toward one-electron oxidation in B-form DNA. *J. Am. Chem. Soc.,* **1999**, *121*, 8712-8719.

[28] Ravanat, J.-L.; Douki, T.; Cadet, J. Direct and indirect effects of UV radiation on DNA and its components. *J. Photochem. Photobiol. B: Biol.*, **2001**, *63*, 88-102.

[29] O'Neill, P.; Parker, A.W.; Plumb, M.A.; Siebbeles, L.D.A. Guanine modifications following ionization of DNA occurs predominantly *via* intra- and not interstrand charge migration: An experimental and theoretical study. *J. Phys. Chem. B*, **2001**, *105*, 5283-5290.

[30] Hall, D.B.; Holmlin, R.E.; Barton, J.K. Oxidative DNA damage through long-range electron transfer. *Nature*, **1996**, *382*, 731-735.

[31] Nunez, M.E.; Hall, D.B.; Barton, J.K. Long-range oxidative damage to DNA: effects of distance and sequence. *Chem. Biol.*, **1999**, *6*, 85-97.

[32] Douki, T.; Angelov, D.; Cadet, J. UV laser photolysis of DNA: effect of duplex stability on charge-transfer efficiency. *J. Am. Chem. Soc.*, **2001**, *123*, 11360-11366.

[33] Kasai, H.; Yamaizumi, Z.; Berger, M.; Cadet, J. Photosensitized formation of 7,8-dihydro-8-oxo-2'-deoxyguanosine (8-hydroxy-2'-deoxyguanosine) in DNA by riboflavin: a non singlet oxygen mediated reaction. *J. Am. Chem. Soc.*, **1992**, *114*, 9692-9694.

[34] Bruner, S.D.; Norman, D.P.G.; Verdine, G.L. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature*, **2000**, *403*, 859-866.

[35] Shibutani, S.; Takeshita, M.; Grollman, A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature*, **1991**, *349*, 431-434.

[36] Cullis, P.M.; Malone, M.E.; Merson-Davies, L.A. Guanine radical cations are precursors of 7,8-dihydro-8-oxo-2'-deoxyguanosine but are not precursors of immediate strand breaks in DNA. *J. Am. Chem. Soc.*, **1996**, *118*, 2775-2781.

[37] Kino, K.; Saito. I.; Sugiyama, H. Product analysis of GG-specific photooxidation of DNA *via* electron transfer: 2-aminoimidazolone as a major guanine oxidation product. *J. Am. Chem. Soc.*, **1998**, *120*, 7373-7374.

[38] Kino, K.; Sugiyama, H. Possible cause of G-C-->C-G transversion mutation by guanine oxidation product, imidazolone. *Chem. Biol.*, **2001**, *8*, 369-378.

[39] Cadet, J.; Berger, M.; Buchko, G.W.; Joshi, P.C.; Raoul, S.; Ravanat, J.-L. 2,2-Diamino-4-[(3,5-di-O-acetyl-2-deoxy-β-D-erythro-pentofuranosyl)amino]-5-(2H)-oxazolone: a novel and predominant radical oxidation product of 3',5'-di-O-acetyl-2'-deoxyguanosine. *J. Am. Chem. Soc.*, **1994**, *116*, 7403-7404.

[40] Raoul, S.; Berger, M.; Buchko, G.W.; Joshi, P.C.; Morin, B.; Weinfeld, M.; Cadet, J. ^{1}H, ^{13}C and ^{15}N nuclear magnetic resonance analysis and chemical features of the two main radical oxidation products of 2'-deoxyguanosine: oxazolone and imidazolone nucleosides. *J. Chem. Soc. Perkin Trans. 2*, **1996**, 371-381.

[41] McBride, T.J.; Schneider, J. E.; Floyd, R.A.; Loeb, L.A. Mutations induced by methylene blue plus light in single-stranded M13mp2. *Proc. Natl. Acad. Sci. USA*, **1992**, *89*, 6866-6870.

[42] Negishi, K.; Hao, W. Spectrum of mutations in single-stranded DNA phage M13mp2 exposed to sunlight: predominance of G-to-C transversion. *Carcinogenesis*, **1992**, *13*, 1615-1618.

[43] Arimoto-Kobayashi, S.; Kaji, K.; Sweetman, G. M.; Hayatsu, H. Mutation and formation of methyl- and hydroxylguanine adducts in DNA caused by N-nitrosodimethylamine and N-nitrosodiethylamine with UVA irradiation. *Carcinogenesis*, **1997**, *18*, 2429-2433.

[44] Timpe, H.J.; Kronfeld, K.P.; Lammel, U; Fouassier, J.P.; Lougnot, D.J. Excited states of ketones as electron donors - ketone-iodonium salt system as photoinitiators for radical polymerization. *J. Photochem. Photobiol. A Chem.*, **1990**, *52*, 111-122.

[45] Klecak, G.; Urbach, F.; Urwyler, H. Fluoroquinolone antibacterials enhance UVA-induced skin tumors. *J. Photochem. Photobiol. B: Biol.*, **1997**, *37*, 174-181.

[46] Makinen, M.; Forbes, P.D.; Stenback, F. Quinolone antibacterials: a new class of photochemical carcinogens. *J. Photochem. Photobiol. B: Biol.*, **1997**, *37*, 182-187.

[47] Kawanishi, S.; Inoue, S.; Sano, S.; Aiba, H. Photodynamic guanine modification by hematoporphyrin is specific for single-stranded DNA with singlet oxygen as a mediator. *J. Biol. Chem.*, **1986**, *261*, 6090-6095.

[48] Hirakawa, K.; Suzuki, H.; Oikawa, S.; Kawanishi, S. Sequence-specific DNA damage induced by ultraviolet A-irradiated folic acid *via* its photolysis product. *Arch. Biochem. Biophys.*, **2003**, *410*, 261-268.

[49] Hirakawa, K.; Ochiai, S.; Oikawa, S.; Kawanishi, S. Oxygen-independent DNA damage photosensitized by rhodamine-6G. *Trends Photochem. Photobiol.*, **2011**, *13*, 29-35.

[50] Slaga, T.J. Nutrition and bio/technology in heart disease and cancer; Longenecker, J.B.; Kritchevski, D.; Drezner M.K. Ed.; Plenum Publishing Corp: New York, 1995; pp. 167-174.

[51] Yamashita, N.; Murata, M.; Inoue, S.; Burkitt, M.J.; Milne, L.; Kawanishi, S. α-Tocopherol induces oxidative damage to DNA in the presence of copper(II) ions. *Chem. Res. Toxicol.*, **1998**, *11*, 855-862.

[52] Murata, M.; Kawanishi, S. Oxidative DNA damage by vitamin A and its derivative *via* superoxide generation. *J. Biol. Chem.*, **2000**, *275*, 2003-2008.

[53] Nitta, Y.; Kamiya, K.; Tanimoto, M.; Kagimoto, O.; Niwa, O.; Yokoro, K. Effects of administration natural vitamin E on spontaneous hepatocarcinogenesis and N-nitrosoamine initiated tumors in mice. *J. Toxicol. Pathol.*, **1991**, *4*, 55-61.

[54] The alpha-tocopherol, beta carotene cancer prevention study group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N. Engl. J. Med.*, **1994**, *330*, 1029-1035.

[55] Omenn, G.S.; Goodman, G.E.; Thornquist, M.D.; Balmes, J.; Cullen, M.R.; Glass, A.; Keogh, J.P.; Meyskens, F.L.Jr.; Valanis, B.; Williams, J.H.Jr.; Barnhart, S.; Cherniack, M.G.; Brodkin, C.A.; Hammar, S. Risk factors for lung cancer and for intervention effects in CARET, the beta-carotene and retinol efficacy trial. *J. Natl. Cancer Inst.*, **1996**, *88*, 1550-1559.

[56] Menkovic, N.; Savikin-Fodulovic, K.; Bulatovic, V.; Aljancic, I.; Juranic, N.; Macura, S.; Vajs, V.; Milosavljevic, S. Xanthones from *Swertia punctata*. *Phytochemistry*, **2002**, *61*, 415-420.

[57] Ya, B.Q.; Nian, L.C.; Li, C.; Gen, X.P. Protective effect of swerchirin on hematopoiesis in ^{60}Co-irradiated mice. *Phytomedicine*, **1999**, *6*, 85-88.

[58] Bajpai, M.B.; Asthana, R.K.; Sharma, N.K.; Chatterjee, S.K.; Mukherjee, S.K. Hypoglycemic effect of swerchirin from the hexane fraction of *Swertia chirayita*. *Planta Med.*, **1991**, *57*, 102-104.

[59] Basnet, P.; Kadota, S.; Shimizu, M.; Takata, Y.; Kobayashi, M.; Namba, T. Bellidifolin stimulates glucose uptake in rat 1 fibroblasts and ameliorates hyperglycemia in streptozotocin (STZ)-induced diabetic rats. *Planta Med.*, **1995**, *61*, 402-405.

[60] Basnet, P.; Kadota, S.; Shimizu, M.; Namba, T. Bellidifolin: a potent hypoglycemic agent in streptozotocin (STZ)-induced diabetic rats from Swertia japonica. *Planta Med.*, **1994**, *60*, 507-511.

[61] Hirakawa, K.; Yoshida, M.; Nagatsu, A.; Mizukami, H.; Rana, V.; Rawat, M.S.M.; Oikawa, S.; Kawanishi, S. Chemopreventive action of xanthone derivatives on photosensitized DNA damage. *Photochem. Photobiol.*, **2005**, *81*, 314-319.

[62] Kawai, K.; Takada, T.; Tojo, S.; Majima, T. Kinetics of weak distance-dependent hole transfer in DNA by adenine-hopping mechanism. *J. Am. Chem. Soc.*, **2003**, *125*, 6842-6843.

[63] Hirakawa, K.; Yoshida, M.; Oikawa, S.; Kawanishi, S. Base oxidation at 5' site of GG sequence in double-stranded DNA induced by UVA in the presence of xanthone analogues: relationship between the DNA-damaging abilities of photosensitizers and their HOMO energies. *Photochem. Photobiol.*, **2003**, *77*, 349-355.

[64] Massad, W.; Criado, S.; Bertolotti, S.; Pajares, A.; Gianotti, J.; Escalada, J.P.; Amat-Guerri, F.; García, N.A. Photodegradation of the herbicide Norflurazon sensitised by Riboflavin. A kinetic and mechanistic study. *Chemosphere*, **2004**, *57*, 455-461.

[65] DeRosa, M.C.; Crutchley, R.J. Photosensitized singlet oxygen and its applications. *Coordination Chem. Rev.*, **2002**, *233-234*, 351-371.

[66] Schweitzer, C.; Schmidt, R. Physical mechanisms of generation and deactivation of singlet oxygen. *Chem. Rev.*, **2003**, *103*, 1685-1758.

[67] Greer, A. Christopher Foote's discovery of the role of singlet oxygen [$^1O_2(^1\Delta_g)$] in photosensitized oxidation reactions. *Acc. Chem. Res.*, **2006**, *39*, 797-804.

[68] Hatz, S.; Lambert, J.D.; Ogilby, P.R. Measuring the lifetime of singlet oxygen in a single cell: addressing the issue of cell *via*bility. *Photochem. Photobiol. Sci.*, **2007**, *10*, 1106-1116.

[69] Cadet, J.; Douki, T.; Gasparutto, D.; Ravanat, J.-L. Oxidative damage to DNA: formation, measurement and biochemical features. *Mutat. Res.*, **2003**, *531*, 5-23.

[70] Kawanishi, S.; Hiraku, Y.; Oikawa, S. Mechanism of guanine-specific DNA damage by oxidative stress and its role in carcinogenesis and aging. *Mutat. Res.*, **2001**, *488*, 65-76.

[71] Hirakawa, K.; Kawanishi, S.; Hirano, T. The mechanism of guanine specific photo-oxidation in the presence of berberine and palmatine: activation of photosensitized singlet oxygen generation through DNA-binding interaction. *Chem. Res. Toxicol.*, **2005**, *18*, 1545-1552.

[72] Hirakawa, K.; Kawanishi, S.; Hirano, T.; Segawa, H. Guanine-specific DNA oxidation photosensitized by the tetraphenylporphyrin phosphorus(V) complex *via* singlet oxygen generation and electron transfer. *J. Photochem. Photobiol. B: Biol.*, **2007**, *87*, 209-217.

[73] Nonell, S.; Braslavsky, S.E. Time-resolved singlet oxygen detection. *Methods Enzymol.*, **2000**, *319*, 37-49.

[74] Mikata, Y.; Takagi, S.; Tanahashi, M.; Ishii, S.; Obata, M.; Miyamoto, Y.; Wakita, K.; Nishisaka, T.; Hirano, T.; Ito, T.; Hoshino, M.; Ohtsuki, C,; Tanihara, M.; Yano, S. Detection of 1270 nm emission from singlet oxygen and photocytotoxic property of sugar-pendant [60] fullerenes. *Bioorg. Med. Chem. Lett.*, **2003**, *6*, 3289-3292.

[75] Hirakawa, K.; Hirano, T. The microenvironment of DNA switches the activity of singlet oxygen generation photosensitized by berberine and palmatine. *Photochem. Photobiol.*, **2008**, *84*, 202-208.

[76] Tanaka, K.; Miura, T.; Umezawa, N.; Urano, Y.; Kikuchi, K.; Higuchi, T.; Nagano, T. Rational design of fluorescein-based fluorescence probes. Mechanism-based design of a maximum fluorescence probe for singlet oxygen. *J. Am. Chem. Soc.*, **2001**, *123*, 2530-2536.

[77] Gomes, A.; Fernandes, E.; Lima, J.L.F.C. Fluorescence probes used for detection of reactive oxygen species. *J. Biochem. Biophys. Methods*, **2005**, *65*, 45-80.

[78] Costa, D.; Fernandes, E.; Santos, J.L.; Pinto, D.C.; Silva, A.M.; Lima, J.L.F.C. New noncellular fluorescence microplate screening assay for scavenging activity against singlet oxygen. *Anal. Bioanal. Chem.*, **2007**, *387*, 2071-2081.

[79] Soh, N. Recent advances in fluorescent probes for the detection of reactive oxygen species. *Anal. Bioanal. Chem.*, **2006**, *386*, 532-543.

[80] Hirakawa, K.; Aoshima, M.; Hiraku, Y.; Kawanishi, S. Photohydrolysis of methotrexate produces pteridine, which induces poly-G specific DNA damage through photo-induced electron transfer. *Photochem. Photobiol.*, **2002**, *76*, 467-472.

[81] Hirakawa, K. Fluorometry of singlet oxygen generated *via* a photosensitized reaction using folic acid and methotrexate. *Anal. Bioanal. Chem.*, **2009**, *393*, 999-1005.

[82] Akhtar, M.J.; Khan, M.A.; Ahmad, I. Photodegradation of folic acid in aqueous solution. *J. Pharm. Biomed. Anal.*, **1999**, *19*, 269-275.

[83] Akhtar, M.J.; Khan, M.A.; Ahmad, I. Effect of riboflavin on the photolysis of folic acid in aqueous solution. *J. Pharm. Biomed. Anal.*, **2000**, *23*, 1039-1044.

[84] Akhtar, M.J.; Khan, M.A.; Ahmad, I. Identification of photoproducts of folic acid and its degradation pathways in aqueous solution. *J. Pharm. Biomed. Anal.*, **2003**, *31*, 579-588.

[85] Hirakawa, K. Fluorometry of hydrogen peroxide using oxidative decomposition of folic acid. *Anal. Bioanal. Chem.*, **2006**, *386*, 244-248.

[86] Cabrerizo, F.M.; Dántol, M.L.; Petroselli, G.; Capparelli, A.L.; Thomas, A.H.; Braun, A.M.; Lorente, C.; Oliveros, E. Reactivity of conjugated and unconjugated pterins with singlet oxygen ($O_2(^1\Delta_g)$): physical quenching and chemical reaction. *Photochem. Photobiol.*, **2007**, *83*, 526-534.

[87] Chatterji, D.C.; Gallelli, J.F. Thermal and photolytic decomposition of methotrexate in aqueous solutions. *J. Pharm. Sci.*, **1978**, *67*, 526-531.

[88] Chahidi, C.; Morliere, P.; Aubailly, M.; Dubertret, L.; Santus, R. Photosensitization by methotrexate photoproducts. *Photochem. Photobiol.*, **1983**, *38*, 317-322.

[89] Chahidi, C.; Giraud, M.; Aubailly, M.; Valla, A.; Santus, R. 2,4-Diamino-6-pteridinecarboxaldehyde and an azobenzene derivative are produced by UV photodegradation of methotrexate. *Photochem. Photobiol.*, **1986**, *44*, 231-233.

[90] Akhtar, M.J.; Khan, M.A.; Ahmad, I. High performance liquid chromatographic determination of folic acid and its photodegradation products in the presence of riboflavin. *J. Pharm. Biomed. Anal.*, **1997**, *16*, 95-99.

[91] Hirakawa, K.; Mori, M.; Yoshida, M.; Oikawa, S.; Kawanishi, S. Photo-irradiated titanium dioxide catalyzes site specific DNA damage *via* generation of hydrogen peroxide. *Free Radic. Res.*, **2004**, *38*, 439-447.

[92] Hirakawa, K.; Oikawa, S.; Hiraku, Y.; Hirosawa, I.; Kawanishi, S. Catechol and hydroquinone have different redox properties responsible for their differential DNA-damaging ability. *Chem. Res. Toxicol.*, **2002**, *15*, 76-82.

[93] Land, E.J.; Ebert, M. Pulse radiolysis studies of aqueous phenol. Water elimination from dihydroxycyclohexadienyl radicals to form phenoxyl. *Trans. Faraday Soc.*, **1967**, *63*, 1181-1190.

[94] Liao, J.C.; Roider, J.; Jay, D.G. Chromophore-assisted laser inactivation of proteins is mediated by the photogeneration of free radicals. *Proc. Natl. Acad. Sci. USA*, **1994**, *91*, 2659-2663.

[95] Thomas, A.H.; Lorente, C.; Capparelli, A.L.; Pokhrel, M.R.; Braun, A.M.; Oliveros, E. Fluorescence of pterin, 6-formylpterin, 6-carboxypterin and folic acid in aqueous solution: pH effects. *Photochem. Photobiol. Sci.*, **2002**, *1*, 421-426.

[96] Cadet, J.; Ravanat, J.-L.; Martinez, G.R.; Medeiros, M.H.; Di Mascio, P. Singlet oxygen oxidation of isolated and cellular DNA: product formation and mechanistic insights. *Photochem. Photobiol.*, **2006**, *82*, 219-225.

[97] Besaratinia, A.; Bates, S.E.; Synold, T.W.; Pfeifer, G.P. Similar mutagenicity of photoactivated porphyrins and ultraviolet A radiation in mouse embryonic fibroblasts: involvement of oxidative DNA lesions in mutagenesis. *Biochemistry*, **2004**, *43*, 15557-15566.

[98] Nishigori, C. Cellular aspects of photocarcinogenesis. *Photochem. Photobiol. Sci.*, **2006**, *5*, 208-214.

[99] Ji, Z.; Yang, G.; Vasovic, V.; Cunderlikova, B.; Suo, Z.; Nesland, J.M.; Peng, Q. Subcellular localization pattern of protoporphyrin IX is an important determinant for its photodynamic efficiency of human carcinoma and normal cell lines. *J. Photochem. Photobiol. B: Biol.*, **2006**, *84*, 213-224.

[100] Lang, K.; Mosinger, J.: Wagneroviá, D.M.; Photophysical properties of porphyrinoid sensitizers non-covalently bound to host molecules; models for photodynamic therapy. *Coordination Chem. Rev.*, **2004**, *248*, 321-350.

[101] Cló, E.; Snyder, J.W.; Voigt, N.V.; Ogilby, P.R.; Gothelf, K.V. DNA-programmed control of photosensitized singlet oxygen production. *J. Am. Chem. Soc.*, **2006**, *128*, 4200-4201.

[102] Weber, H.A.; Zart, M.K.; Hodges, A.E.; White, K.D.; Barnes, S.M.; Moody, L.A.; Clark, A.P.; Harris, R.K.; Overstreet, J.D.; Smith, C.S. Method validation for determination of alkaloid content in goldenseal root powder. *J. AOAC Int.*, **2003**, *86*, 476-483.

[103] Iwasa, K.; Moriyasu, M.; Yamori, T.; Turuo, T.; Lee, D.U.; Wiegrebe, W. *In vitro* cytotoxicity of the protoberberine-type alkaloids. *J. Nat. Prod.*, **2001**, *64*, 896-898.

[104] Inbaraj, J.J.; Kukielczak, B.M.; Bilski, P.; Sandvik, S.L.; Chignell, C.F. Photochemistry and phototoxicity of alkaloids from Goldenseal (*Hydrastis Canadensis L.*). 1. Berberine. *Chem. Res. Toxicol.*, **2001**, *14*, 1529-1534.

[105] Inbaraj, J.J.; Kukielczak, B.M.; Bilski, P.; He, Y.Y.; Sik, R.H.; Chignell, C. F. Photochemistry and photocytotoxicity of alkaloids from Goldenseal (*Hydrastis canadensis L.*). 2. Palmatine, hydrastine, canadine, and hydrastinine. *Chem. Res. Toxicol.*, **2006**, *19*, 739-774.

[106] Taira, Z.; Matsumoto, M.; Ishida, S.; Ichikawa, T.; Sakiya, Y. Aggregation of DNA enhanced by the protoberberine alkaloids, coralyne and berberine. *Chem. Pharm. Bull.*, **1994**, *42*, 1556-1561.

[107] Pilch, D.S.; Yu, C.; Makhey, D.; LaVoie, E.J.; Srinivasan, A.R.; Olson, W.K.; Sauers, R.R.; Breslauer, K.J.; Geacintov, N.E.; Liu, L.F. Minor groove-directed and intercalative ligand-DNA interactions in the poisoning of human DNA topoisomerase I by protoberberine analogs. *Biochemistry*, **1997**, *36*, 12542-1253.

[108] Mazzini, S.; Bellucci, M.C.; Mondelli, R. Mode of binding of the cytotoxic alkaloids beberine with the double helix nucleotide d(AAGAATTCTT)(2). *Bioorg. Med. Chem.*, **2003**, *11*, 505-514.

[109] Chen, W.H.; Qin, Y.; Cai, Z.; Chan, C.L.; Luo, G.A.; Jiang, Z.H. Spectrometric studies of cytotoxic protoberberine alkaloids binding to double-stranded DNA. *Bioorg. Med. Chem.*, **2005**, *13*, 1859-1866.

[110] Yu, Y.; Long, C.Y.; Sun, S.Q.; Liu, J.P. Application of an alkaloid as a novel fluorescence probe in the determination of DNA. *Anal. Lett.*, **2001**, *34*, 2659-2669.

[111] Jin, R.; Breslauer, K.J. Characterization of the minor groove environment in a drug-DNA complex: bisbenzimide bound to the poly[d(AT)].poly[d(AT)]duplex. *Proc. Natl. Acad. Sci. USA*, **1988**, *85*, 8939-8942.

[112] Minaev, B.F.; Ågren, H. Spin-catalysis phenomena. *Int. J. Quantum Chem.*, **1996**, *57*, 519-532.

[113] Minaev, B.F.; Murugen, N.A.; Ågren, H. Dioxygen spectra and bioactivation. *Int. J. Quantum Chem.*, **2013**, *113*, 1847-1867.

[114] Prabhakar, R.; Siegbahn, P.E.M.; Minaev, B.F.; Ågren, H. Activation of triplet dioxygen by glucose oxidase: spin-orbit coupling in the superoxide ion. *J. Phys. Chem. B*, **2002**, *106*, 3742-3750.

[115] Prabhakar, R.; Siegbahn, P.E.M.; Minaev, B.F. A theoretical study of the dioxygen activation by glucose oxidase and copper amine oxidase. *Biochim. Biophys. Acta*, **2003**, *1647*, 173-178.

[116] Hirakawa, K.; Hirano, T.; Nishimura, Y.; Arai, T.; Nosaka, Y. Dynamics of singlet oxygen generation by DNA-binding photosensitizers. *J. Phys. Chem. B*, **2012**, *116*, 3037-3044.

[117] Niedre, M.; Patterson, M.S.; Wilson, B.C. Direct near-infrared luminescence detection of singlet oxygen generated by photodynamic therapy in cells *in vitro* and tissues *in vivo*. *Photochem. Photobiol.*, **2002**, *75*, 382-391.

[118] Gorman, A.; Killoran, J.; O'Shea, C.; Kenna, T.; Gallagher, W.M.; O'Shea, D.F. *In vitro* demonstration of the heavy-atom effect for photodynamic therapy. *J. Am. Chem. Soc.*, **2004**, *126*, 10619-10631.

[119] Hurley, L.H. DNA and its associated processes as targets for cancer therapy. *Nat. Rev. Cancer*, **2002**, *2*, 188-200.

[120] Chen, J.; Stubbe, J. Bleomycins: towards better therapeutics. *Nat. Rev. Cancer*, **2005**, *5*, 102-112.

[121] Cló, E.; Snyder, J.W.; Ogilby, P.R.; Gothelf, K.V. Control and selectivity of photosensitized singlet oxygen production: challenges in complex biological systems. *ChemBioChem*, **2007**, *8*, 475-481.

[122] Lovell, J.F.; Chen, J.; Huynh, E.; Jarvi, M.T.; Wilson, B.C.; Zheng, G. Facile synthesis of advanced photodynamic molecular beacon architectures. *Bioconjug. Chem.*, **2010**, *21*, 1023-1025.

[123] Tørring, T.; Toftegaard, R.; Arnbjerg, J.; Ogilby, P.R.; Gothelf, K.V. Reversible pH-regulated control of photosensitized singlet oxygen production using a DNA i-motif. *Angew. Chem. Int. Ed. Engl.*, **2010**, *18*, 7923-7925.

[124] Hirakawa, K.; Hirano, T.; Nishimura, Y.; Arai, T.; Nosaka, N. *Photochem. Photobiol.*, **2011**, *87*, 833-839.

[125] Jin, B.; Lee, H.M.; Lee, Y.-A.; Ko, J.H.; Kim, C.; Kim, S.K. Simltaneous binding of meso-tetrakis(N-methylpyridinium-4-yl)porphyrin and 4',6-diamino-2-phenylindole at the minor grooves of poly(dA)-poly(dT) and poly[d(A-T)$_2$]: fluorescence resonance energy transfer between DNA bound drugs. *J. Am. Chem. Soc.*, **2005**, *127*, 2417-2424.

[126] Usui, Y.; Kamogawa, K. A standard system to determine the quantum yield of singlet oxygen formation in aqueous solution. *Photochem. Photobiol.*, **1974**, *19*, 245-247.

[127] Hirakawa, K.; Harada, M.; Okazaki, S.; Nosaka, Y. Controlled generation of singlet oxygen by a water-soluble meso-pyrenylporphyrin photosensitizer through interaction with DNA. *Chem. Commun.*, **2012**, *48*, 4770-4772.

[128] Petroselli, G.; Dántola, M.L.; Cabrerizo, F.M.; Capparelli, A.L.; Lorente, C.; Oliveros, E.; Thomas, A.H. Oxidation of 2'-deoxyguanosine 5'-monophosphate photoinduced by pterin: Type I *versus* type II mechanism. *J. Am. Chem. Soc.*, **2008**, *130*, 3001-3011.

[129] Petroselli, G.; Erra-Balsells, R.; Cabrerizo, F.M.; Lorente, C.; Capparelli, A.L.; Braun, A.M.; Oliveros, E.; Thomas, A.H. Photosensitization of 2'-deoxyadenosine-5'-monophosphate by pterin. *Org. Biomol. Chem.*, **2007**, *5*, 2792-2799.

[130] Hirakawa, K.; Segawa, H. Acid dissociation of axial hydroxyl group of hydroxy(1-pyrenebutoxy)-phosphorus(V) porphyrin controls the intramolecular excitation energy transfer. *Photochem. Photobiol. Sci.*, **2010**, *9*, 704-709.

CHAPTER 3

How to Judge Predictive Quality of Classification and Regression Based QSAR Models?

Kunal Roy[1,2,*] and Supratik Kar[1]

[1]Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India and [2]Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, Great Britain

Abstract: Quantitative structure-activity relationship (QSAR) is a statistical modelling approach that can be used in drug discovery, environmental fate modeling, property and activity prediction of new, untested compounds. Validation has been identified as one of the important steps for checking the robustness and reliability of QSAR models. Various methodological aspects of validation of QSARs have been a subject of strong debate within the academic and regulatory communities. One of the principles (Principle 4) of the Organization for Economic Cooperation and Development (OECD) refers to the need to establish "*appropriate measures of goodness-of-fit, robustness and predictivity*" for any QSAR model. Validation strategies are recognized decisive steps to check the statistical acceptability and applicability of the constructed models on a new set of data in order to judge the confidence of predictions. Validation is a holistic practice that comprises evaluation of issues such as quality of data, applicability of the model for prediction purpose and mechanistic interpretation in addition to statistical judgment. Validation strategies are largely dependent on various validation metrics. Viewing the importance of QSAR validation approaches and different validation parameters in the development of successful and acceptable QSAR models, we herein focus to have an overview of different traditional as well as relatively new validation metrics used to judge the quality of the regression as well as classification based QSAR models.

Keywords: Applicability domain, OECD, QSAR, randomization, validation, virtual screening.

INTRODUCTION

Quantitative structure-activity relationships (QSARs) have a significant role in

***Corresponding author Kunal Roy:** Manchester Institute of Biotechnology, University of Manchester Manchester M1 7DN, Great Britain; Tel: +91-98315 94140; Fax: +91-33-2837 1078; E-mails: kunalroy_in@yahoo.com; kroy@pharma.jdvu.ac.in; kunal.roy@manchester.ac.uk;
URL: http://sites.google.com/site/kunalroyindia/

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

drug design, property prediction and environmental fate modeling of chemicals and pharmaceuticals [1-3]. Predictive QSAR models are also used by different regulatory agencies to assess physical, chemical, and biological properties of chemicals using applications precise for decision-making frameworks in chemical safety assessment [4]. In a broad perspective, QSAR models may be divided into two major categories, regression-based QSAR models and classification-based QSAR models. Another key application of a statistically important QSAR model is the development of focused libraries based on the features of 3D-pharmacophores and the attributes appearing in the QSAR models followed by subsequent virtual screening of libraries in search of compounds with enhanced potency [5]. New chemical entities with superior potency can also be obtained from application of the lead optimization technique using knowledge derived from QSAR models.

The most important purpose of QSAR modeling is to predict the activity/property/toxicity of new chemical entity (NCE) falling within the domain of applicability of the developed models. To check reliability of the QSAR models for their predictions is an important aspect for the applicability of the models and therein arises the importance of the validation process [6]. Validation has been recognized as one of the backbone steps for QSAR model development. Validation of QSAR models plays a vital step in the identification of predictive and robust models which may be utilized for future screening of new and/or untested molecules. This crucial step was ignored for a long time and the techniques of identification of statistical robustness of the model were only practised. However, only recently, a huge number of researches have been conducted for to the design of NCE with the exploitation of QSAR techniques where validation of the models has been considered as the most noteworthy step [7] for reviewing the quality of the input data as well as applicability and mechanistic interpretability of the constructed models. Various methodological aspects of validation of QSARs have been the subject of strong debate within the academic and regulatory communities. The following questions are often asked before successful validation and subsequent application of a QSAR model: 1) which of the validation principles should be followed to judge the quality as well as predictive power of the QSAR model?, 2) what are the major criteria for

establishing scientific validity of a QSAR model?, 3) how to use QSAR models for regulatory purposes?; and 4) is it possible to use any QSAR model for any given set of new untested chemicals?

The Organization for Economic Cooperation and Development (OECD) [8] has suggested five principles that should be followed to establish the scientific validity of a QSAR, thereby facilitating its acceptance for regulatory purposes. One of these principles (Principle 4) refers to the need to establish "*appropriate measures of goodness-of-fit, robustness and predictivity*" for any QSAR model. It identifies the need of the internal validation (as represented by goodness-of-fit and robustness) as well as the external validation of the QSAR model (predictivity). Validation strategies are recognized as the decisive steps to check the acceptability of the constructed models for their probable use on a new set of data, in order to judge the confidence of predictions. Basically, four strategies are adopted for validation of QSAR models [9]: i) internal or cross-validation; ii) division of the parent data set into training and test compounds; iii) application of the model on a (true) external data from a different source (true external validation); and iv) data randomization or Y-scrambling. The last method (randomization) can be regarded as a type of internal validation.

Validation strategies are largely dependent on various validation metrics. The statistical quality of the regression based and classification based QSAR models can be examined by different statistical metrics developed over the years [10]. Like various validation metrics, another important validation criterion is to check the chance correlation of the QSAR model by Y-randomization test. The randomization test is executed in order to guarantee the robustness of the QSAR model. The necessity to identify an applicability domain [11] (OECD Principle 3) arises due to the fact that QSARs are inescapably connected with restrictions in terms of the variation of chemical structures, properties and mechanisms of action for which the models can generate trustworthy predictions. The developed model can predict a new compound reliably only if the new compound lies in the applicability domain of the model. It is extremely useful that the QSAR model user has information about the applicability domain of the developed model to identify interpolation (true prediction) or extrapolation (less reliable prediction). Steps for development of reliable and acceptable QSAR model are demonstrated in Fig. **1.**

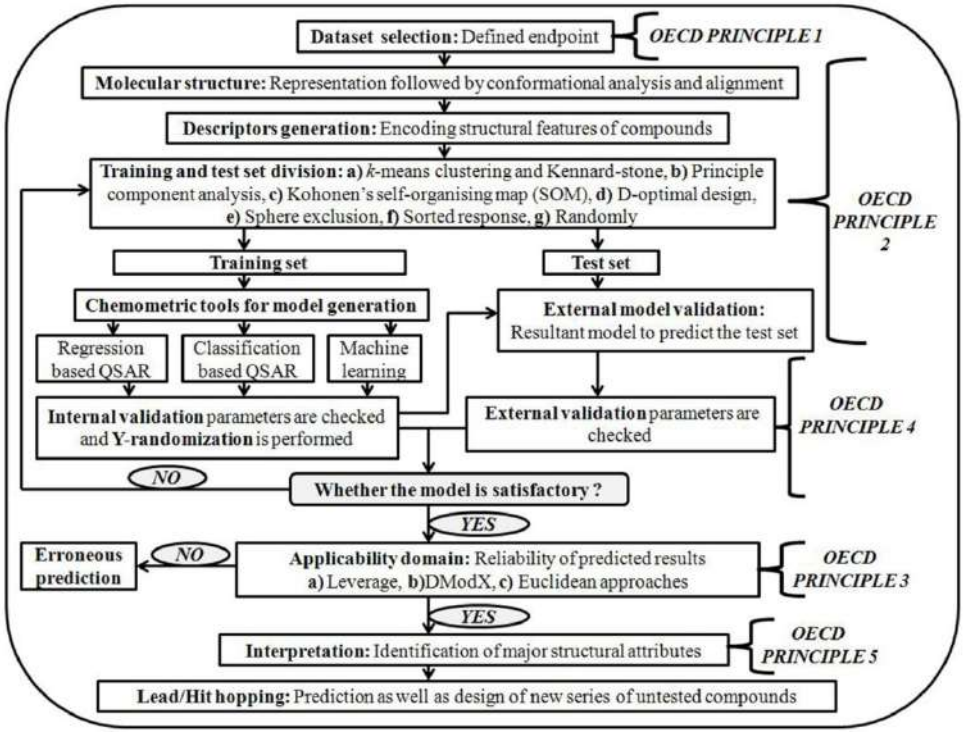


Figure 1: Steps for development of a reliable and acceptable QSAR model.

Viewing the importance of QSAR validation approaches and different validation parameters in the development of successful and acceptable QSAR models, we herein focus in overviewing of different traditional as well as relatively new validation metrics used to judge the quality of the QSAR models. This book chapter will help QSAR learners to have a bird's eye view on different available validation metrics useful for evaluating predictive quality of models.

IMPORTANCE OF VALIDATION OF A QSAR MODEL

With the advancement of cheminformatics, it is now possible to compute a large number of descriptors using various software tools [12]. Moreover, using various optimization procedures, it is now possible to obtain models that can fit well the experimental data but there may be a significant risk of overfitting. Fitting of data does not in any way confirm the prediction ability of a model. This is the main reason behind the requirement of validation of the developed models in terms of predictivity and robustness. A QSAR model is fundamentally judged in provisions

of its predictivity, representing how well it is capable to predict compounds which are not employed to develop the model. QSAR models that have been appropriately validated internally and externally can be considered trustworthy for both researchers and regulatory bodies [13, 14]. The meeting organized by QSAR experts formulated a set of guiding principles for the validation of QSAR models in Setúbal, Portugal in March 2002 [15]. The five guidelines adopted by the OECD [16] denoting validity of QSAR model are as follows: "*(i) a defined endpoint, (ii) an unambiguous algorithm, (iii) a defined domain of applicability, (iv) appropriate measures of goodness-of–fit, robustness and predictivity and (v) a mechanistic interpretation, if possible*". The mentioned guidelines are now considered as OECD Principles for the validation of QSAR models. The OECD has also offered a checklist to provide direction on the interpretation of the principles [17]. Thus, the existing challenge in the development of a QSAR model is not only make statistically sound and robust model to predict the activity within the domain of applicability constructed by the training set, but in developing a model with the ability to accurately predict the activity of untested chemicals [18].

Validation Strategies

The quality parameters quantify the fitness of a QSAR model as well as its robustness and predictive capabilities on a pure statistical basis. Apart from the use of the fitness parameters, validation of QSAR models consists of three strategies [9]: (i) internal validation utilizing the training set compounds, (ii) external validation employing the test set compounds, and (iii) true external validation by means of an external dataset from a different source. For evaluating the predictive ability of the developed models, both the internal and external validation methods have been considered by different groups of researchers. Randomization or Y-scrambling implemented on the data matrix provides a valuable technique for evaluating the existence of any chance correlation in the QSAR model. Along with these validation techniques, determination of the applicability domain of the model and selection of outliers are other vital aspects in the course of developing a reliable QSAR model with the spirit of OECD principles. As different metrics are used for regression based and classification based QSAR models, we have explained them in two different sections for the better understanding of the readers. The mathematical expressions of different validation metrics are summarized in Table **1**.

Table 1: Mathematical definitions of various statistical validation metrics for the classification and regression based QSAR models

Metrics Defining Statistical Quality of the Regression Based Models		
Sl. No.	**Mathematical definition**	
1	$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{calc})^2}{\sum (Y_{obs} - \overline{Y}_{training})^2}$	***Goodness-of-fit and quality measures***
2	$R_a^2 = \frac{\{(n-1) \times R^2\} - p}{n - p - 1}$	
3	$RMSE_c = \sqrt{\frac{\sum (Y_{obs(training)} - Y_{calc(training)})^2}{n_{training}}}$ $RMSE_p = \sqrt{\frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}}$	
4	$\rho = \frac{n_{training}}{p}$	
5	$Q^2 (Q^2_{LOO}) = 1 - \frac{\sum (Y_{obs(training)} - Y_{pred(training)})^2}{\sum (Y_{obs(training)} - \overline{Y}_{training})^2}$	***Internal parameters for robustness checking***
6	${}^c R_p^2 = R_{non\text{-}random} \times \sqrt{(R^2_{non\text{-}random} - R^2_{random})}$	
7	$Q^2_{(F1)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{training})^2}$	***External predictivity parameters/***
8	$Q^2_{(F2)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{test})^2}$	

Table 1: contd…

9	$Q^2_{(F3)} = 1 - \dfrac{\left[\sum\left(Y_{obs(test)} - Y_{pred(test)}\right)^2\right] / n_{test}}{\left[\sum\left(Y_{obs(train)} - \overline{Y}_{train}\right)^2\right] / n_{train}}$	
10	$CCC = \overline{\rho}_c = \dfrac{2\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{x} - \overline{y})}$	
11	**Golbraikh and Tropsha criteria** for external validation: i) $Q^2_{training} > 0.5$· ii) $R^2_{test} > 0.6$· iii) $\dfrac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \le k \le 1.15$ **or** $\dfrac{r^2 - r_0'^2}{r^2} < 0.1$ and $0.85 \le k' \le 1.15$. iv) $\left\|r_0^2 - r_0'^2\right\| < 0.3$ ·	
12	***r_m^2 metric*** $\overline{r_m^2} = \left(r_m^2 + r_m'^2\right)/2$ and $\Delta r_m^2 = \left\|r_m^2 - r_m'^2\right\|$, where $r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2})$ $r_m'^2 = r^2 \times (1 - \sqrt{(r^2 - r_0'^2)})$ The parameters r^2 and r_0^2 are defined as follows: $r_0^2 = 1 - \dfrac{\sum\left(Y_{obs} - k \times Y_{pred}\right)^2}{\sum\left(Y_{obs} - \overline{Y}_{obs}\right)^2}$ & $r_0'^2 = 1 - \dfrac{\sum\left(Y_{pred} - k' \times Y_{obs}\right)^2}{\sum\left(Y_{pred} - \overline{Y}_{pred}\right)^2}$ The terms k and k' are defined as:	***Scaled r_m^2 metrics for internal, external and overall predictivity***

Table 1: contd…

	$k = \frac{\sum\left(Y_{obs} \times Y_{pred}\right)}{\sum\left(Y_{pred}\right)^2}$ & $k' = \frac{\sum\left(Y_{obs} \times Y_{pred}\right)}{\sum\left(Y_{obs}\right)^2}$ The Y_{obs} and Y_{pred} values have been scaled at the beginning using the following formula: $Y_{i(scaled)} = \frac{Y_i - Y_{min(obs)}}{Y_{max(obs)} - Y_{min(obs)}}$	
13	$r_m^{2}{}_{(rank)} = r^2{}_{(rank)} \times \left(1 - \sqrt{r^2{}_{(rank)} - r_0^2{}_{(rank)}}\right)$	
Metrics defining statistical quality of the classification based QSAR models		
Sl. No.	**Mathematical definition**	
14	$\lambda = \det\left(\frac{W_g}{B_g + W_g}\right)$	***Goodness-of-fit and quality measures***
15	$R_c = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}$	
16	$\chi^2 = \sum_{i=1}^{t} \frac{(f_i - F_i)^2}{F_i}$	
17	$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$	
18	$d_{Mahalanobis}(x_i, x_j) = \sqrt{(x_i - x_j)^T \sum^{-1} (x_i - x_j)}$	
19	$\rho = \frac{n_{training}}{p}$	
20	$Sensitivity = \frac{TP}{TP + FN}$	***Internal and external validation metrics***

Table 1: contd…

21	$Specificity = \frac{TN}{TN + FP}$	***and parameters for ROC analysis***
22	$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$	
23	$Precision = \frac{TP}{TP + FP}$	
24	$F\text{ - }measure = \frac{2}{1/Precision + 1/Sensitivity}$	
25	$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$	
26	$G\text{-}means = \sqrt{Sensitivity \times Specificity}$	
27	$Cohen's\ \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$ $P_r(a) = \frac{(TP + TN)}{(TP + FP + FN + TN)}$ $P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2}$	
28	$AUC\text{-}ROC = 1 - \frac{\sum_{i=1}^{n} r_i}{n \times (N - n)} + \frac{n + 1}{2 \times (N - n)}$	
29	$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{1}{N}\left[\frac{1 - e^{-\alpha}}{e^{\frac{\alpha}{N}} - 1}\right]}$	

Table 1: contd…

30	$BEDROC = RIE \times \frac{\frac{1}{N}\sinh\left(\alpha/2\right)}{\cosh\left(\alpha/2\right)-\cosh\left(\alpha/2-\alpha\frac{n}{N}\right)} + \frac{1}{1-e^{\alpha\left(\frac{N-n}{N}\right)}}$	
31	$AUC\text{-}pROC = -\frac{1}{n}\times\sum_{i=1}^{n}\log_{10}\left(\frac{r_i - i}{N-n}\right)$	
32	$ROCED = \left(\left\|d_{training} - d_{test}\right| + 1\right)\times\left(d_{training} + d_{test}\right)\times\left(d_{test} + 1\right)$	
33	$FIT(\lambda) = \frac{(1-\lambda)\times(n-p-1)}{(n+p^2)\times\lambda}$	
34	$ROCFIT = \frac{ROCED}{FIT(\lambda)}$	
35	$E_{toxicity} = \frac{\%\ of\ \ toxic}{(\%\ of\ \ non\text{-}toxic + 100)}$	***Metrics for PDD analysis***
36	$E_{non-toxicity} = \frac{\%\ of\ \ non\text{-}toxic}{(\%\ of\ \ toxic + 100)}$	

*Each notation is mentioned in the text.

In Table **2**, we have demonstrated how one can easily calculate different metrics from a regression based QSAR model. In Table **3**, the methods of computation of different validation metrics for a classification based QSAR model are demonstrated.

Table 2: A simple example demonstrating calculation of different validation metrics from a regression based QSAR model*

Model equation (MLR)	$Y = 5.16 - 0.159 \times X_1 - 1.44 \times X_2$ $N_{training} = 11$, $N_{test} = 6$						
Training set							
Compound ID	***Y (Observed)***	X_1^*	X_2^*	***Y (Calculated/LOO predicted)***	$(Y_{obs(train)} - Y_{calc(train)})^2$	$(Y_{obs(train)} - \overline{Y}_{training})^2$	$(Y_{obs(train)} - Y_{LOO-pred(train)})^2$
1	3.45	1.65	1	3.46/3.44	0.000	0.624	0.000
3	3.14	1.63	1.5	2.74/3.59	0.159	0.230	0.201
5	2.82	2.02	1.5	2.68/3.01	0.020	0.026	0.035
7	2.64	2.05	1.5	2.67/2.60	0.001	0.000	0.001
9	2.29	1.83	1.5	2.711.82	0.176	0.137	0.218
10	2.2	1.9	2	1.98/2.56	0.049	0.212	0.132
11	2.15	1.33	2	2.07/2.29	0.007	0.260	0.019
13	1.74	1.54	2	2.04/1.35	0.087	0.846	0.153
15	3.45	1.91	1	3.42/3.51	0.001	0.624	0.003
16	2.51	1.66	1.5	2.74/2.27	0.051	0.023	0.060
17	2.87	1.1	1.5	2.83/2.98	0.002	0.044	0.012
	$\overline{Y}_{training} = 2.66$				$\sum (Y_{obs(train)} - Y_{calc(train)})^2 = 0.553$	$\sum (Y_{obs(train)} - \overline{Y}_{training})^2 = 3.026$	$\sum (Y_{obs(train)} - Y_{LOO-pred(train)})^2 = 0.834$
Test set							
Compound ID	***Y (Observed)***	X_1^*	X_2^*	***Y (Predicted)***	$(Y_{obs(test)} - Y_{pred(test)})^2$	$(Y_{obs(test)} - \overline{Y}_{training})^2$	$(Y_{obs(test)} - \overline{Y}_{test})^2$
2	3.2	1.9	1	3.42	0.047	0.292	0.144
4	2.87	1.22	1.5	2.81	0.004	0.044	0.003
6	2.81	1.78	1.5	2.72	0.009	0.023	0.000
8	2.49	1.61	1.5	2.74	0.065	0.029	0.109

Table 2: contd…

12	2.01	1.96	2	1.97	0.002	0.423	0.656
14	3.51	1.88	1	3.42	0.008	0.722	0.476
	$\overline{Y}test$ = 2.82				$\sum(Y_{obs(test)} - Y_{pred(test)})^2$ = 0.134	$\sum(Y_{obs(test)} - \overline{Y}_{training})^2$ =1.532	$\sum(Y_{obs(test)} - \overline{Y}_{test})^2$ = 1.388
Calculation of metrics							
S	$s = \sqrt{\frac{\sum(Y_{obs} - Y_{calc(train)})^2}{n - p - 1}} = \sqrt{\frac{0.553}{11 - 2 - 1}} = 0.26$						
R^2	$R^2 = 1 - \frac{\sum(Y_{obs(train)} - Y_{calc(train)})^2}{\sum(Y_{obs(train)} - \overline{Y}_{training})^2} = 1 - \frac{0.553}{3.026} = 0.817$						
R_a^2	$R_a^2 = \frac{(n-1)R^2 - p}{n-p-1} = \frac{(11-1)\times 0.817 - 2}{11-2-1} = 0.771$						
$RMSE_c$	$RMSE_c = \sqrt{\frac{\sum(Y_{obs(training)} - Y_{calc(training)})^2}{n_{training}}} = \sqrt{\frac{0.553}{11}} = 0.224$						
$RMSE_p$	$RMSE_p = \sqrt{\frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{n_{test}}} = \sqrt{\frac{0.134}{6}} = 0.177$						
Q^2	$Q^2 = 1 - \frac{\sum(Y_{obs(train)} - Y_{pred(train)})^2}{\sum(Y_{obs(train)} - \overline{Y}_{training})^2} = 1 - \frac{0.834}{3.026} = 0.632$						
R^2_{pred}	$R^2_{pred} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y}_{training})^2} = 1 - \frac{0.134}{1.532} = 0.912$						
$Q^2_{(F2)}$	$Q^2_{(F2)} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \overline{Y}_{test})^2} = 1 - \frac{0.134}{1.388} = 0.903$						
$Q^2_{(F3)}$	$Q^2_{(F3)} = 1 - \frac{[\sum(Y_{obs(test)} - Y_{pred(test)})^2]/n_{ext}}{[\sum(Y_{obs(train)} - \overline{Y}_{train})^2]/n_{tr}} = 1 - \frac{0.134/6}{3.026/11} = 0.919$						

* Y is the response variable; X_1 and X_2 are the descriptors involved in the QSAR model

Table 3: A simple example demonstrating calculation of different validation metrics from a classification based QSAR model*

Compound ID	Y (Observed)	Classification threshold based on *Y* (Observed): (L≤3.69<H)	Posterior Probabilities (PP)	Predicted classification threshold based on PP: (L<0.50<H)
Training set compounds				
1	4.17	H	0.94	H
3	4.11	H	0.95	H
4	3.58	L	0.97	H
8	3.82	H	0.63	H
9	3.63	L	0.11	L
10	3.89	H	0.87	H
11	3.82	H	0.81	H
12	3.55	L	0.10	L
13	3.98	H	0.83	H
14	3.5	L	0.01	L
15	3.78	H	0.87	H
16	4.44	H	0.78	H
17	3.36	L	0.01	L
18	3.91	H	0.97	H
20	3.69	L	0.88	H
22	3.55	L	0.60	H
23	3.64	L	0.01	L
25	3.94	H	0.63	H
26	3.94	H	0.94	H
29	2.77	L	0.02	L
30	2.41	L	0.00	L
31	3.38	L	0.10	L
33	3.29	L	0.01	L

Table 3: contd…

<table>
<tr><td>34</td><td>4.24</td><td>H</td><td>0.98</td><td>H</td></tr>
<tr><td>35</td><td>4.1</td><td>H</td><td>0.99</td><td>H</td></tr>
<tr><td colspan="5">Test set compounds</td></tr>
<tr><td>2</td><td>4.08</td><td>H</td><td>0.60</td><td>H</td></tr>
<tr><td>5</td><td>3.51</td><td>L</td><td>0.06</td><td>L</td></tr>
<tr><td>6</td><td>3.83</td><td>H</td><td>0.53</td><td>H</td></tr>
<tr><td>7</td><td>3.59</td><td>L</td><td>0.94</td><td>H</td></tr>
<tr><td>19</td><td>3.93</td><td>H</td><td>0.87</td><td>H</td></tr>
<tr><td>21</td><td>3.63</td><td>L</td><td>0.98</td><td>H</td></tr>
<tr><td>24</td><td>4.03</td><td>H</td><td>0.88</td><td>H</td></tr>
<tr><td>27</td><td>3.36</td><td>L</td><td>0.00</td><td>L</td></tr>
<tr><td>28</td><td>3.95</td><td>H</td><td>0.49</td><td>L</td></tr>
<tr><td>32</td><td>3.23</td><td>L</td><td>0.00</td><td>L</td></tr>
<tr><td>Classification metrics</td><td colspan="2">Training set
<table><tr><td>Confusion matrix</td><td>P</td><td>N</td></tr><tr><td>P</td><td>13</td><td>0</td></tr><tr><td>N</td><td>3</td><td>9</td></tr></table></td><td colspan="2">Test set
<table><tr><td>Confusion matrix</td><td>P</td><td>N</td></tr><tr><td>P</td><td>4</td><td>1</td></tr><tr><td>N</td><td>2</td><td>3</td></tr></table></td></tr>
<tr><td>TP</td><td colspan="2">13</td><td colspan="2">4</td></tr>
<tr><td>FN</td><td colspan="2">0</td><td colspan="2">1</td></tr>
<tr><td>FP</td><td colspan="2">3</td><td colspan="2">2</td></tr>
<tr><td>TN</td><td colspan="2">9</td><td colspan="2">3</td></tr>
<tr><td>Sensitivity (%)
$Sensitivity = \frac{TP}{TP + FN}$</td><td colspan="2">$\left(\frac{13}{13+0}\right)*100 = 100$</td><td colspan="2">$\left(\frac{4}{4+1}\right)*100 = 80$</td></tr>
</table>

Table 3: contd…

Specificity (%) $Specificity = \frac{TN}{TN + FP}$	$\left(\frac{9}{9+3}\right)*100 = 75$	$\left(\frac{3}{3+2}\right)*100 = 60$
Precision (%) $Precision = \frac{TP}{TP + FP}$	$\left(\frac{13}{13+3}\right)*100 = 81.25$	$\left(\frac{4}{4+2}\right)*100 = 66.67$
Accuracy (%) $Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$	$\left(\frac{13+9}{13+0+3+9}\right)*100 = 88$	$\left(\frac{4+3}{4+1+2+3}\right)*100 = 70$
F-measure (%) $F\text{-}measure = \frac{2}{1/Precision + 1/Sensitivity}$	$\left(\frac{2}{1/0.81 + 1/1}\right)*100 = 89.66$	$\left(\frac{2}{1/0.667 + 1/0.80}\right)*100 = 72.72$
G-means $G\text{-}means = \sqrt{Sensitivity \times Specificity}$	$\sqrt{(1 \times 0.75)} = 0.87$	$\sqrt{(0.93 \times 0.89)} = 0.69$
Cohen's κ $P_r(a) = \frac{(TP + TN)}{(TP + FP + FN + TN)}$ $P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2}$ $Cohen's\ \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$	$P_r(a) = \left(\frac{13+9}{13+3+0+9}\right) = 0.88$ $P_r(e) = \frac{\langle (13+3) \times (13+0) \rangle + \langle (9+3) \times (9+0) \rangle}{(13+0+3+9)^2}$ $= 0.5056$ $Cohen's\ \kappa = \frac{0.88 - 0.5056}{1 - 0.5056} = 0.76$	$P_r(a) = \left(\frac{4+3}{4+2+1+3}\right) = 0.70$ $P_r(e) = \frac{\langle (4+2) \times (4+1) \rangle + \langle (3+2) \times (3+1) \rangle}{(4+1+2+3)^2}$ $= 0.50$ $Cohen's\ \kappa = \frac{0.821 - 0.179}{1 - 0.179} = 0.40$
MCC $MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$	$\left(\frac{(13 \times 9) - (3 \times 0)}{\sqrt{(13+3) \times (13+0) \times (9+3) \times (9+0)}}\right) = 0.781$	$\left(\frac{(4 \times 3) - (2 \times 1)}{\sqrt{(4+2) \times (4+1) \times (3+2) \times (3+1)}}\right) = 0.41$

*Y is the response variable

VALIDATION METRICS FOR REGRESSION BASED QSAR MODELS

MEASUREMENTS OF QUALITY OF QSAR MODELS

A QSAR model is required to be checked for various quality measures before it can be applied for screening of new chemical entities. For assessing the quality of a QSAR model, various statistical parameters can be used. The acceptability of a regression based QSAR model relies upon different statistical parameters [10] such as (i) standard error of estimate (***s***), (ii) determination coefficient (R^2) and (iii) explained variance (R_a^2). The error in the inference of individual activity values of the molecules under study utilizing the regression method can be quantified based on their residual data. The standard error of estimate (*SEE* or *s*) is calculated from the square root of sum of squares of the residuals divided by the degree of freedom. The standard error of the estimate is a measure of the precision of fitting. Lower values of SEE correspond to improved model fitting.

$$s = \sqrt{\frac{\sum (Y_{obs} - Y_{calc})^2}{n - p - 1}} \tag{1}$$

In Eq. 1, Y_{obs} and Y_{calc} are the observed (experimental) and estimated scores respectively, while n is the number of compounds and p is the number of descriptors (it should not be confused with the probability '*p*' value denoting the level of significance).

The determination coefficient R^2 [10] measures the variation of the observed or experimental data with the predicted ones. Errors either in the data or in the model will lead to a bad fit. The maximum possible value for R^2 is 1, which defines a perfect correlation. R^2 is calculated from the following equation:

$$R^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{calc(train)})^2}{\sum (Y_{obs(train)} - \overline{Y}_{training})^2} \tag{2}$$

Here, $\overline{Y}_{training}$ is the mean observed activity of the training set compounds.

Another parameter used for testing the quality of developed regression equation is adjusted R^2 (R_a^2) [10]. R_a^2 (Eq. 3) is calculated to overcome the drawbacks associated with the increasing value of R^2 with an increase in number of variables. The explained variance of a model can be adjusted in the form of the modified R^2 as shown below.

$$R_a^{\,2} = \frac{(n-1)R^2 - p}{n - p - 1} \tag{3}$$

VIF

The descriptors in a multiple linear regression (MLR) equation should show minimum intercorrelation. Variance inflation factor (VIF) [19] of predictor variables should be checked for MLR models in order to check the presence of multicollinearity along with the model performance. VIF can be calculated from the following equation:

$$VIF = \frac{1}{1 - R_i^2} \tag{4}$$

where, R_i^2 the unadjusted R^2 when one regresses X_i against the remaining descriptive variables of the model. Multicollinearity is considered as very high when VIF value is greater than 5.

F-Ratio

The *F* ratio is used in order to quantify the statistical importance of the regression model. A higher value of *F* implies that a more significant correlation has been achieved. It is defined in the following Eq. 5 [10]:

$$F = \frac{\sum \left(Y_{calc} - \overline{Y}\right)^2 \Big/ p}{\sum \left(Y_{obs} - Y_{calc}\right)^2 \Big/ \left(n - p - 1\right)} \tag{5}$$

where, Y_{obs} is the observed response, Y_{calc} is the calculated response, n defines the total number of compounds and predictor variables is denoted as p.

FIT Kubinyi Function

The correlation coefficient R is clearly not the ultimate criterion because it tends to choose as many variables as possible. The standard error s value also tries to include too many variables, and the F value occasionally selects less number of variables than that frequently accepted by a QSAR researcher as large F values are often achieved by including only one or two variables in the model. The FIT Kubinyi function is closely connected to the F value [20].

$$FIT = \frac{R^2\left(n-p-1\right)}{\left(n+p^2\right)\left(1-R^2\right)} \tag{6}$$

In Eq. (6), n defines the number of training set compounds and p denotes the number of generated variables in the model.

Akaike's Information Criteria (AIC)

The goodness of fit and the number of variables that has to be estimated to attain that degree of fit are considered by Akaike's information criteria (AIC). AIC is calculated using the following equation:

$$AIC = RSS\frac{\left(n+p'\right)}{\left(n-p'\right)^2} \tag{7}$$

where, RSS is the sum of squared differences between the observed and estimated response; n defines the number of training set compounds; and p' denotes the number of adjustable variables (the best possible combination of predictor variables for the model) in the model [21]. When comparing a number of models, the model that creates the least value of this statistic should be identified as the most useful one. However, it may be noted that AIC can be employed to evaluate classification based QSAR models also if the RSS parameter in eq. 7 is substituted by any error estimation metric, *i.e.* classification error.

Additional metrics need to be employed to judge the predictive ability of the QSAR models as acceptable values of these statistical parameters only are not always adequate. To optimally determine the predictability of the models, they are required to be further assessed utilizing diverse validation metrics. Thus, internal and external validation experiments are performed in order to check the predictability the models.

INTERNAL VALIDATION

Internal validation of a QSAR model is done employing the molecules present in the training set [22-23]. It involves activity prediction of the training set molecules followed by a check of the precision of predictions. The cross-validation approach involves the leave-group-out cross-validation (LGO-CV) method where a set of *n* number of observations is divided into calibration and validation subsets. The calibration subset is used to develop the model, while the validation set is used to test the predictability of the model for the new data which are not used in the calibration. The method comprises two commonly used techniques namely, leave-one-out and leave-many-out cross-validation techniques.

Leave-One-Out (LOO) Cross-Validation

The training data set is initially modified by eliminating one compound from the set for LOO cross-validation. The QSAR model is then constructed based on the remaining molecules of the training set and the activity of the omitted compound is computed based on the resulting QSAR equation. This process is repeated until all the molecules of the training set have been deleted once, and the predicted activity data are obtained for all the training set compounds. The model predictivity is evaluated using the predicted residual sum of squares (*PRESS*) and cross-validated R^2 (Q^2) [24, 25] for the model. *PRESS* is a sum of squared differences between experimental and predicted data while the value of standard deviation of error of prediction (*SDEP*) [8, 26, 27] is calculated from *PRESS*. Eqns. 8, 9 and 10 give the expressions for *PRESS*, *SDEP* and Q^2 respectively.

$$PRESS = \sum (Y_{obs} - Y_{pred})^2 \tag{8}$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \tag{9}$$

$$Q^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{pred(train)})^2}{\sum (Y_{obs(train)} - \overline{Y}_{training})^2} = 1 - \frac{PRESS}{\sum (Y_{obs(train)} - \overline{Y}_{training})^2} \tag{10}$$

In Eq. 8, Y_{obs} and Y_{pred} correspond to the observed and LOO predicted activity values while in Eq. 9, n refers to the number of observations. In Eq. 10, $Y_{obs(train)}$ is the observed activity, $Y_{pred(train)}$ is the predicted activity of the training set molecules based on the LOO technique. The accepted threshold value of Q^2 is 0.5.

It is interesting to point out that only a higher value of Q^2 is an inadequate criterion to assess the predictive potential of a QSAR model. Structural redundancy of the training set may be a cause for overestimation of the value of Q^2 [28]. The models developed may suffer from the problem of overfitting. Thus, despite bearing a significant correlation between the descriptors and response parameter, the developed model may fail to predict correctly the activity of new compounds. So, LOO-Q^2 can serve only as a crude criterion to judge the quality and robustness of a model while external validation plays a key role in detecting the ability of the model to predict the new set of molecules.

Leave-Many-Out (LMO) Cross-Validation

The basic principle of the leave-many-out technique (LMO) method is that a specific section ($1 \leq M < N$ where N is a sample size) of the training compounds is held out and omitted in each cycle [29]. For each cycle, the model is constructed employing the reduced dataset and the deleted compounds are predicted utilizing the developed model. After completion of all cycles, the predicted activity values of the compounds are considered for the calculation of the LMO-Q^2. Based on the predicted values of the deleted compounds in each of the cycles, predictive R^2 may be calculated for each cycle. Thus, within the process of internal validation, the LMO technique gives to some extent a flavour of external validation. The major steps for LOO and LMO cross-validation are presented in Fig. **2**.

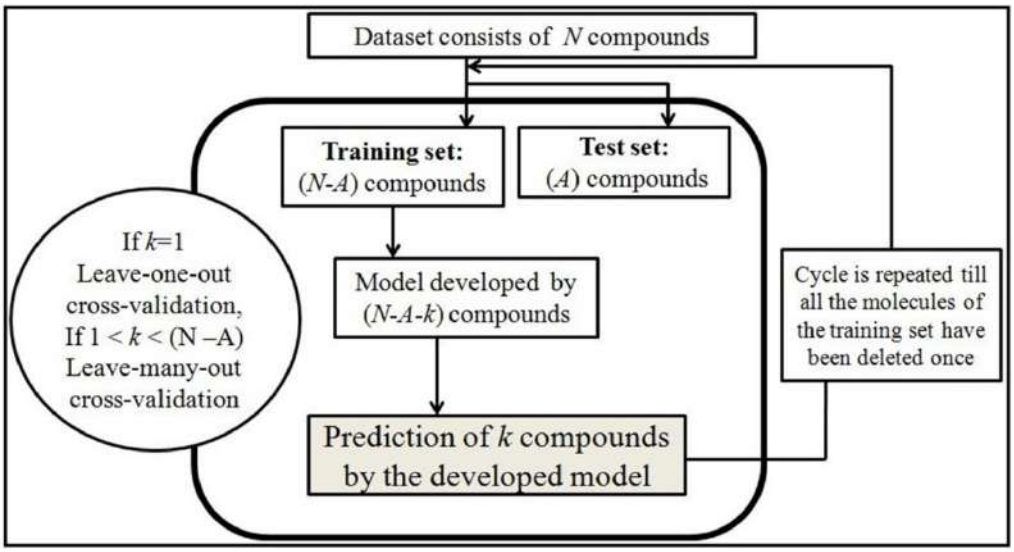


Figure 2: Major steps for the calculation of LOO and LMO cross-validation metrics.

Hou Fitness Function

Hou fitness function [30, 31] is defined as the metric (Eq. 11) which unites a multiple correlation coefficient R and the leave-one-out Q (Q_{LOO}).

$$R_p = RQ_{LOO} \tag{11}$$

The acceptability of a model is tested with the leave-one-out cross-validated correlation coefficient (Q^2), which can be defined as:

$$Q^2 = (\text{SSY-PRESS})/\text{SSY} \tag{12}$$

where, SSY is the sum of the squared deviations of the dependent variable values from their mean, PRESS is the predicted sum of squares obtained from the leave-one-out cross-validation method.

True Q^2

The process of selection of the training and test sets may be prejudiced in many cases though external validation is largely accepted by various research groups. Moreover, the division of a dataset may result in loss of information (for the test set) which otherwise could have been used for developing the QSAR model.

Thus, according to Hawkins *et al.* [32], holding out a part of the dataset is useless in case of a small dataset. Again in case of cross-validation, the model is constructed using the training set where the separation is done internally. It is critical that each compound should be removed for prediction so that it is not used in any way in the model fitting applied to the remaining retained *n*-1 compounds. The activity of the deleted compound is then predicted using the developed equation employing retained *n*-1 compounds. Thus, Hawkins *et al.* proposed the concept of "true Q^2" parameter, which can be computed employing the variable selection strategy at each validation cycle. In case of small data sets, compared to the traditional approach of the splitting of the data set into training and test sets, this may be a better metric for assessment of predictability [32, 33].

Hawkins *et al.* [32] arrived at the conclusion that (1) if one has 100 compounds available in a particular dataset, then he can use all compounds for calibration with accurate and precise Q^2; (2) if the dataset consists of 120 compounds, splitting the set into 100 for training set and 20 for test set is not a worthy one as the information in the test compounds is far substandard to what one gets from the training set's Q^2; (3) if one has 150 compounds, one can use 100 for training set and the remaining ones for test set; the test set will give an approximation of R^2 moderately equivalent with the Q^2.

r_m^2 Metric for Internal Validation

The mean response value of the training set molecules and the distance of the mean from the response values of the each compound plays a critical role in computing the Q^2 value. The Q^2 value increases with an increase of the value of the denominator ($\sum (Y_{obs(train)} - \overline{Y_{training}})^2$) in the right hand side of the equation 10. Thus, even for great disparity in the predicted and observed response values, satisfactory Q^2 values may be obtained if the molecules exhibit a considerably broad range of response data. Hence, an acceptable Q^2 does not guarantee that the predicted activity values lie in close proximity to the observed ones although there may exist a good overall correlation between the values. Thus, to prevent this error and to better point out the model predictability, the r_m^2 metrics (Eqs. 13 and 14) for internal validation, introduced by Roy and co-workers [34-36], may be calculated.

$$\overline{r_m^2} = \frac{(r_m^2 + r'^2_m)}{2} \tag{13}$$

$$\Delta r_m^2 = | r_m^2 - r'^2_m | \tag{14}$$

Here, $r_m^2 = r^2 \times (1 - \sqrt{(r^2 - r_0^2)})$ and $r'^2_m = r^2 \times (1 - \sqrt{(r^2 - r'^2_0)})$. The parameters r^2 and r_0^2 are the squared correlation coefficients between the observed and (leave-one-out) predicted values of the compounds with and without intercept respectively. In the initial studies, observed values were considered in *y*-axis whereas predicted values were considered in the *x*-axis. r'^2_0 bears the same meaning but uses the reversed axes. It is interesting to note that during the change of axes, the value of r^2 remain same while it is not true for the case of r_0^2. When the observed values (*y*-axis) are plotted against the calculated values of the compounds (*x*-axis) setting intercept to zero, the slope of the fitted line gives the value of *k*. Interchange of the axes gives the value of k'. The following equations are employed for the calculation of r^2, r_0^2, *k* and k'.

$$r_0^2 = 1 - \frac{\sum (Y_{obs} - k \times Y_{pred})^2}{\sum (Y_{obs} - \overline{Y_{obs}})^2} \tag{15}$$

$$r'^2_0 = 1 - \frac{\sum (Y_{pred} - k' \times Y_{obs})^2}{\sum (Y_{pred} - \overline{Y_{pred}})^2} \tag{16}$$

$$k = \frac{\sum (Y_{obs} \times Y_{pred})}{\sum (Y_{pred})^2} \tag{17}$$

$$k' = \frac{\sum (Y_{obs} \times Y_{pred})}{\sum (Y_{obs})^2} \tag{18}$$

$\overline{r_m^2}$ is the average value of r_m^2 and r'^2_m, and Δr_m^2 is the absolute difference between r_m^2 and r'^2_m. In general, the difference between r_m^2 and r'^2_m values of the training set should be null for better models. $\overline{r_m^2}_{(LOO)}$ and $\Delta r_m^2{}_{(LOO)}$ parameters can be used for the internal validation of the training set and it has been shown that the

value of $\Delta r_m^{2}{}_{(LOO)}$ should be less than 0.2 provided that the value of $\overline{r_m^{2}}_{(LOO)}$ is more than 0.5. In Fig. **3** we have tried to demonstrate the major steps for the calculation of r_m^2 metrics.

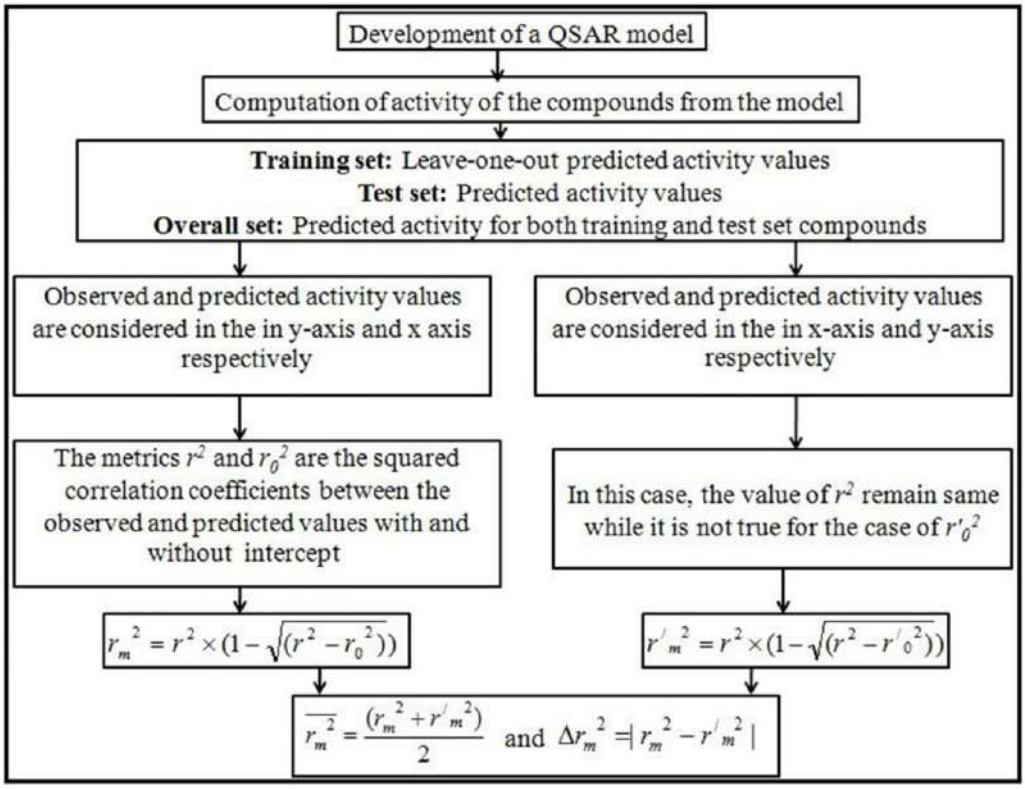


Figure 3: Major steps for the calculation of r_m^2 metrics.

True $r_m^{2}{}_{(LOO)}$

True $r_m^{2}{}_{(LOO)}$ is calculated from the model developed utilizing the undivided data set employing the variable selection stratagy at each cycle of validation. External validation characteristics are truly reflected by the 'true $r_m^{2}{}_{(LOO)}$' for the QSAR model as reported by Mitra *et al.* [37]. In case of small datasets, the predictability and accuracy of predictions can be efficiently judged employing this metric.

EXTERNAL VALIDATION

Although internal validation is the most accepted approach for validation of a QSAR model, but it is not the ultimate approach to judge the predictive power of

a QSAR model for NCE [13]. The true predictive ability of the model is judged based on the values of predictive R^2 (R^2_{pred}) and r_m^2 metrics $\overline{r_m^2{}_{(test)}}$ and $\overline{r_m^2{}_{(overall)}}$ from test set predictions.

Predictive R^2 (R^2_{pred})

The first step of external validation is splitting of the entire dataset into training and test sets. The choice and number of compounds in the training and test sets are the principal conditions for development of a statistically noteworthy QSAR model [33]. The QSAR models are constructed based on the training set compounds, and the activity of the test set molecules is predicted utilizing the developed model. Thus, the metric R^2_{pred} [13] (Eq. 19) replicates the measure of correlation between the observed and predicted data.

$$R^2_{pred} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{training})^2} \tag{19}$$

Here, $Y_{obs(test)}$ and $Y_{pred(test)}$ signify the observed and predicted response data for the test set molecules, while $\overline{Y}_{training}$ denotes the mean observed activity of the training set compounds. The stipulated threshold value of R^2_{pred} is 0.5 for acceptability of any QSAR model.

Validation Based on Golbraikh and Tropsha's Criteria

In case of least squares, the experimental *versus* fitted and the fitted *versus* experimental plots are not always equivalent [38]. According to Golbraikh and Tropsha [13], regressions of y against y' or y' against y through the origin, should be characterized by either k or k' (slopes of the corresponding regression lines) being close to 1. Consequently, the regression lines through origin are defined by $y^{r_0} = ky'$ and $y'^{r_0} = k'y$ while the slopes k and k' are given by Eqs. (20) and (21) respectively.

$$k = \frac{\sum y_i y_i'}{\sum y_i'^2} \tag{20}$$

$$k^{/} = \frac{\sum y_i y_i^{/}}{\sum y_i^2} \tag{21}$$

Golbraikh and Tropsha proposed more strict conditions for a QSAR model in order to ensure high predictability. They suggested that either of the squared correlation coefficients of these two regression lines (y against $y^{/}$ or $y^{/}$ against y through the origin) r_0^2 or r'^2_0 (given by Eqs. 22 and 23) respectively should be close to the value of r^2 for the developed model. The values of r^2 and r_0^2 specify the squared correlation coefficients between the observed and the predicted values with and without intercept respectively while r'^2_0 represents the same information as r_0^2 does but with inverted axes.

$$r_0^2 = 1 - \frac{\sum (y_i^{/} - y_i^{r_0})^2}{\sum (y_i^{/} - \overline{y^{/}})^2} \tag{22}$$

$$r'^2_0 = 1 - \frac{\sum (y_i - y_i^{/r_0})^2}{\sum (y_i - \overline{y})^2} \tag{23}$$

In Eqs. (22) and (23), $\overline{y^{/}}$ and $\overline{y}$ signify the mean values of the predicted and observed data, respectively. Therefore, according to Golbraikh and Tropsha [13], acceptable QSAR models should maintain the following conditions:

i) $Q^2_{training} > 0.5$.

ii) $R^2_{test} > 0.6$.

iii) $\frac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \le k \le 1.15$ or

$\frac{r^2 - r_0'^2}{r^2} < 0.1$ and $0.85 \le k' \le 1.15$.

iv) $\left| r_0^2 - r_0'^2 \right| < 0.3$.

$r_m^2{}_{(test)}$ for External Validation

As shown by the expression of R^2_{pred}, similar to Q^2, the value of R^2_{pred} relies on the average activity value of the training set molecules. Thus, high values of R^2_{pred} may be obtained when the test set molecules bear a wide range of activity data;

but this may not signify that the predicted values are very near to the corresponding observed data. In order to find out the propinquity between the observed and predicted activity data for the test set compounds, the metric $r_m^{2}{}_{(test)}$ [35], parallel to $r_m^{2}{}_{(LOO)}$ employed in internal validation, has been introduced. The $r_m^{2}{}_{(test)}$ is calculated utilizing the squared correlation coefficients between the predicted and observed values of the test set molecules. As suggested by Roy *et al*., the value of $\overline{r^2_{m(test)}}$ should be more than 0.5 provided that the value of $\Delta r_m^{2}{}_{(test)}$ is lower than 0.2 [35].

It is interesting to point out that, the r_m^2 metrics are not limited to the training and test set only. Roy and coworkers [35] showed that it can be extended to the entire dataset employing the LOO predicted data for the training set and predicted data for the test set molecules. The metrics have been defined as $\overline{r_m^{2}{}_{(overall)}}$ and $\Delta r_m^{2}{}_{(overall)}$ which reflect the predictability of the model for the entire dataset. The major advantages of this parameter include among other things: (i) unlike external validation, the $r_m^{2}{}_{(overall)}$ metric includes both training and test set molecules and thus the statistic is based on prediction of comparably large number of compounds which imparts greater reliability to the model; (ii) when many equivalent models are obtained, where few models explain superior reliability in terms of the internal validation metrics while others may demonstrate better external validation statistics, choice of the best model becomes complicated. Since $\overline{r_m^{2}{}_{(overall)}}$ and $\Delta r_m^{2}{}_{(overall)}$ are based on the entire dataset, the values of these parameters enable selection of the best model based on an overall contribution of both internal and external validation measures. The r_m^2 statistics have been discussed by Roy and coworkers [34-37, 39-50] in many reports. The parameter r_m^2 has also been used by various other groups of authors [51-59] to verify the robustness and predictability of QSAR models.

ADDITIONAL FUNCTIONS FOR MODEL PREDICTIVE ABILITY RMSEP

External predictability of a QSAR model may further be calculated by an assessment of the observed activity and the predictions of the test set molecules through calculation of a metric referred to as root mean square error in prediction (rmsep) [60] given by Eq. 24.

$$RMSEP = \sqrt{\frac{\sum (y_{obs(test)} - y_{pred(test)})^2}{n_{ext}}} \tag{24}$$

Here, n_{ext} signifies the number of test set compounds. The parameter depends exclusively on the divergence between the predicted and observed activity values and can also be calculated when there is only single test compound is present.

$Q^2_{(F2)}$

Schüürmann *et al.* [61] proposed an additional metric for computation of external Q^2, *i.e.*, Q^2 based on prediction of test set compounds ($Q^2_{(F2)}$) as given by Eq. 25.

$$Q^2_{(F2)} = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \overline{Y}_{test})^2} \quad (25)$$

Here, Y_{test} signifies the mean observed activity of the test set compounds. In case of $Q^2_{(F2)}$, the mean activity value used in the denominator represents external set compounds. Almost equal or close values of $Q^2_{(F2)}$ and $Q^2_{(F1)}$ infer that the training set mean lies in the close propinquity to that of the test set which indicates that the test set utilized for modeling covers the whole response domain of the model. The threshold value 0.5 is defined for this parameter.

$Q^2_{(F3)}$

Another parameter ($Q^2_{(F3)}$ with a threshold value of 0.5) for validation of a QSAR model has been proposed by Consonni *et al.* [60]. This metric can be computed by the following equation:

$$Q^2_{(F3)} = 1 - \frac{[\sum (Y_{obs(test)} - Y_{pred(test)})^2] / n_{ext}}{[\sum (Y_{obs(train)} - \overline{Y}_{train})^2] / n_{tr}} \quad (26)$$

In Eq. 26, n_{tr} and n_{ext} denote the number of training set and test set compounds, respectively. However, although $Q^2_{(F3)}$ measures the model predictability, it is sensitive to selection of training data set and tends to penalize models fitted to a very homogeneous dataset even if predictions are close to the truth [60-62].

Concordance Correlation Coefficient (CCC)

The CCC parameter [63] can also be calculated in order to check the model reliability by the following equation:

$$\overline{\rho}_c = \frac{2\sum_{i=1}^{n}(x_{obs(test)} - \overline{x_{obs(test)}})(y_{pred(test)} - \overline{y}_{pred(test)})}{\sum_{i=1}^{n}(x_{obs(test)} - \overline{x_{obs(test)}})^2 + \sum_{i=1}^{n}(y_{pred(test)} - \overline{y_{pred(test)}})^2 + n(\overline{x_{obs(test)}} - \overline{y_{pred(test)}})} \quad (27)$$

Here, $x_{obs(test)}$ and $y_{pred(test)}$ refer to the observed and predicted values of the test set, n denotes the number of compounds, and $\overline{x_{obs(test)}}$ and $\overline{y_{pred(test)}}$ signify the mean of observed and predicted values, respectively. The CCC coefficient measures both precision and accuracy detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin respectively. Any deviation of the regression line from the concordance line (line passing through the origin) give a value of CCC smaller than 1.

$r_{m}^{2}{}_{(rank)}$ Metric

The $r_{m}^{2}{}_{(rank)}$ parameter was introduced to measure the proximity between the order of the observed data and that of the corresponding predicted data [64]. The $r_{m}^{2}{}_{(rank)}$ metric was introduced by Roy and co-workers with the intention of incorporating the rank-order predictions of molecules. The $r_{m}^{2}{}_{(rank)}$ metric is computed utilizing the correlation of the ranks generated from the observed and the corresponding predicted data. First, the observed and predicted data of the molecules are ranked and the (Pearson's) correlation coefficients of the corresponding ranks are determined with ($r^{2}{}_{(rank)}$) and without intercept ($r_{0}^{2}{}_{(rank)}$). The $r^{2}{}_{(rank)}$ and $r_{0}^{2}{}_{(rank)}$ calculated employing the rank-order are used to compute the $r_{m}^{2}{}_{(rank)}$ metric. The values of $r^{2}{}_{(rank)}$ and $r_{0}^{2}{}_{(rank)}$ vary from each other considering the variation in ranking of the two variables. In an ideal ranking, where the observed and predicted response data perfectly match with each other for all molecules, the obtained $r_{m}^{2}{}_{(rank)}$ metric value is 1. The minimum acceptable threshold value for the $r_{m}^{2}{}_{(rank)}$ metric is 0.5.

ASSESSMENT OF CHANCE CORRELATION: Y-RANDOMIZATION

Y-randomization test is a practice to make sure the robustness of the QSAR model. In Y-randomization, validation is carried out by permuting the response values (Y) with respect to the descriptor matrix which has been kept unaltered.

There are two types of randomization tests namely process and model randomization that can be performed at varying confidence levels. In case of process randomization, the response variables are arbitrarily jumbled, and variable selection is done newly from the entire descriptor matrix. In contrast, for model randomization, the dependent variables are scrambled and new models are constructed employing the same set of variables as present in the nonrandom model.

$^{c}R_{p}^{2}$

The degree of variation in the values of the squared mean correlation coefficient of the randomized model (R_r^2) and squared correlation coefficient of the nonrandom model (R^2) is reflected in the value of $^{c}R_p^2$ [65] parameter. This metric penalizes the model R^2 for a small difference between the values of the squared correlation coefficients of the non-random (R^2) and the randomized (R_r^2) models as per the following equation (Eq. 28):

$$^{c}R_{p}^{2} = R \times \sqrt{R^2 - R_r^2} \tag{28}$$

The threshold value of $^{c}R_p^2$ is 0.5 and a QSAR model having the corresponding value above the stated limit might be correctly considered that the model is not obtained by chance only.

Another approach for the assessment of chance correlation deals with set of decision inequalities utilizing the values of Q^2_{yrand} and R^2_{yrand} and their relationship as that $R^2_{yrand} > Q^2_{yrand}$. The approach has been suggested by Eriksson and Wold [9]. They have set the following rules to measure any kind of chance correlation in the QSAR model.

"(i) $Q^2_{yrand} < 0.2$ and $R^2_{yrand} < 0.2$ → no chance correlation

(ii) any Q^2_{yrand} and $0.2 < R^2_{yrand} < 0.3$ → negligible chance correlation

(iii) any Q^2_{yrand} and $0.3 < R^2_{yrand} < 0.4$ → tolerable chance correlation

(iv) any Q^2_{yrand} and $R^2_{yrand} > 0.4$ → recognised chance correlation"

Thus, the number of randomization resulting in false positive correlations is counted up as the correlation frequency from the Q^2_{yrand} *vs* R^2_{yrand} plot.

VALIDATION METRICS FOR CLASSIFICATION BASED QSAR MODELS

In order to assess the performance of classification based models, validation is one of the utmost criterion in terms of qualitative predictions [66]. Validation for a classification based model is often performed for two-class problems, where the compounds are classified into actives or inactives.

Goodness-of-fit and Quality Measures

Wilks Lambda (λ) Statistics

The Wilks Lambda is a widely used parameter for the testing of significance of discriminant model function. It is a distance based parameter and is calculated from the scalar transformations of the covariance matrices of between and within-groups variances. In a classification analysis, where at least two groups are present, Wilks lambda is determined as the ratio of within group sum of squares and total sum of squares, *i.e.* within-category to total dispersion [67].

$$Wilks\ \lambda = \frac{Within\ group\ sum\ of\ square}{Total\ sum\ of\ square} \tag{29}$$

Let us consider B_g and W_g are the random $p \times p$ independent variable matrix with the distribution W_p (q, Σ) and W_p (n, Σ), respectively considering $n>p$. Then the Wilks λ will be given by the following equation [33]:

$$\lambda = \det\left(\frac{B_g}{B_g + W_g}\right) \tag{30}$$

Where '*det'* refers to the determinant of the matrix constructed using descriptors. Wilks lambda focuses on the best discriminating property of the analyzed independent variables and it spans from 0 to 1, where 0 corresponds to different values of group means signifying good level of discrimination achieved by variable and 1 referring to similar group mean values meaning no discrimination achieved by the variables. Hence, the value of Wilks lambda for a good discriminant model should preferably be lower.

Canonical Index (R_c)

The measure of the strength of the relationship between the two variates is expressed as a canonical correlation coefficient [68].

$$R_c = \sqrt{\frac{\lambda_i}{1 + \lambda_i}} \tag{31}$$

Here, λ_i is refereed as eigen value of the matrix.

Chi-Square (χ^2)

The chi-square (χ^2) statistic identifies the liberty between two groups signifying that a higher value of this metric will point out superior separability between groups, *i.e.* good classification analysis [69].

$$\chi^2 = \sum_{i=1}^{t} \frac{(f_i - F_i)^2}{F_i} \tag{32}$$

where, f_i is observed response, F_i is predicted response and t is the number of observations.

Squared Mahalanobis Distance

Square of Mahalanobis distance is measured during linear discriminant analysis for the determination of likelihood of a compound to be classified in a specific group in the discriminant space. In a referred discrimination space or transformed space, euclidean distances among data points become equal to Mahalanobis distances. In a multivariate normal distribution with covariance matrix Σ, the Mahalanobis distance between any two data points x_i and x_j can be defined as [67]:

$$d_{Mahalanobis}(x_i, x_j) = \sqrt{(x_i - x_j)^T \sum^{-1} (x_i - x_j)} \tag{33}$$

In eqn. 33, x_i and x_j are two random data points, T is transpose of a matrix and Σ^{-1} is inverse of the covariance matrix.

Model Performance Parameters

The compounds can be classified using the developed classification based QSAR models into four groups based on a assessment between the observed and classified activity values: (i) true positives (TP): the active (positive) compounds which have been correctly classified as actives, (ii) false positives (FP): the active compounds which have been incorrectly classified as inactives, (iii) false negatives (FN): the inactive (negative) compounds erroneously classified as positives, *i.e.* actives, and (iv) true negatives (TN): inactive compounds which have been correctly predicted as inactive [70]. Employing this classification, a two-by-two confusion matrix [71] can be prepared individually for the training and test sets. To assess the classifier model performance and classification competence, various statistical tests are employed.

$$Sensitivity = \mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{34}$$

$$Specificity = \frac{TP}{TP + FP} \tag{35}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{36}$$

$$F - measure = \frac{2(\mathrm{Re}\,call)(\mathrm{Pr}\,ecision)}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \tag{37}$$

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} = fp\ rate \tag{38}$$

G-Means

A simple way to assess the model's capability to correctly classify active and inactive compounds using the combination of sensitivity and specificity into a single value employing the geometric mean (G-means) [72]:

$$G\text{-}means = \sqrt{Sensitivity \times Specificity} \tag{39}$$

Cohen's κ

Cohen's kappa (κ) can be used to determine the conformity between classification models and known classifications [73]. It can be calculated using the following formulae:

$$Cohen's\ \kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} \tag{40}$$

$$P_r(a) = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{41}$$

$$P_r(e) = \frac{\{(TP + FP) \times (TP + FN)\} + \{(TN + FP) \times (TN + FN)\}}{(TP + FN + FP + TN)^2} \tag{42}$$

where, $P_r(a)$ is the relative agreement between the predicted classification of the model and the known classification, and $P_r(e)$ is the hypothetical probability of chance agreement. The values of $P_r(a)$ and $P_r(e)$ can be calculated from the generated confusion matrix. Cohen's kappa returns values between −1 (no agreement) and 1 (complete agreement). Cohen's kappa values between −1.0 and 0.4 point out that the model is a poor predictor, values between 0.4 and 0.6 designate that the model is average, values between 0.6 and 0.8 imply that the model is satisfactory, and values between 0.8 and 1.0 signify that the model is highly predictive.

Receiver Operating Characteristics (ROC) Curve

A ROC curve presents a visual representation of the success and error observed in a classification model. The curve is plotted taking true positive rate (tp) on the *Y* axis and false positive rate (fp) rate on the *X* axis, and the nature of the curve provides easier detection of the correctness of prediction. Apart from classification problems, ROC curves have been a useful measure in signal detection theory since the past intended for determining the tradeoff between hit rates and false alarm rates of classifiers [71].

$$tp\ rate \approx \frac{Positives\ (active\ molecules)\ correctly\ classified}{Total\ postives} = Sensitivity \tag{43}$$

$$fp\ rate = \frac{Negatives\ (inactive\ compounds)\ incorrectly\ classified}{Total\ negatives} = 1 - specificity \qquad (44)$$

Thus, the ROC curve may also be drawn by plotting the *fp rate* and the *tp rate* along the X and Y axes respectively. It signifies the number of correctly and wrongly identified objects by the developed classifier. A sample picture for ROC curve is presented in Fig. **4**. Most classifiers can be varied from "conservative" to "liberal" classifiers. A perfect classifier correctly classifies all positive cases and has no false positives. A conservative classifier (lower left region of the ROC space) requires strong evidence to classify a point as positive while a liberal classifier (upper right region of the ROC space) does not require much evidence to classify an event as positive.

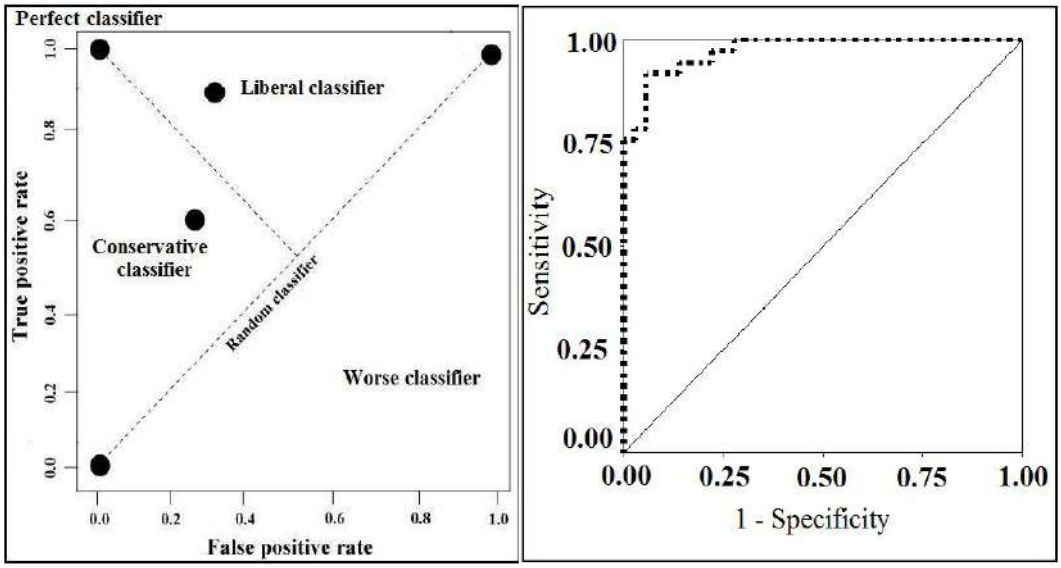


Figure 4: The ROC space (left) and a sample ROC curve (right).

ROCED and ROCFIT

Two new metrics employing distances in a ROC curve for the identification of the acceptable classification models are ROC graph Euclidean distance (ROCED) and ROC graph Euclidean distance corrected with Fitness Function (FIT(λ)) or Wilks λ (ROCFIT) [74]. It is interesting to point out that these two metrics make a proper balance in both training and test sets.

If the best model is one whose depiction is positioned as close as possible to the upper left corner in the ROC graph, a good indicator would be a measure of this distance. The Euclidean distance between the perfect and a real classifier (d_i) expressed by the following equation:

$$d_i = \sqrt{(Se_p - Se_r)^2 + (Sp_p - Sp_r)^2} \tag{45}$$

where Se_p and Se_r are the respective sensitivity values of the perfect and the real classifier, while Sp_p and Sp_r represent the specificity values of the perfect and real classifier, respectively. As both the sensitivity and specificity for a perfect classifier have a value of 1, the euclidean distance can be calculated as following equation:

$$d_i = \sqrt{(1 - Se_r)^2 + (1 - Sp_r)^2} \tag{46}$$

where i = 1 stands for the training set, and i = 2 for the test set. As these two distances corresponding to the training and test sets should be as small as possible, the parameter can be defined as follows:

$$ROCED = (|d_1 - d_2| + 1) \times (d_1 + d_2) \times (d_2 + 1) \tag{47}$$

where, d_1 and d_2 are depiction of the distances in a ROC graph for the training and test sets respectively.

ROCED can take values between 0 (which signifies an ideal classification for both training and test sets) and 4.5 (d_1 = 0.5 random classifier and d_2 = 1). Models with values greater than 2.5 indicate that these models have random responses.

A new parameter ROCFIT (ROC graph Euclidean Distance corrected with Wilks λ) has also been introduced to avoid the probable loss of significance in the variables of the models generated by linear discriminant analysis (LDA) using only Eqn. 47,. ROCFIT is defined as follows:

$$ROCFIT = \frac{ROCED}{Wilks(\lambda)} \tag{48}$$

AUC-ROC

AU-ROC is equal to a straightforward average of the ranks of the actives, the good performance of "*early recognitions*" is offset rapidly by "*late recognitions*" [75, 76]. AUROC is approximately normal distributed, with mean $\mu = 1/2 + 1/2(N-n)$ and variance $\sigma^2 = N+1/12n(N-n)$. Here, the number of actives and the total number of compounds are denoted by n and N respectively. AU-ROC defined in equation (47) is linearly related to the rank sum of actives, which is also called Mann-Whitney U test. r_i is the rank of the i_{th} active.

$$AUC\text{-}ROC = 1 - \frac{\sum_{i=1}^{n} r_i}{n \times (N-n)} + \frac{n+1}{2 \times (N-n)} \tag{49}$$

RIE and BEDROC

Truchon and Bayly [75] have shown that the exponential weighting schemes, BEDROC and robust initial enhancement (RIE) provide good "*early recognition*" of actives. By changing the tuning parameter, α, users can control the earliness of "*early recognition*" to test whether a ranking method is useful in the context of *VS*. BEDROC is bounded by interval [0, 1] and can be interpreted as the probability that an active is ranked before a randomly selected compound exponentially distributed with parameter α, only when $\frac{n}{N}\alpha << 1$. RIE, developed by Sheridan *et al.* [77], used an exponential weighting scheme, that places heavier weight to "*early recognized*" actives.

$$RIE = \frac{\frac{1}{n}\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{1}{N}\left[\frac{1-e^{-\alpha}}{e^{\frac{\alpha}{N}}-1}\right]} \tag{50}$$

In Eq. (50), $x_i = r_i/N$ is the relative rank of the ith active and α is a tuning parameter. BEDROC is derived from RIE and it is bounded by [0, 1]. BEDROC has a linear relationship with RIE,

$$BEDROC = RIE \times \frac{\frac{1}{N}\sinh\left(\alpha/2\right)}{\cosh\left(\alpha/2\right) - \cosh\left(\alpha/2 - \alpha\frac{n}{N}\right)} + \frac{1}{1 - e^{\alpha\left(\frac{N-n}{N}\right)}} \quad (51)$$

Although RIE and BEDROC produce dissimilar values, their distributions are alike up to a scale and a translation factor, and their correlation is 1.

AUC-pROC

Logarithmic transformation shifts the emphasis from "late recognition" to "early recognitions". Clark and Clark [78] proposed a new metric, pROC, on basis of the negative logarithmic transformation of false positive rates, θ. When the false positive rate is zero, they suggested a zero-point continuity modification should be made by replacing zero with 1/N.

$$AUC\text{-}pROC = -\frac{1}{n} \times \sum_{i=1}^{n} \log_{10}\left(\frac{r_i - i}{N - n}\right) = \frac{1}{n}\sum_{i=1}^{n} \log_{10}\left(\frac{1}{\theta_i}\right) \quad (52)$$

RIE, BEDROC and AUC-pROC metrics are used for evaluating the performance of classification models as well as for the virtual screening performance of the models. It is interesting to point out that to employ these metrics to evaluate the performance of a model in typical external sets (about the same number of actives and inactives) will lead to results suffering the “saturation effect”.

Matthews Correlation Coefficient (MCC)

The quality of a binary (two-class) classification is measured by another parameter MCC [79]. MCC consider true and false positives as well as negatives and generally regarded as a balanced measure which can be utilized even if the classes are of very different sizes. The MCC is a correlation coefficient between the observed and predicted binary classifications which returns a value between −1 and +1. A perfect prediction is presented by an MCC coefficient of +1, average random prediction by 0 and −1 is considered as inverse prediction.

The MCC can be calculated considering the confusion matrix using the following formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (53)$$

The meaning of *TP*, *TN*, *FP* and *FN* are same as discussed earlier. The denominator value will be considered one if any of the sums in the denominator is zero. This parameter in particular addresses the issue of inappropriate explanation of a confusion matrix, and the cases where the dataset sizes are higher.

Pharmacological Distribution Diagram (PDD)

Pharmacological distribution diagram (PDD) is a frequency distribution plot of a dependent variable where expectancy values of the variable is plotted in the Y-axis against numeric intervals of the variable in the X-axis. In a classification issue, expectancy refers to the probability of categorization of a compound in a specific group for a specific value of the discriminant function. During LDA, a discriminant function (DF) is developed, which is a mathematical equation, used for the calculation of discriminant scores of every individual compounds (machine learning classification algorithms also provide scores that can be employed in the same way the LDA scores are). Then the discriminant function values of all samples are taken in the abscissa in the form of range, and the expectancy values (probability of activity) are plotted in the ordinate against those ranges. Hence, this graph visually signifies the overlapping regions of the categories *e.g.*, positives and negatives, as well as it shows the regions of DF values that possess maximal probability of finding actives and inactives [80]. For a classification case comprising of two classes like actives and inactives (or positives and negatives), two terms named 'active expectancy' and 'inactive expectancy' may be defined as below where the denominator is added with a numerical value of 100 to avoid division by zero [80]:

$$\text{Activity expectancy} = E_a = \frac{Percentage\ of\ actives}{\left(Percentage\ of\ inactives + 100\right)} \quad (54)$$

$$\text{Inactivity expectancy} = E_i = \frac{Percentage\ of\ inactives}{\left(Percentage\ of\ actives + 100\right)} \quad (55)$$

where '*a'* and '*i'* are the number of occurrences of active and inactive compounds at a specific range. It can be evidently understood that for a perfect classification scheme, the active (positive) and inactive (negative) compounds will always be characterized by different ranges of DF values, and hence in an ideal discriminant operation, the actives will always be separated than the inactives whereas overlapping of them will correspond to error in prediction referring to false positives as well as false negatives. A sample picture for PDDs showing good and bad classifications is represented in Fig. **5**.

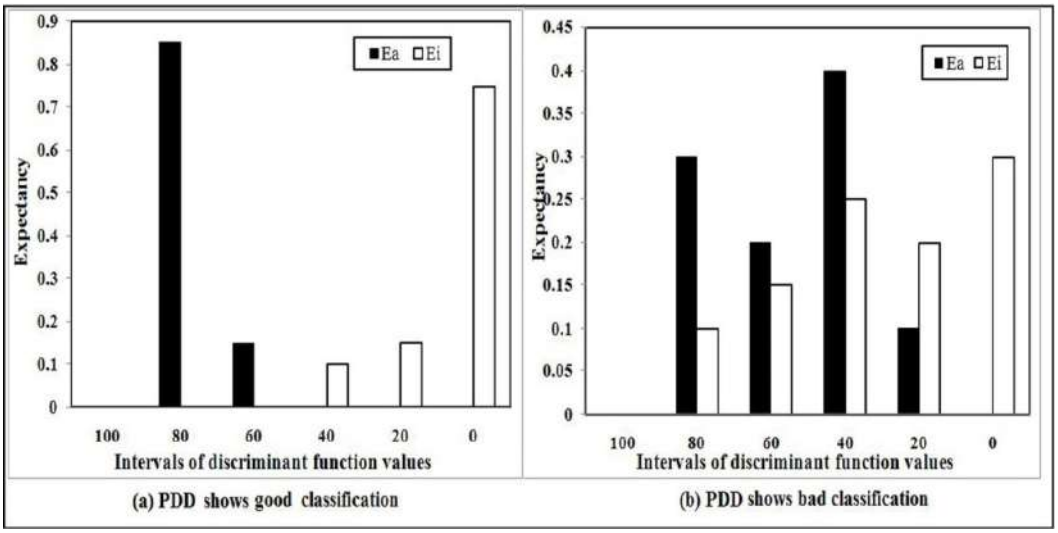


Figure 5: Sample PDDs showing good and bad classifications.

APPLICABILITY DOMAIN (AD)

The applicability domain (AD) [81-86] is a theoretical area in chemical space, defined by the model descriptors and modeled response. The AD of a model plays a crucial role for computing the uncertainty in the prediction for a test molecule based on similarity to the molecules utilized to construct the QSAR model. The prediction of a modelled response using QSAR is valid only when the compound being predicted falls within the AD of the model as it is unfeasible to predict all chemicals using a single model. AD is the physico-chemical, structural or

biological space based on which the training set of the model is constructed, and the model cane be used for predictions of NCEs within the specific domain [87]. To make confident predictions, a QSAR model should always be employed for those molecules which are within its applicability domain.

A conceptual guidance was offered by the Setubal Workshop [42] on defining AD of QSAR models, but it is complicated for use directly. Ideally the models should only be used to make predictions within its domain by interpolation [88]. There are four major approaches for estimating interpolation regions in a multivariate space as discussed thoroughly in Table **4** [86, 89-96].

Table 4: Methods for estimating AD

AD approaches	Methods	Hypothesis	Criteria
Range based	Bounding box (Descriptor ranges)	The range of each descriptor is considered defining an n-dimensional hyper-rectangle with sides parallel to the coordinate axes.	Any test set compounds, which are not present in any of these particular ranges, are considered out of the AD
	PCA bounding box (Principal components ranges)	Principal components (PC) construct a new orthogonal coordinate system and allow to correct the correlations between descriptors. The minimum value and the maximum value of each PC define an n-dimensional hyper-rectangle with sides parallel to the PC. The AD is defined with the created hyper-rectangle.	
	TOPKAT Optimal Prediction Space	The Optimum Prediction Space (OPS) from TOPKAT uses a variation of PCA and therefore creates a novel orthogonal coordinate system. In this system, the OPS boundary is denoted by the minimum and maximum values of the data points on each axis of the OPS coordinate system.	The Property Sensitive object Similarity (PSS) between the training set and a queried point assesses the confidence of the prediction.
Geometric methods	Convex Hull	The coverage of an n-dimensional set using the convex hull calculation is estimated.	Interpolation space is defined by the smallest convex area containing the entire training set.
Distance-based methods	Mahalanobis distance	It considers the distance of an observation from the mean values of the independent variables but not the impact on the predicted value. Mahalanobis distance is unique as it involuntarily takes into account the correlation between descriptor axes.	Observations with values much higher than those of the remaining ones are considered to be outside of the AD.

Table 4: contd….

	Euclidean distance	The distance scores are calculated by the Euclidean distance norm. A distance score, d_{ij}, for two different compounds X_i and X_j can be measured by the Euclidean distance norm. The Euclidean distance can be expressed by the following equation: $$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$ The mean distances of one sample to the residual ones were calculated as follows: $$\overline{d_i} = \sum_{j=1}^{n} d_{ij} \Big/ n-1,$$ where, i=1,2,….,n. The mean distances are then normalized within the interval of zero to one.	Compounds with distance values adequately higher than the most active probes are identified to be outside the AD.
	City block distance	City-block distance is the summed difference across dimensions and is computed as: $$d(x,y) = \sum_{i=1}^{n}\left|x_i - y_i\right|$$ It examines the absolute differences between coordinates of a pair of objects (x_i and y_i). A triangular distribution is assumed by city-block distance and is predominantly helpful for the distinct type of variabls.	
	Leverage approach	The leverage (*h*) of a compound in the original variable space is computed utilizing the HAT matrix as: $H = X(X^TX)^{-1}X^T$, where H is an (n x n) matrix that orthogonally projects vectors into the space spanned by the columns of X.	The AD of the model is defined as a squared area within the ±3 band for standardised residuals and the leverage threshold [h*=3(p+1)/n], where p denotes the number of variables and n signify the number of data points.
	Hotelling T^2 test	It is a multivariate student's *t* test and presumes a normal data distribution. Statistical significance of the disparity on the means of two or more variables between two groups is assessed.	AD is determined with the *t* value. The significant compounds remain within the determined AD.

Table 4: contd….

Probability density distribution	Parametric method	It assumes a standard distribution such as Gaussian and Poisson distributions	Real data distribution is captured without help of any reference data point. Individual data points are checked whether they belong to the set.
	Non parametric method	It does not rely on such assumptions considering the data distribution	

CONCLUSION

With the introduction of modernized chemometric tools, a great number of researches have been originated for to the expansion of a noteworthy relationship between molecular structure and function, specifically in the fields of medicine and predictive toxicology. Exciting and encouraging outcomes of QSAR approaches are noticed in the process of drug design for the last two decades. Here, we have tried to discuss current best practices validation metrics for constructing robust and predictive QSAR models among the QSAR researchers. Like any computational approach, it is vital that the QSAR method is taken as a technically trustworthy device for predicting the biological activities/properties/ toxicities of untested drug/chemicals. To make it transferable and acceptable to the scientific community, validation metrics along with the mechanistic interpretation are of the utmost importance for any developed QSAR model. A wide range of validation techniques have also been proposed for identifying the capability of QSAR models to predict the activity of NCEs. Predictive potential of QSAR models is judged from the internal as well as by the external validation tests based on the training and external test set molecules respectively. Moreover, a single metric may often prove to be inadequate for assessing the performance of a model and therefore overall validation parameters should be considered for selection of the ideal QSAR model for a particular endpoint. Therefore, this book chapter has focused on the critical metrics of QSAR modelling to allow their proper utilization. The discussed metrics should be helpful to both computational and synthetic chemists as well as experimental biologists who are working in biological screening of chemical libraries using QSAR models.

ACKNOWLEDGEMENTS

SK is thankful to the Department of Science and Technology (DST), Government of India for awarding him a Research fellowship under the INSPIRE scheme. KR

thanks the Council of Scientific and Industrial Research (CSIR), New Delhi for awarding a major research project.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

DISCLOSURE

Some parts of this article have been previously published in *Combinatorial Chemistry & High Throughput Screening*, YEAR: 2014, VOLUME: 14, ISSUE: 6, Page nos. 450-474. DOI: 10.2174/138620711795767893

ABBREVIATIONS

AD = Applicability Domain

AIC = Akaike's information criteria

AUC = Area Under Curve

CCC = Concordance Correlation Coefficient

GA = Genetic algorithm

ITS = Internal test set

LDA = Linear Discriminatory Analysis

LMO = Leave-many-out

LOO = Leave-one-out

MCC = Matthews correlation coefficient

MIC = Model Instability Coefficient

MLR = Multiple Linear Regression

MVIC = Model Value Instability Coefficient

NCE = New Chemical Entity

OECD = Organization for Economic Cooperation and Development

PCA = Principal Component Analysis

PDD = Pharmacological distribution diagram

PRESS = Predicted residual sum of squares

QSAR = Quantitative structure-activity relationship

RIE = Robust initial enhancement

RMSEP = Root mean square error in prediction

ROC = Receiver operating characteristics

ROCED = ROC graph Euclidean distance

ROCFIT = ROC graph Euclidean distance corrected with Fitness Function

SDEP = Standard deviation of error of prediction

SLR = Sum of the log of ranks test

VIF = Variance inflation factor

REFERENCES

[1] Selassie, C.D. In: *Burger's Medicinal Chemistry and Drug Discovery*; Abraham D.J, Ed.; John Wiley, New York, USA, **2003**; pp. 1-48.

[2] Perkins, R.; Fang, H.; Tong, W.; Welsh, W.J. Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. *Environ. Toxicol. Chem.,* **2003**, *22*(8), 1666-1679.

[3] Walker, J.D.; Jaworska, J.; Comber, J.H.I.; Schultz, T.W.; Dearden, J.C. Guidelines for developing and using Quantitative Structure-Activity Relationships. *Environ. Toxicol. Chem.,* **2003**, *22*(8), 1653-1665.

[4] Kar, S.; Roy, K. Predictive toxicology using QSAR: A perspective. *J. Indian. Chem. Soc.,* **2010**, *87*, 1455-1515.

[5] Tikhonova, I.G.; Baskin, I.I.; Palyulin, V.A.; Zefirov, N.S. Virtual screening of organic molecule databases. Design of focused libraries of potential ligands of NMDA and AMPA receptors. *Russ. Chem. B*+ **2004**, *53*, 1335-1344.

[6] Balls, M.; Blaauboer, B.J.; Fentem, J.H.; Bruner, L.; Combes, R.D.; Ekwall, B.; Fielder, R.J.; Guillouzo, A.; Lewis, R.W.; Lovell, D.P.; Reinhardt, C.A.; Repetto, G.; Sladowski,

D.; Spielman, H.; Zucco, F. Practical aspects of the validation of toxicity test procedures–the report and recommendations of ECVAM workshop 5. *ATLA* **1995**, *23*, 129-147.

[7] Aptula, A.O; Jeliazkova, N.G.; Schultz, T.W.; Cronin, M.T.D. The better predictive model: high q^2 for the training set or low root mean square error of prediction for the test set? *QSAR Comb. Sci.,* **2005**, *24*(3), 385-396.

[8] OECD Document. Guidance Document on the Validation of (Quantitative) 1226 Structure-Activity Relationships (Q)SARs] Models, ENV/JM/MONO(2007)2, 2007.

[9] Wold, S.; Eriksson, L. In: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H.E, Ed.; VCH, Weinheim, Germany, **1995**, pp. 195-218.

[10] Snedecor, G.W.; Cochran, W.G. *Statistical Methods*. Oxford & IBH: New Delhi, **1967**.

[11] Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.,* **2001**, *58*, 109-130.

[12] Topliss, J.G.; Costello, R.J. Chance correlation in structure-activity studies using multiple regression analysis. *J. Med. Chem.,* 1972, *15*(10), 1066-1068.

[13] Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Mod.,* **2002**, *20*(4), 269- 276.

[14] Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.,* **2004**, *112*(12), 1249-1254.

[15] Jaworska, J.S.; Comber, M.; Auer, C.; Van Leeuwen, C.J. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ. Health Persp.,* **2003**, *111*(10), 1358-1360

[16] OECD Principles for the Validation of (Q)SARs, http://www.oecd.org/dataoecd/33/37/37849783.pdf (Accessed on: September **2014**).

[17] OECD, Environment Directorate, Joint Meeting of The Chemicals Committee and The Working Party on Chemicals, Pesticides and Biotechnology, http://www.olis.oecd.org/olis/2004doc.nsf/ (Accessed on: September **2014**).

[18] Tropsha, A.; Gramatica, P.; Gombar, V.K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.,* **2003**, *22*(1), 69-77.

[19] O'Brien, M., Robert, A. caution regarding rules of thumb for variance inflation factors. *Qual. Quant.,* **2007**, *41*, 673-690.

[20] Kubinyi H. Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. *Quant. Struct. Act. Relat.,* **1994**, *13*, 393-401.

[21] Akaike, H. A new look at the statistical model identification. *IEEE Trans Autom. Control.,* **1974**, *19*, 716-723.

[22] Wold, S. Cross-validation estimation of the number of components in factor and principal components models. *Technometrics* **1978**, *20*, 397-405.

[23] Wold, S. Validation of QSAR's. *Quant Struct Act Relat* **1991,** 10, 191-193.

[24] Debnath, A.K. In: *Combinatorial Library Design and Evaluation*. Ghose, A.K.; Viswanadhan, V.N., Eds.; Marcel Dekker: New York, **2001**.

[25] Roy, K. On some aspects of validation of predictive quantitative structure-activity relationship models. *Expert. Opin. Drug. Discov.* 2007; 2(12); 1567-1577.

[26] Tichý, M.; Rucki, M. Validation of QSAR models for legislative purposes. *Interdisc. Toxicol.* 2009, *2*(3), 184-186.

[27] Cruciani, G.; Baroni, M.; Bonelli, D.; Clementi, S.; Ebert, C.; Skagerberg, B. Comparison of chemometric models for QSAR. *QSAR Comb. Sci.* 2006, *9*(2), 101-107.

[28] Clark, R.D.; Sprous, D.G.; Leonard, J.M. In: *Rational Approaches to Drug Design, Proceedings of the 13th European Symposium on Quantitative Structure–Activity Relationships*. Höltje, H.D; Sippl, W, Eds.; Prous Science: Dusseldorf, **2001**, pp. 475-485.

[29] Geisser, S. The predictive sample reuse method with application. *J. Amer. Stat. Ass.* **1975**, 70, 320-328.

[30] Hou, T.J.; Wang, J.; Xu, X. Applications of genetic algorithms on the structure– activity correlation study of a group of non-nucleoside HIV-1 inhibitors. *Chemom. Intell. Lab. Syst.,* **1999**, *45*, 303-310.

[31] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*. Wiley- VCH: Weinheim, Germany, **2009**.

[32] Hawkins, D.M., Basak, S.C., Mills, D. Assessing model fit, by cross-validation. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*(2), 579-586.

[33] Roy, P.P.; Leonard, J.T.; Roy, K. Exploring the impact of the size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2008**, *90*(1), 31-42.

[34] Roy, PP.; Paul, S.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive QSAR models. *Molecules* 2009, 14, 1660-1701.

[35] Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H. Comparative studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.,* 2012, 52, 396-408.

[36] Ojha, P.K.; Mitra, I.; Das, R.N.; Roy, K. Further exploring r_m^2 metrics for validation of QSPR models dataset. Chemom. Intell. Lab. Sys., **2011**, 107, 194-205.

[37] Mitra, I.; Roy, P.P.; Kar, S.; Ojha, P.; Roy, K. On further application of r_m^2 as a metric for validation of QSAR models. *J. Chemometrics* **2010**, *24*(1), 22-33.

[38] Besalu, E.; de Julia´n-Ortiz, J.V.; Pogliani, L. Trends and Plot Methods in MLR Studies. *J. Chem. Inf. Model.* **2007**, 47, 751-760.

[39] Roy, K.; Paul, S. Docking and 3D QSAR studies of protoporphyrinogen oxidase inhibitor 3H-pyrazolo[3,4-d][1,2,3]triazin-4-one derivatives. *J. Mol. Model.,* 2009, *16*(1), 137-153.

[40] Mitra, I.; Roy, K.; Saha, A. QSAR of antilipid peroxidative activity of substituted benzodioxoles using chemometric tools. *J. Comput. Chem.,* **2009**, *30*(16), 2712- 2722.

[41] Roy, K.; Roy, P.P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA- MLR, PLS, GFA, G/PLS and ANN techniques. *Eur. J. Med. Chem.,* **2009**, *44*(7): 2913-2922.

[42] Roy, K.; Mitra, I.; Saha, A. Molecular shape analysis of antioxidant and squalene synthase inhibitory activities of aromatic tetrahydro-1,4-oxazine derivatives. Chem. Biol. Drug. Des. **2009**, *74*(5), 507-516.

[43] Roy, K.; Popelier, P.L.A. Exploring predictive QSAR models for hepatocyte toxicity of phenols using QTMS descriptors. Bioorg. Med. Chem. Lett. **2008**, *18*(8), 2604-2609.

[44] Roy, K.; Popelier, P.L.A. Exploring predictive qsar models using quantum topological molecular similarity (QTMS) descriptors for toxicity of nitroaromatics to *Saccharomyces cerevisiae*. QSAR Comb. Sci., **2008**, *27*(8), 1006-1012.

[45] Roy, K.; Popelier, P.L.A. Predictive QSPR modeling of acidic dissociation constant (pKa) of phenols in different solvents. *J. Phys. Org. Chem.,* **2009**, *22*(3), 186-196.

[46] Kar, S.; Roy, K. QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors. *J. Hazard. Mater.,* **2010**, *177*(1-3): 344-351.

[47] Kar, S.; Harding, A.P.; Roy, K.; Popelier, P.L.A. QSAR with Quantum Topological Molecular Similarity Indices: Toxicity of Aromatic Aldehydes to Tetrahymena pyriformis. *SAR QSAR Environ. Res* **2010**, *21*(1-2), 149-168.

[48] Kar, S.; Roy, K. First report on predictive chemometric modeling, 3D- toxicophore mapping and in silico screening of in vitro basal cytotoxicity of diverse organic chemicals. *Toxicol. in vitro.,* **2013**, *27*, 597-608.

[49] Kar, S.; Deeb, O.; Roy, K. Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor. *Ecotox. Environ. Saf.,* **2012**, *82*, 85-95.

[50] Kar, S.; Roy, K. First report on development of quantitative interspecies structure-carcinogenicity relationship models and exploring discriminatory features for rodent carcinogenicity of diverse organic chemicals using OECD guidelines. *Chemosphere* **2012**, *87*, 339-355.

[51] Hu, R.; Doucet, J.P.; Delamar, M.; Zhang, R. QSAR models for 2-amino-6-arylsulfonylbenzonitriles and congeners HIV-1 reverse transcriptase inhibitors based on linear and nonlinear regression methods. *Eur. J. Med Chem.,* **2009**, *44*(5), 2158-2171.

[52] Toropov, A.A.; Toropova, A.P.; Benfenati, E. QSPR modeling bioconcentration factor (BCF) by balance of correlations. *Eur. J. Med. Chem.,* **2009**, *44*(6), 2544- 2551.

[53] Basu, A.; Jasu, K.; Jayaprakash, V.; Mishra, N.; Ojha, P.; Bhattacharya, S. Development of CoMFA and CoMSIA models of cytotoxicity data of anti-HIV-1- phenylamino-1H-imidazole derivatives. *Eur. J. Med. Chem.,* **2009**, *44*(6), 2400- 2407.

[54] Nargotra, A.; Koul, S.; Sharma, S.; Khan, I.A.; Kumar, A.; Thota, N.; Koul, J.L.; Taneja, S.C.; Qazi, G.N. Quantitative structure activity relationship (QSAR) of aryl alkenyl amides/imines for bacterial efflux pump inhibitors. *Eur. J. Med. Chem.,* **2009**, *44*(1), 229-238.

[55] Liao, S.Y.; Chen, J.C.; Miao, T.F.; Shen, Y.; Zheng, K.C. Binding conformations and QSAR of CA-4 analogs as tubulin inhibitors. *J. Enz. Inh. Med. Chem.,* 2010, *25*(3): 421-429.

[56] Toropov, A.A., Toropova, A.P., Benfenati, E., Leszczynska, D., Leszczynski, J. SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations. *J. Comput. Chem.,* **2010**, *31*(2), 381-392.

[57] Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. InChI-based optimal descriptors: QSAR analysis of fullerene [C60]-based HIV-1 PR inhibitors by correlation balance. *Eur. J. Med. Chem.,* **2010**, *45*(4), 1387-1394.

[58] Lu, P.; Wei, X.; Zhang, R. CoMFA and CoMSIA 3D-QSAR studies on quionolone caroxylic acid derivatives inhibitors of HIV-1 integrase. *Eur. J. Med. Chem.,* **2010**, *45*(8), 3413-3419.

[59] Dashtbozorgi Z, Golmohammadi H. Prediction of air to liver partition coefficient for volatile organic compounds using QSAR approaches. *Eur. J. Med. Chem.,* **2010**, *45*(6), 2182-2190.

[60] Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemometrics* **2010**, *24*(3-4), 194-201.

[61] Schu¨u¨rmann, G.; Ebert, R.U.; Chen, J.; Wang, B.; Ku¨hne, R. External validation and prediction employing the predictive squared correlation coefficient-Test-set activity mean vs training set activity mean. *J. Chem. Inf. Model.,* **2008**, *48*(11), 2140-2145.

[62] Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.,* **2009**, *49*(7), 1669-1678.

[63] Chirico, N.; Gramatica, P. Real External predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.,* **2011**, *51*, 2320-2335.

[64] Roy, K.; Mitra, I.; Ojha, P.K.; Kar, S.; Das, R.N.; Kabir, H. Introduction of $r_{m}^{2}{}_{(rank)}$ metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. *Chemom. Intell. Lab. Sys.,* **2012**, *118,* 200-210.

[65] Mitra, I.; Saha, A.; Roy, K. Exploring quantitative structure-activity relationship (QSAR) studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol. Simulat*., **2010**, *36,* 1067-1079.

[66] Walkera, J.D.; Carlsenb, L.; Jaworskac, J. Improving opportunities for regulatory acceptance of QSARs: the importance of model domain, uncertainty, validity and predictability. *QSAR Comb. Sci.,* **2003**, 22, 346-350.

[67] Gálvez-Llompart, M.; Recio, M.C.; García-Domenech, R. Topological virtual screening: a way to find new compounds active in ulcerative colitis by inhibiting NF-kB. *Mol. Divers*., **2011**, *15,* 917-926.

[68] Prado-Prado, F.J.; Uriarte, E.; Borges, F.; Gonza´lez-Dı´az, H. Multi-target spectral moments for QSAR and complex networks study of antibacterial drugs. *Eur. J. Med. Chem.,* **2009,** *44,* 4516-4521.

[69] Speck-Planche, A.; Kleandrova, V.V.; Luan, F.; Cordeiro, M.N.D.S. Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads. *Eur. J. Med. Chem.,* **2011**, *46*(12), 5910-5916.

[70] Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P.A.; Igglessi-Markopoulou, O.; Kollias, G. A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs. *Mol. Divers*., **2010,** *14*(2), 225-235.

[71] Fawcett, T. An introduction to ROC analysis. *Pattern. Recognit. Lett*., **2006**, 27, 861-874.

[72] Kubat, M.; Holte, R.; Matwin, S. Machine learning for the detection of oil spills in satellite radar images. *Mach. Learn.* **1998,** *30*, 195-215.

[73] Cohen, J.A. Coefficient of agreement for nominal scales. *Educ. Psychol. Meas.,* **1960**, *20*, 37-46.

[74] Perez-Garrido, A.; Helguera, A.M.; Borges, F.; Cordeiro, M.N.D.S.; Rivero, V.; Escudero, A.G. Two New Parameters Based on Distances in a Receiver Operating Characteristic Chart for the Selection of Classification Models. *J. Chem. Inf. Model.,* **2011**, *51*, 2746-2759.

[75] Truchon, J.F.; Bayly, C.I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model* **2007**, *47*, 488- 508.

[76] Zhao, W.; Hevener, K.E.; White SW, Lee RE, Boyett JM. A statistical framework to evaluate virtual screening. *BMC Bioinformatics* **2009**, *10*, 225.

[77] Sheridan RP, Singh SB, Fluder EM, Kearsley SK. Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*(5), 1395-406.

[78] Clark, R.D.; Webster-Clark, D.J. Managing bias in ROC curves. *J. Comput. Aid. Mol. Des.,* **2008**, *22*, 141-146.

[79] Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.,* **1975**, 405, 442-451

[80] Galvez, J.; Garcia-Domenech, R.; de Gregorio Alapont, C.; De Julian-Ortiz, V.; Popa, L. Pharmacological distribution diagrams: a tool for de novo drug design. *J. Mol. Graph.,* **1996,** *14,* 272-276.

[81] Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb Sci.,* **2007**, *26*(6), 694-701.

[82] Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *ATLA (Altern. Lab. Anim.)* **2005**, *33*(5), 445-459.

[83] Tetko, I.V.; Bruneau, P.; Mewes, H.W.; Rohrer, D.C.; Poda, G.I. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today* **2006**, *11,* 700-707.

[84] Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17,* 4791-4810.

[85] Netzeva, T.I.; Worth, A.P.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.; Perkins, R.; Roberts, D.W.; Schultz, T.W.; Stanton, D.T.; van de Sandt, J.J.M.; Tong, W.D.; Veith, G.D.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA (Altern. Lab. Anim.)* **2005**, *33,* 155-173.

[86] Jaworska, J.S.; Nikolova-Jeliazkova, N.; Aldenberg, T. Review of methods for applicability domain estimation. Report, The European Commission-Joint Research Centre, Ispra (Italy), **2004**.

[87] Weaver, S.; Gleeson, M.P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.,* **2008,** *26*(8), 1315-1326.

[88] Jaworska, J.; Comber, M.; Van Leeuwen, C.; Auer, C. Summary of the workshop on regulatory acceptance of QSARs. *Environ. Health. Perspect.,* **2003**, *111*(10), 1358-1360.

[89] Seber, G.A.F. *Multivariate Observations*. John Wiley & Sons: New York, **1984**.

[90] Nikolova-Jeliazkova, N.; Jaworska, J. An approach to determining applicability domain for QSAR group contribution models: an analysis of SRC KOWWIN. *ATLA (Altern. Lab. Anim.)* **2005**, *33*(5), 461-470.

[91] TOPKAT OPS. U.S. Patent 6, 036, 349, March 14, **2000**.

[92] Preparata, F.P.; Shamos, M.I. In: *Computational Geometry: An Introduction*. Preparata, F.P.; Shamos, M.I. Ed.; Springer-Verlag: New York, **1991**.

[93] Hair, J.F.(Jr.).; Anderson, R.E.; Tatham, R.L.; Black, W.C. *Multivariate Data Analysis*. Pearson Education: Singapore, **2005**.

[94] Silverman, B.W. *Density Estimation for Statistics and Data Analysis*, CRC Press: Boca Raton, **1986**.

[95] Stanforth, R.W., Kolossov, E., Mirkin, B. A measure of domain of applicability for QSAR modeling based on intelligent K-means clustering. QSAR Comb. Sci. **2007**, *26*(7), 837-844.

[96] Guha, R.; Jurs, P.C. Determining the validity of a QSAR model-a classification approach. *J. Chem. Inf. Model.,* **2005**, *45*(1), 65-73.

CHAPTER 4

Density Functional Studies of Bis-alkylating Nitrogen Mustards

Pradip Kr. Bhattacharyya*, Sourab Sinha, Nabajit Sarmah and Bhabesh Chandra Deka

Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India

Abstract: Nitrogen mustards are the most extensively used chemotherapeutic agent since their evolution in the mid-1940s. The high degree of cytotoxicity of these drugs is attributed to their ability to form DNA interstrand cross-linked adducts, thereby inhibiting DNA replication. Interstrand cross-linking occurs *via* formation of an unstable intermediate, the *aziridinium ion* and formation of mono-adducts. Mustine, the first member of this family, suffers from some serious drawbacks such as high rate of hydrolysis. Therefore its stable analogs have been sought; and since its discovery hundreds of analogs have been synthesized.

This article presents a brief introduction to nitrogen mustards and deliberates on the works already devoted to establishing the mechanism of action of this class of drug. A brief discussion on DFT and DFRT is also furnished in section 1.2. Further, computational studies performed on nitrogen mustards are discussed in section 1.3 and 1.4. Section 1.4 of the article consists of research works from our group and has special reference to DFT and DFRT.

Keywords: Anticancer drug, aziridinium ion, bis-alkylating agent, cancer chemotherapy, chemical hardness, chemical potential, chlorambucil, computational chemistry, density functional theory (DFT), density functional reactivity theory (DFRT), DNA alkylation, external electric field, maximum hardness principle, minimum electrophilicity principle, melphalan, Mustine, nitrogen mustards, reactivity descriptors, steroid-linked nitrogen mustard, structural variation.

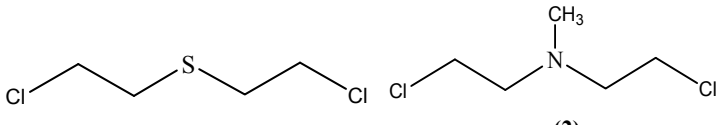
INTRODUCTION TO NITROGEN MUSTARDS

The world has always witnessed some great accidental discoveries. Evolution of nitrogen mustards as chemotherapeutic agent is a perfect example of such miracle.

***Corresponding author Pradip Kr. Bhattacharyya:** Department of Chemistry, Arya Vidyapeeth College, Assam-781016, India; Tel: +919864087494; Fax: +913612474065; E-mail: prdpbhatta@yahoo.com

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

The mustard gas (sulphur mustard, (**1**)), initially used as a chemical weapon, was responsible for thousands of deaths during World War I [1]. In a subsequent development, Louis S. Goodman and Alfred Gilman were recruited by the United states Department of Defence to investigate the potential therapeutic application of this chemical warfare agent. They observed that sulphur mustard is too volatile to be used in laboratory experiments, so they replaced the sulphur atom with nitrogen and synthesized a more stable product, mustine (**2**), the first member of the family called '*nitrogen mustard'*.



(1) **(2)**

It was ostensibly one of the greatest discoveries in the field of medicine that ushered in the advent of cancer chemotherapy. Mustine, also called mechlorethamine or mustragen (**2**), is the earliest member of the nitrogen mustard family and is still being used as a potent anticancer drug [2]. A number of drugs of this family have been synthesized and their antitumor activity has been analyzed.

Though, the mode of action of both sulphur and nitrogen mustards are similar, due to highly volatile nature of sulphur mustard, it has been thrown out of the frame.

Nitrogen mustard is termed as *biological alkylator* or an *alkylating agent.* According to Ross, an *alkylating agent* is a chemical compound that can replace a hydrogen atom in a molecule by an alkyl group [3]. In Oncology, the term *alkylating agent* refers to any antineoplastic compound that irreversibly binds (by formation of covalent bonds) to a variety of susceptible biomolecules such as nucleic acids, proteins *etc*. The simplest result of covalent bond formation with DNA is the cell death.

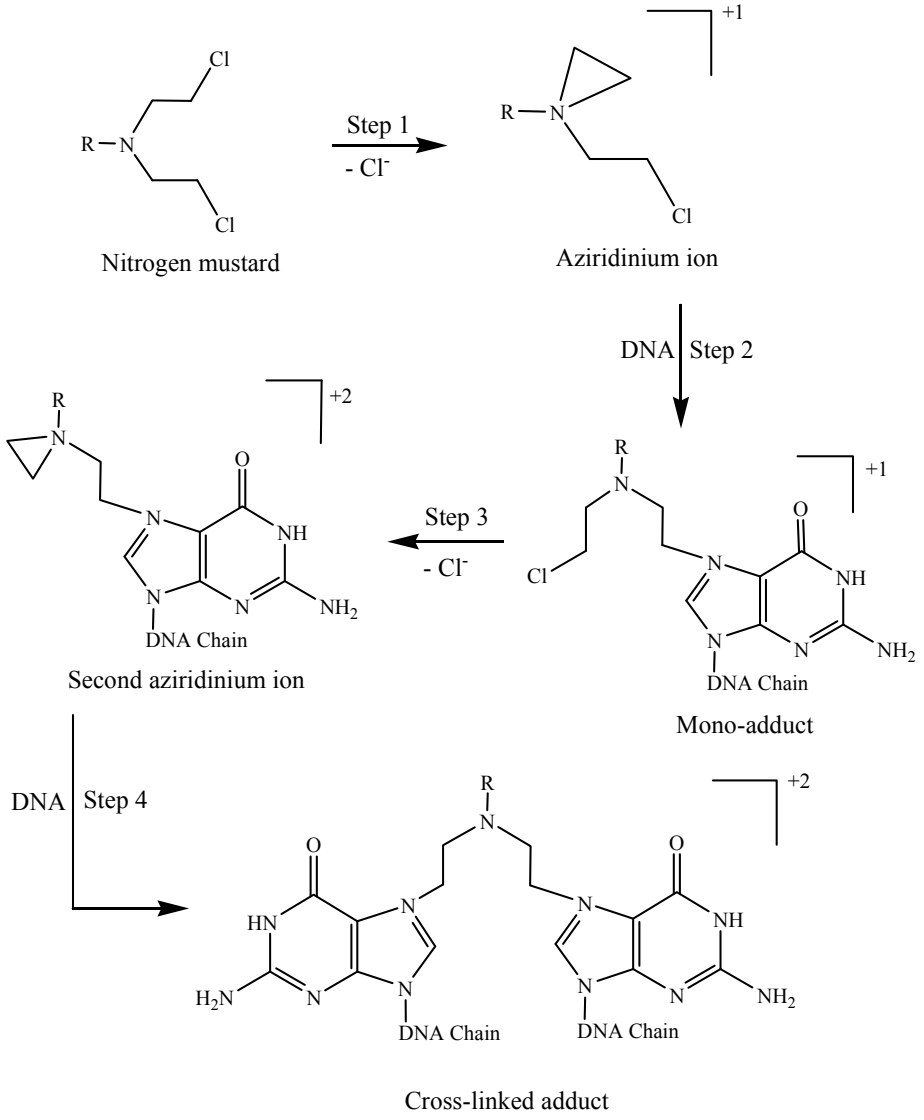
Establishing the mechanism of action of nitrogen mustards has been a formidable challenge for the researchers. A plethora of attempts have been made to understand the mechanism of action of these drugs, the first been made by Chanutin and Gjessing in 1946 [4]. They used UV-spectroscopy to show that

mustine reacts with DNA bases, but could not specify the group reacting with. Nucleic acids contain various centers which are vulnerable to attack by nitrogen mustards under physiological condition. On analyzing the product between the mustard gas and nucleic acid, Elmore *et al.* [5] confirmed that both basic and phosphate groups are prone to attack. But a conclusion regarding specific site of alkylation still eluded. This was followed by a number of studies to clarify the alkylation site; [6-10]; Wheeler and Skipper performed experimental studies on *in-vivo* alkylation of DNA [8]. Brookes and Lawley studied alkylation of ^{35}S-labelled mustard gas with nucleic acids in neutral aqueous solution at physiological temperature (37 °C) [11] and confirmed the reaction of mustard gas with DNA and RNA bases.

A number of studies have confirmed that alkylation of DNA by nitrogen mustards passes through the formation of aziridinium ion [12-14]. The first step of the alkylation reaction involves donation of a lone pair of electron from the N-center of the nitrogen mustard to the chloroethyl side chain with the release of a chloride ion (Cl^-), forming an aziridinium (Az^+) ion (step 1, **Scheme 1**). Being a positively charged species, the Az^+ ion is highly unstable and reacts immediately with the nucleophilic centers in biomolecules, (attack at guanine N7 is shown in step 2, **Scheme 1**) [15-16]. This leads to the formation of a drug-DNA mono-adduct. The mono-adduct further cyclizes to form a second aziridinium (Az^{2+}) ion (step 3, **Scheme 1**) which can bind to a second DNA strand resulting in a drug-DNA cross-linked adduct (step 4, **Scheme 1**).

Usually, all the heteroatoms in DNA/RNA bases exhibit higher tendency to interact with electrophiles like aziridinium ion [17-18]. Among different nucleophilic centers present in DNA bases, N1, N3 and N7 in adenine, N1, N3 and O2 in cytosine, N3 and O4 in thymine, N3, N7 and O6 in guanine are some of the highly preferred sites for alkylation. However, alkylation at these reactive sites depends on the reaction conditions [19-26]. It is expected that, due to steric hindrance, access to some sites in a double-stranded DNA is limited as compared to a single stranded DNA. Factors determining the selectivity of an alkylating drug towards DNA bases are very complicated [27-28]. The order of reactivity of the nucleophilic sites of the bases towards alkylating agent is: guanine N7 >

adenine N1 > cytosine N1 > adenine N3 in RNA and guanine N7 > adenine N3 > cytosine N1 in DNA [29].



Scheme 1: Mechanism of DNA alkylation by nitrogen mustards.

Lawley and Brookes [30] confirmed that cytotoxic action of sulphur and nitrogen mustards is not associated with the inhibition of cell growth as measured by RNA and protein synthesis but is caused due to interstrand cross-linking in DNA.

Cytotoxicity of nitrogen mustards is associated with decreased DNA synthesis because of DNA interstrand cross-linking, which prevents DNA replication. Interestingly, intensity of cross-linking of DNA by nitrogen mustards depends upon the size of the DNA fragment reacting with. Cross-linking by nitrogen mustards takes place through bis-alkylation at guanine N7 in right handed B DNA in 5′-GC-3′ or 3′CG-5′ sequence.

Inspite of the substantial amount of research works devoted to the understanding of alkylation of DNA, prior to work of Mattes *et al.* [31], DNA sequence selectivity towards guanine N7 was not properly understood. The group attached to N atom in nitrogen mustards affects the reactivity of aziridinium ion. This in turn might have some influence on sequence selectivity [32].

The dynamics of DNA alkylation and reactivity of different nucleic acids towards alkylation depends on different factors. Reactivity of different sites of guanine depends on the type of bases it is surrounded with. Compared to an isolated guanine, an enhanced rate of the reaction was observed for guanine N7 when it is surrounded by other guanines. The reaction of the drugs with isolated DNA is completely different from that in intact cells. The cellular environment also plays an important role during the alkylation reaction although it does not alter the sequence specificity. Limited diffusion of the drug molecule into the cell, alkylation reaction to other cellular components, *etc*. are some of the factors which determine the extent of alkylation in intact cells and in isolated DNA [33]. Pullman *et al.* [34] studied the effect of adjacent base pairs on molecular electrostatic potential (MEP) in the vicinity of various positions in B-DNA for both single and double stranded DNA. Another important point to note is that, both aliphatic and aromatic nitrogen mustards show similar sequence selectivity.

Price *et al.* [35] studied the reactivity of mono-functional nitrogen mustards (having only one chloroethyl chain) and observed that this class of nitrogen mustards was not much effective in preventing DNA replication due to its inability to form cross-linked adducts.

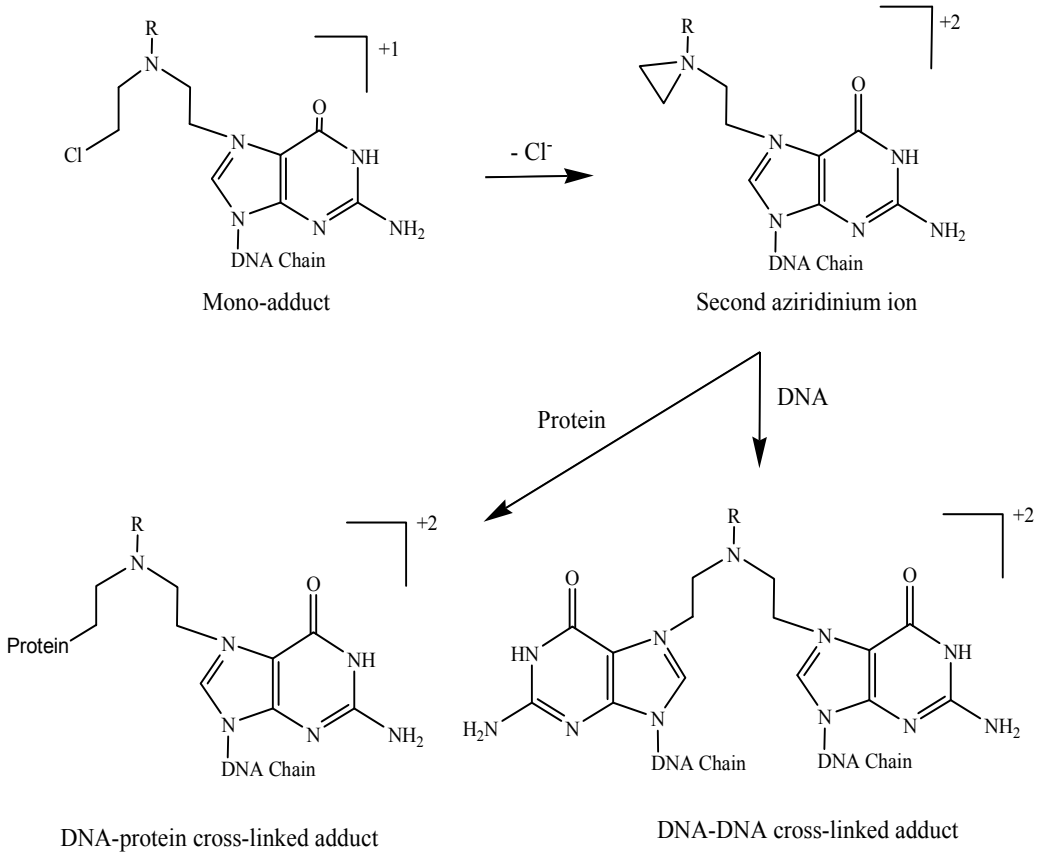
Nitrogen mustards form interstrand as well as intrastrand cross-linked adducts with DNA. However, formation of interstrand cross-linked adducts is favored

over intrastrand adducts. This is because, the distance between the two chlorine atoms in nitrogen mustards is ~7 Å and thus it becomes difficult for the second aziridinium ion to alkylate a second guanine of the same strand which is separated from the first by a distance of 8.9 Å in a B-DNA. This makes intrastrand cross-linking more difficult and thus mono-adduct happens to be the major portion (~90%) of the alkylated products [36]. The cytotoxicity of nitrogen mustards is proved to be allied with its ability to form interstrand cross-linked adducts that prohibits DNA replication and transcription and ultimately leads to cell death [37, 38]. However, a mono-adduct is found to be less cytotoxic compared to cross-linked adducts.

Apart from DNA-DNA intra- and interstrand cross-linking, there is another possibility of formation of a cross-linked adduct *i.e.* the DNA-protein cross-linked adduct, **Scheme 2**. A number of studies have confirmed the formation of such DNA-Protein cross-linked adducts [39-42].

The mystery behind the mode of action of the drug has been unfolded but then another controversy crops up: what leads to cytotoxicity of nitrogen mustards: intra- or interstrand cross-linking? Kohn tried to provide an answer to this question. He suggested that out of the different lesions formed during alkylation of DNA, interstrand cross-links are presumed to be particularly cytotoxic [43]. His assumption was based on the fact that the bis-functional drugs are more cytotoxic compared to mono-functional drugs [44]. Moreover, potency to form cross-linked adduct is not the sole criterion for cytotoxicity of a drug; instead, low rate of repairing of damaged DNA is also responsible for cytotoxicity. Thus, cytotoxicity of a drug molecule is affected by DNA repair [45]. In cultured cells, DNA interstrand cross-links comprise of 30-40% of the total DNA cross-links (excluding intrastrand cross-links), while the remaining constitutes DNA-protein cross-links [46].

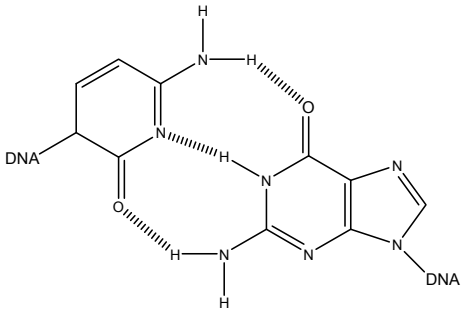
Cytotoxicity of a drug depends on the rate of removal of the cross-linked adducts which varies from drug to drug. For example, the rate of removal of cross-linked adducts for melphalan is higher than that of mustine [46]. Moreover, relationship between DNA damage and repair of damaged DNA is also an important factor.



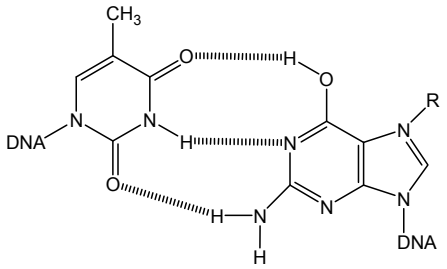
Scheme 2: DNA-protein cross-linking.

Consequences of DNA Alkylation

Under physiological conditions, keto-form of guanine is more stable than its enol-form. An important consequence of guanine alkylation is the formation of ammonium ion which makes guanine more acidic and hence shifts the keto-enol equilibrium towards the enol-tautomer [47, 48]. Now a different chemistry starts emerging in hydrogen bond formations between guanine and other bases. In its usual physiological form, (keto-form) guanine pairs up with cytosine forming three hydrogen bonds (**3a**). However, in its enol-form, guanine is not in a suitable position to pair up with cytosine, instead it has now gained the structure which is very much suitable for pairing up with thymine (**3b**). Thus alkylation of DNA at guanine N7 leads to the miscoding in DNA.



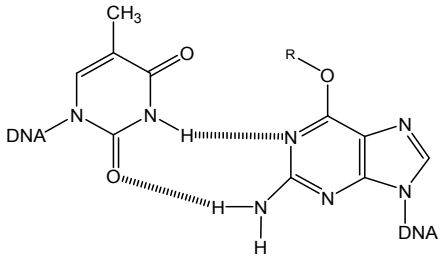
(**a**) Guanine (keto)-cytosine pairing



(**b**) Guanine (enol)-thymine pairing

(**3**) Miscoding due to alkylation at guanine N7

Similar miscoding in DNA results when alkylation occurs at the guanine O6; it alters the normal hydrogen bonding between guanine-cytosine base pair and results in the formation of guanine-thymine base pair (**4**) [49-53].



(**4**) Miscoding due to alkylation at guanine O6 (guanine-thymine pairing)

As a result of alkylation at the endocyclic nitrogen atoms in DNA bases by aziridinium ions or other electrophiles, the nucleobases acquire a positive charge and result in destabilization. For example, alkylation at guanine N3 or N7, adenine N1, N3 or N7 and cytosine N3 results in unstable lesions. To neutralize the additional formal charge(s) imparted to the nucleobases, they undergo further reaction. There are three types of reactions involved in decomposition of alkylated bases [54]:

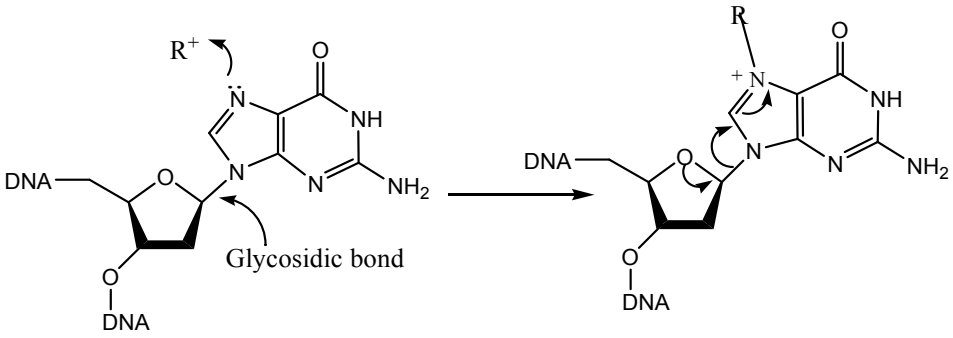
a) Deglycosylation or depurination or depyrimidination: it involves hydrolytic loss of the alkylated base from the DNA backbone.

b) Ring opening: opening of pyridine ring.

c) Reverse alkylation: loss of the alkyl group from the base.

Deglycosylation (Depurination or Depyrimidination)

Under normal physiological conditions, the glycosidic bonds holding the DNA base to sugar-phosphate backbone are quite stable and resistant to hydrolysis [55]. Breaking of these glycosidic bonds is referred to as *deglycosylation* (**5**). Under physiological conditions, the rate of deglycosylation of cytosine and thymine is extremely slow (rate constant=1.5×10-12 s^{-1}, half life=14,700 years), and in case of guanine and adenine it is slightly faster, (rate constant = 3.0×10-11 s^{-1}, half life = 730 years). In contrast, the rate of deglycosylation is quite high in alkylated bases. The degree of destabilization depends on the site of alkylation as well as on the nature of the attacking electrophiles and varies among the bases. For example, the half lives for deglycosylation of bases (with simple alkyl groups) are: N7 of dA is 3 h, N3 of dA is 24 h, N7 of dG is 150 h, O2 of dC is 750 h; for O2 of dT and N3 of dC, the half lives are slightly higher: 6300 h and 7700 h respectively [54-58].

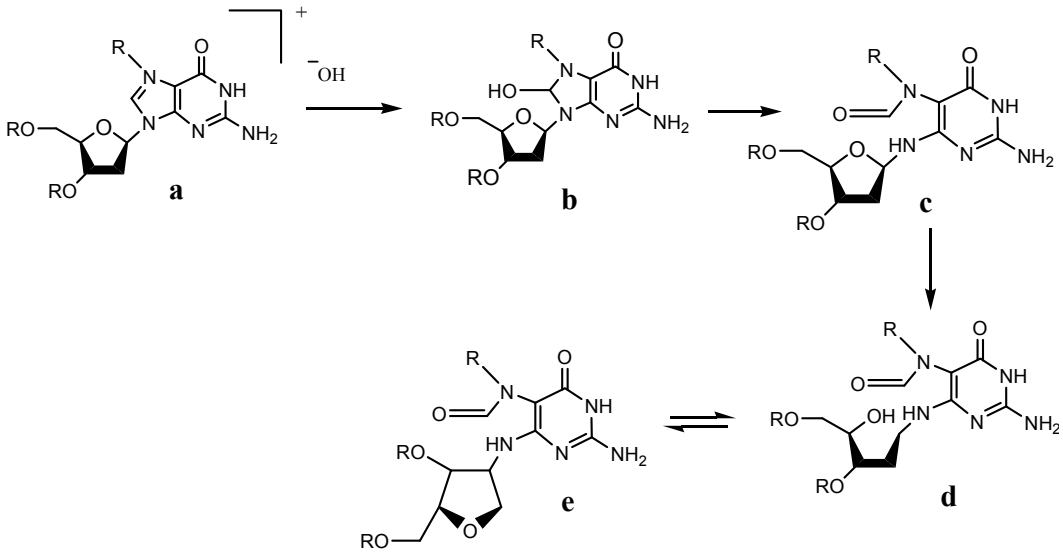


(**5**) Deglycosylation

In alkylated DNA, attachment of an electrophile (*e.g.* aziridinium ion) to the guanine N7 increases the electrophilicity at the adjacent positions. As a result of this, leaving group ability of the alkylated base increases resulting in scission of the C-N bond (glycosidic bond) and because of this an extraordinary increase in deglycosylation rate is observed [54, 59-60]. However, in DNA duplex, the rate of deglycosylation is slower as compared to a single stranded DNA or monomeric nucleosides. Gates *et al.* [54] observed that the reaction is 50-100 times slower in DNA duplex compared to monomeric nucleosides.

Purine Ring Opening

Another possibility of degradation of alkylated DNA is through purine ring opening as illustrated in (**6**). As a result of guanine N7 alkylation, electrophilicity of the C8 center in the purine ring increases and this facilitates purine hydrolysis. Hydrolysis at the C8 position of the purine ring (**6a**) leads to an intermediate (**6c**), which undergoes further fragmentation. For example, attack on C8 position of alkylated guanine by hydroxide ion leads to the fragmentation of imidazole ring [61-64]. Intermediate (**6c**) further undergoes anomerization to afford intermediate **6d** and subsequent DNA strand scission.



(**6**) Ring opening in alkylated guanine

Ring opening mechanism in case of N1 alkylated adenine is different from that of N7 alkylated guanine. It passes through Dimroth rearrangement [65]. N3 alkylated adenine also undergoes ring opening under basic condition but at a slower rate as compared to deglycosylation [66]. Usually, under physiological conditions, the rate of ring opening is quite slow compared to deglycosylation [67]. Therefore, deglycosylation is expected to be primarily responsible for degradation of alkylated DNA.

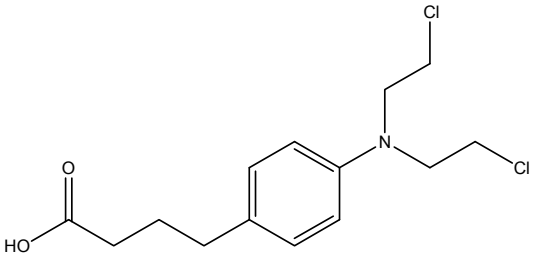
Reverse Alkylation

The alkylated DNA may undergo reverse alkylation (loss of alkyl group). Reverse alkylation has in fact been observed for some alkylating drugs (not in case of nitrogen mustards). Especially few drug molecules, such as CC-1065 (alkylate adenine N3), duocarmycin (alkylate adenine N3), leinamycin (alkylate guanine N7) alkylate DNA to show subsequent reverse alkylation [68-70].

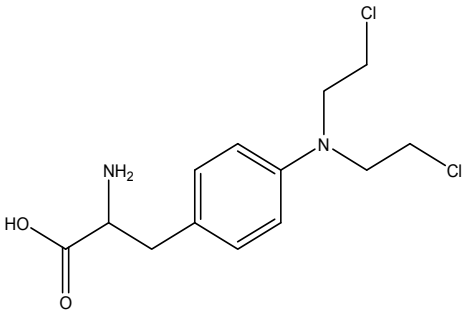
Nitrogen Mustard Derivatives

Mustine suffers from some demerits that arise due to highly reactive nature of the aziridinium ion (Az^+) produced by it. Because of its highly reactive nature, it is very prone to hydrolysis and reacts immediately with the nucleophilic centers in biomolecules. Because of this it is marketed as a dry solid and just prior to injection, its aqueous solution is prepared. Therefore, more stable analogs of mustine were looked for. Substitution of the methyl group (which is not possible in case of sulfur mustard) on the N-atom of mustine by aryl groups makes the N-atom less nucleophilic and slows down the rate of Az^+ ion formation [32]. This lowers the reactivity of the nitrogen mustards and, as a result of this stabilization, some of the drugs can be administered orally. Again, compounds prepared by simple aryl substitution at the N-center of mustine are water insoluble and hence are not suitable to be used as drugs. However, carboxyl and/or amine-containing aryl substituted mustine are water soluble and are not as reactive to water as mustine. Examples of such drugs are chlorambucil (**7**) and melphalan (**8**) *etc*. The carboxyl group in these drugs is not connected to the phenyl ring directly so as to make them more reactive or otherwise the lone pair present at the N-center may delocalize. To avoid this, methylene groups are inserted in between phenyl and carboxylic groups. Some other examples are phosphoramide mustard (**9**), uracil mustard (**10**), quinacrine mustard (**11**), bendamustine (**12**) *etc*.

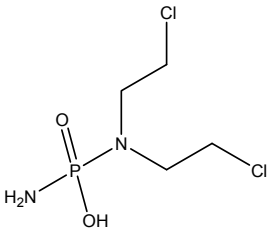
Number of attempts has been made in the past few decades which yielded many successful outcomes. As for example, Ross and coworkers [71] synthesized four aromatic nitrogen mustards (**13**) (with n = 0-4) containing carboxylic substituted derivatives of *N,N*-di-2-chloroethylaniline. Robert *et al.* [72] prepared cinchophen derivatives of nitrogen mustard (**14**).


Cl
N
Cl
O
HO

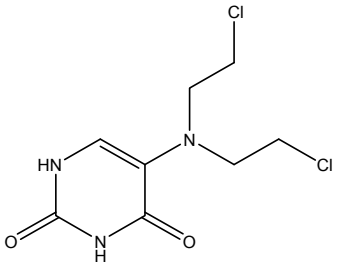
(**7**)


Cl
NH2
N
Cl
HO
O

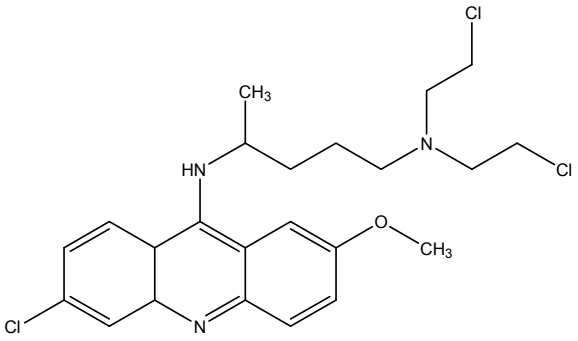
(**8**)


Cl
O
N
Cl
H2N
P
OH

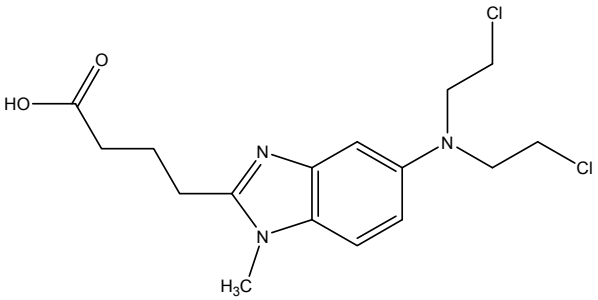
(**9**)


Cl
N
Cl
HN
O
N
H
O

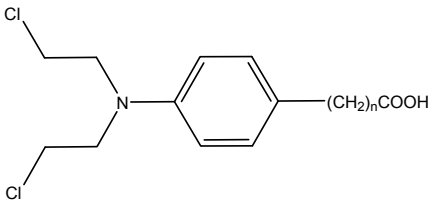
(**10**)


Cl
CH3
HN
N
Cl
O
CH3
Cl
N

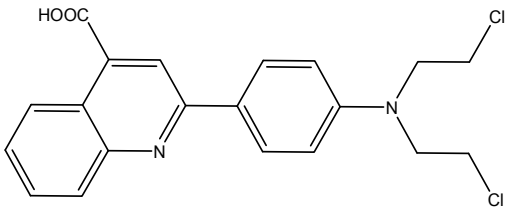
(**11**)


Cl
O
HO
N
N
Cl
N
H3C

(**12**)



(**13**)



(**14**)

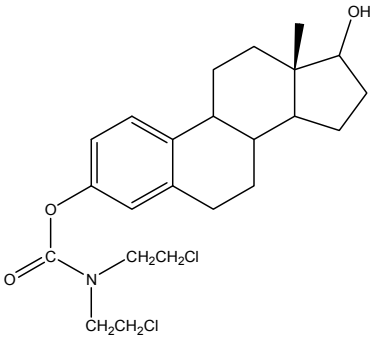
Attempt were also made to introduce iodo- group instead of chloro- group in chlorambucil, however it was found to be less active than chlorambucil against the Walker tumour and Sarcoma 180 [73]. Creech *et al.* [74] made an important study and tested the activity of 140 compounds, including mustine, chloroquine mustard, camoquine mustard, quinacrine mustard, nitromin, uracil mustard, chlorambucil, melphalan, cyclophosphamide, nor-mustard, cinchophen mustard, and numerous new analogs of nitrogen mustard. Interestingly, most of the quinoline and acridine analogs were observed to be highly active from the stand point of low molar dosage. Derivatives of chlorambucil and melphalan were also synthesized by incorporating into several peptide hormones, including luteinizing hormone-releasing hormone (LH-RH) [75].

After the discovery of mercaptopurine, which has very high antitumour activity, a number of attempts have been made to synthesize its derivatives [76]. Substitution at 1-, 3- or 7- positions of mercaptopurine exhibited no significant activity [77-79]. However, substitution at the 9-position imparts anticancer activity to the molecule [80, 81].

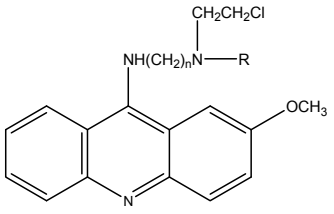
Other important chemotherapeutic agent that has been successfully applied against cancer includes uracil mustard (**10**) [82-85] and its derivatives [86-88]; and estramustine (**15**), used in the treatment of advanced prostatic carcinoma [89].

Intercalating drugs are another important class of drugs that have been used successfully in cancer chemotherapy. These drugs exhibit their cytotoxicity by

intercalating between the DNA base pairs. Intercalating ability of 9-amino acridine (**16**) is well established and a number of nitrogen mustard analogs have been synthesized by incorporating the intercalating chromophore [90, 91].



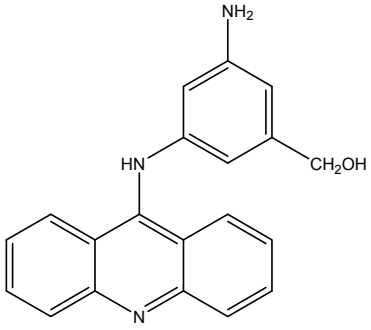
(**15**)



(**16**)

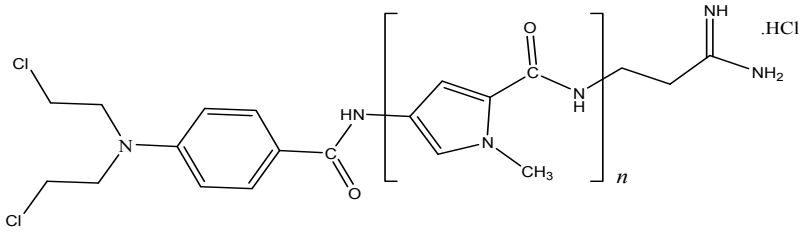
The mechanism of action of these drugs consists of two steps. Initially, the drug intercalates between the DNA base pairs and then alkylation takes place. Such a two step reaction (intercalation followed by alkylation) results in 10-100 fold lower concentration of acridine mustard required to alkylate DNA compared to mustine. Aniline mustard analog, m-AMSA (**17**) was prepared by linking a mustard residue to the aniline ring or acridine chromophore [92], but the derivatives did not show any enhanced activity as compared to the parent compound, AMSA. However, introduction of a short spacer between the nitrogen mustard and 9-anilinoacridine increases the reactivity [93]. 9-anilinoacridine and acridine derivatives, bearing nitrogen mustard residue at C4 of the acridine chromophore, were found to possess potent cytotoxicity against human leukemia and various solid tumors *in-vitro* [94]. To improve the chemical stability and therapeutic efficiency, efforts have been made to synthesize aniline nitrogen mustards linked to 9-anilinoacridines *via* urea linkage [95]. Derivatives with the nitrogen mustard residue linked to the C3′ or C4′ position of the anilino ring with an O-ethylene (O-C2), O-butylene (O-C4), and methylene (C1) spacer were prepared. A few of them were reported to possess approximately 100-fold more potency than its parent analog AHMA [96]. Series of compounds were also synthesized with nitrogen mustard pharmacophore on both anilino (C3′ or C4′)

and acridine (C4) rings with O-ethyl (O-C2) or O-butyl (O-C4) spacer and these compounds exhibited significant *in-vitro* cytotoxicity [97].

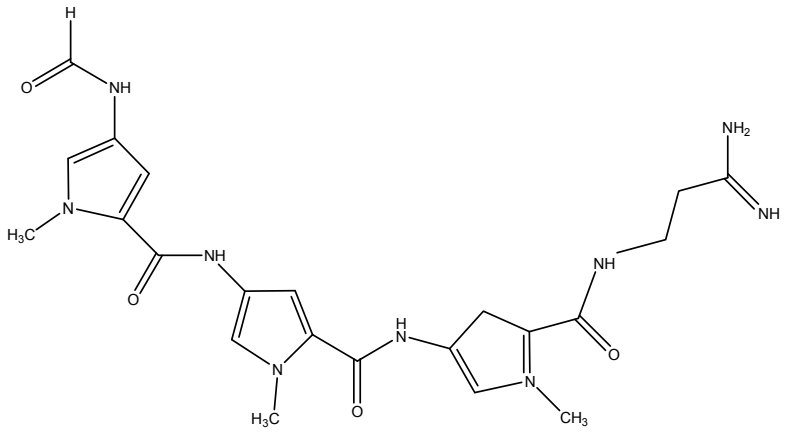


(**17**)

Tallimustine (**18**), a benzoic acid nitrogen mustard derivative of distamycin A (**19**) is also an effective antitumor drug that has been used as an important model for designing new nitrogen mustards containing pyrrole-amide unit [98].

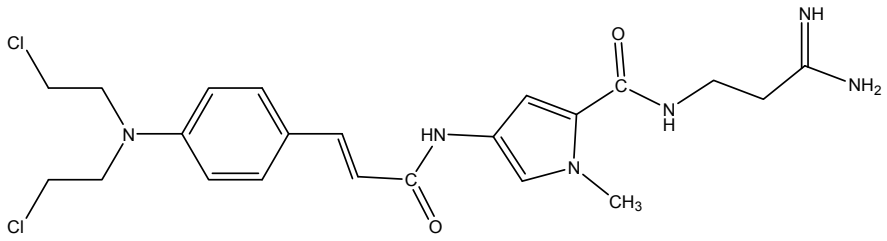


(**18**)



(**19**)

Tallimustine consists of three pyrrole-amide units (n = 3, **18**). Cytotoxicity of benzoic acid mustard conjugated with pyrrole ring was examined [99]. The di-pyrrole and tri-pyrrole conjugates did not produce any detectable guanine N7 alkylation but only alkylate AT tracts. Baraldi *et al.* [100] synthesized benzoic acid mustard (BAM) derivatives of distamycin A, bearing one or more pyrazole rings and tested their *in-vitro* and *in-vivo* activities against L1210 leukemia. Some of these derivatives showed activity comparable to tallimustine. All the compounds bearing the pyrazole ring close to the BAM moiety showed reduced cytotoxicity in comparison to derivatives characterized by the BAM linked to a pyrrole ring. No such effects were observed when conjugation was made to the amidine terminus of the oligopeptidic frame. Benzoheterocyclic analogs of tallimustine were not showing any enhanced activity [101].



(**20**)

PNU-157911 (**20**) is one of the most important nitrogen mustard derivatives. In an effort to examine the role of amidino group present in PNU-157911, a series of cinnamoyl nitrogen mustards (pyrazole analogs of tallimustine) were synthesized in which the amidino moiety was replaced by other moieties [102]. These modifications on the amidino moiety showed significant growth inhibitory activity against mouse leukemia L1210 cells. They also showed the capability to interact with DNA with sequence selectivity for certain AT-rich sequences. Compounds of this series possess a pattern of alkylation similar to that of tallimustine, but they apear to be less reactive. Therefore, presence of the amidino moiety or a basic moiety in general is not an absolute requirement for biological activity. *In-vitro* and *in-vivo* activities of novel benzoyl and cinnamoyl nitrogen mustard and half-mustard derivatives of distamycin A, in which the amidino moiety was replaced by moieties of different physico-chemical features, were also reported [103, 104]. As reported by Wang *et al.* [105], anticancer activity of

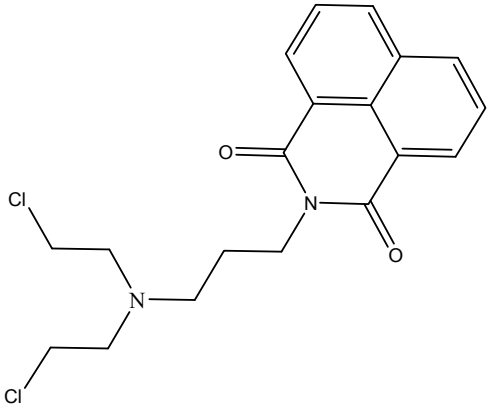
distamycin A and nitrogen mustard conjugates, in which the nitrogen mustard unit is coupled to the C-terminus of the pyrrole is observed to depend on the number of pyrrole rings; compound bearing three pyrroles is more potent than compounds bearing one or two pyrrole ring(s). Generally, it is observed that for distamycin nitrogen mustard, potency of the compound increases approximately 10-fold with the addition of one pyrrole unit up to a total of four pyrrole units. Interestingly, switching the nitrogen mustard unit from the N-terminus to the C-terminus of the pyrrole did not compromise with the cytotoxicity of the compounds. Distamycin nitrogen mustard derivatives with different substituents at the amidino moiety located at the C-terminal of the peptide were synthesized by Wang *et al.* [106]; compounds bearing a terminal ethylamido group show good antitumor activity against human chronic leukemia K562 cells.

Although the nitrogen mustard derivatives possess revolutionized anticancer activity, it also has got some serious drawbacks. These compounds are too polar to cross the highly lipophilic Blood Brain Barrier (BBB) and because of this, these drugs have very low brain penetration. Therefore, anticancer agents with highly hydrophobic nature are sought for the treatment of cerebral tumors. One of the most promisimg approaches to design CNS active anticancer drug is the prodrug approach. This approach is based on redox system, analogus to the $NADH \leftrightarrow NAD^+$ coenzyme system. In a novel attempt Sing *et al*. synthesized nicotinic nitrogen mustards that are hydrophilic in nature and can easily enter into brain and then are oxidized to quarternary salts which can not efflux from brain cells [107]. Based on redox prodrug approach a number of anticancer drugs have been synthesized [108-113].

Because of two important properties *viz*. preferential accumulation in neoplastic cells and inhibition of glycolysis, deoxygluocose, can also be attached to nitrogen mustard for the treatment of brain tumors. Successful attempts have been made to synthesize such derivatives by attaching the chlorambucil moiety to the glucose unit [114, 115]. Singh *et al.* prepared such a class of nitrogen mustard derivatives linked to CNS (central nervous system) active compound 1,4-benzodiazepine [116]. Efforts have also been made to decrease the lipophilic nature of chlorambucil and melphalan by preparing quaternary ammonium conjugates [117]. Though, quaternary ammonium functionalization does not alter the

cytotoxicity, it modifies the cell uptake by decreasing lipophilicity of the drug molecules.

Napromustine, (**21**) derivative of napthalimide, displays an excellent antitumor activity *in-vivo* against Sarcoma-180 and Ehrlich ascites carcinoma compared to that of fluorouracil. Other derivatives of Napromustine, such as nitro (mitonafide) and amine (amonafide), have also been tested, but no significant enhancement in cytotoxicity is observed [118].



(**21**)

Structurally modified carnitine analogs have been observed to show enhanced anticancer activity compared to chlorambucil against A375 human melanoma, HT29 resistant type colon carcinoma and MCF7 human breast carcinoma cells [119]. Cyclic nitrogen mustards, structurally related to L-carnitine are effective [120] and their trans-isomer produces better cytotoxicity than the cis-isomer.

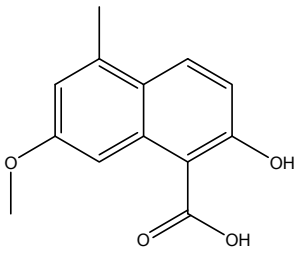
Neocarzinostatin (NCS), an antitumor antibiotic, is a protein-chromophore complex that exhibits cytotoxic action through DNA cleavage *via* H-abstraction [121]. Cytotoxicity of such drugs resides with the chromophore moiety alone, while the protein (apoNCS) protects and transports the labile chromophore. The naphthoate portion (**22**) of NCS chromophore is the most important site for binding to apoNCS and DNA intercalation. Attempts have been made to use apoNCS to improve the hydrolytic stability of the novel bis-functional DNA alkylating agents, melphalan and chlorambucil [121]. Though the melphalan conjugate (**23**) exhibits higher *in-vitro* cytotoxic activity against human leukaemia

cell line K562 than the unmodified melphalan, the inverse was observed in case of chlorambucil conjugate.

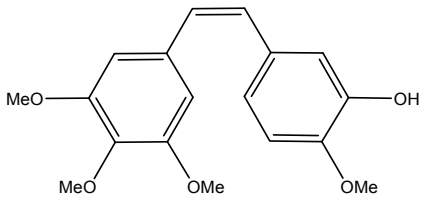
Another new series of anticancer drugs called chimeric compounds, bearing the combretastatin (**24**) and the nitrogen mustard moieties have been synthesized. Amazingly, when combretastatin is attached to chlorambucil *via* an ester linkage (**25**), the resultant compound proves to be significantly more potent than the combined potency of the individual drugs. However, when combretastatin is conjugated to nitrogen mustard *via* an ether linkage, loss in its potency is reported [122].

Benzoic acid derivatives are observed to be potential EGFR (Epidermal Growth Factor Receptor) and HER-2 (Human Epidermal Growth Factor Receptor 2) kinase inhibitor [123]. Especially, derivatives containing arylamine moiety show better inhibitory activity than those containing fatty amine moiety.

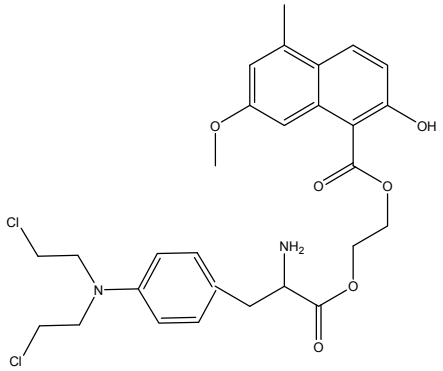
Melphalan and chlorambucil derivatives of 2,2,6,6-tetramethyl-1-piperidinyloxy radicals also show good potency against human leukaemia K562 cell line [124].



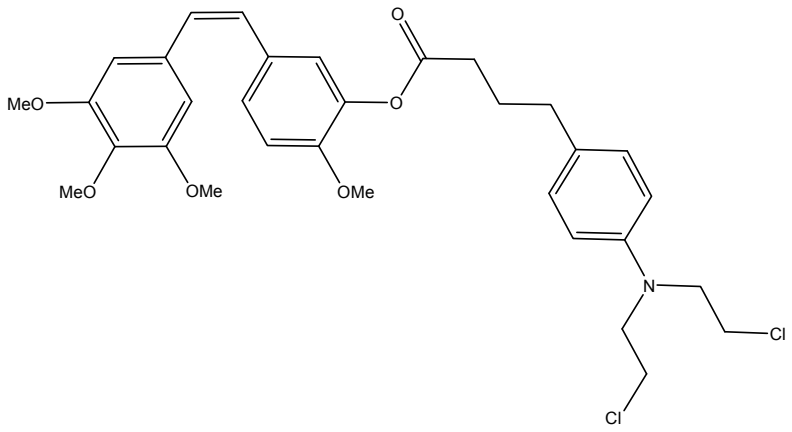
(**22**)



(**24**)



(**23**)



(**25**)

Nitrogen mustard derivatives of 4-anilinoquinazoline, where the nitrogen mustard pharmacophore is attached to the C6 of the 4-anilinoquinazolines *via* a urea linkage, are found effective against breast cancer [125]. Recently, aromatic ring of natural pyrimidine base has been replaced to obtain a thymine derivative of chlorambucil [126].

Aromatic bisamidines, such as berenil, pentamidine, propamidine and furamidine, are well-known antimicrobial and antifungal drugs [127-129]. Amino analogs of pentamidine with tetra- and hexa-methylene chain between aromatic units show antiproliferative activity against MCF7 (Michigan Cancer Foundation-7) breast cancer cell line of mammalian tumour and inhibitory influence on the activity of topoisomerase I and II [130]. Amino analogs of pentamidine with a polymethylene (n=3-6) chain and their chlorambucil analogs exhibit cytotoxic effect on MCF-7 human breast cancer cell line [131].

Because of their important biological and pharmacological properties, isoflavonoids have attracted considerable research interests for a long time. Formononetin, a type of isoflavonoids, has been reported to show many biological activities including antioxidant, antidiabetic, antiestrogenic, antibacterial, antiangiogenic effects and so on. Studies have shown that formononetin and its derivatives exhibit potent antiproliferative activities against two human tumor cells (Jurkat and HepG-2) *in-vitro* [132]. Recently, a series of formononetin nitrogen mustard derivatives have been synthesized and their cytotoxicity has

been evaluated *in-vitro* against five cancer cell lines (SH-SY5Y, HCT-116, DU-145, Hela and SGC-7901). Many of them display more potent cytotoxicity compared to melphalan [133].

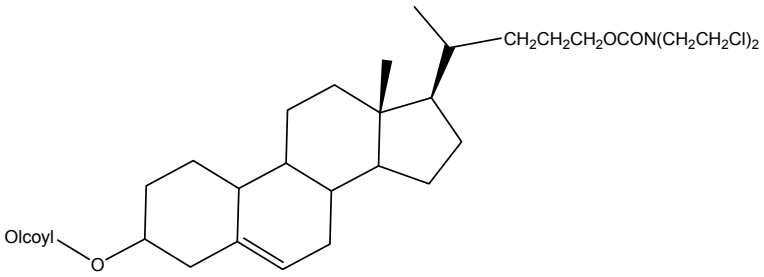
Recent findings show that the uptake of polyamine compounds such as amino acids (*e.g*. methionine) in cancer cells is high. Omoomi *et al.* [134] have synthesized chlorambucil-methionine conjugate and tested it against breast cancer MCF-7 cell line and observed high antineoplastic properties without any abnormal toxicity. The conjugate has also showed very good anticancer activity comparable to chlorambucil and less toxicity. Hence chlorambucil-methionine conjugate has been considered to be a better option for the treatment of breast cancer than Chlorambucil.

Further, in order to increase their antitumor potency and tumor selectivity, bendamustine and melphalan have been esterified with *N*-(2-hydroxyethyl) maleimide and connected by diamines with various chain lengths. Expectedly, the two new derivatives showed higher cytotoxicity compared to bendamustine and melphalan against breast cancer [135].

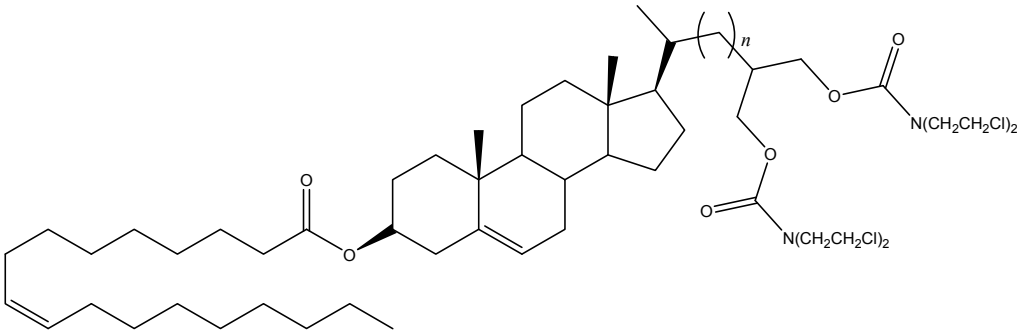
Studies of different hormone-linked antineoplastic agents reveal highly effective results in receptor positive tumors *in-vivo* [136]. Among these conjugates, the antitumor steroid hormone-nitrogen mustard combination is found to be quite successful [137]. It is expected that a lipophilic steroid carrier molecule would aid in the transport of the nitrogen mustard moiety to a specific target tissue more efficiently. Nitrogen mustard derivatives of androstane [138, 139], estrone [140] and estramustine phosphate [141-143] also exhibit high anticancer activity.

In order to deliver a cytotoxic molecule to the cancerous cells, the most efficient technique is to replace the cholesteryl ester core by a suitable lipophilic cytotoxic agent. This leads to the formation of cytotoxic-LDL (Low-density lipoprotein) particles (reconstituted LDL or rLDL) that targets cancerous cell. Several such cytotoxic compounds were synthesized and observed to be suitable for cancer treatment [144]. Another important task is to improve the potency of rLDL by increasing the number of cytotoxic molecules on each carrier without reducing its hydrophobicity. Dubowchik *et al.* [145] synthesized mono-, bis- and tris- nitrogen

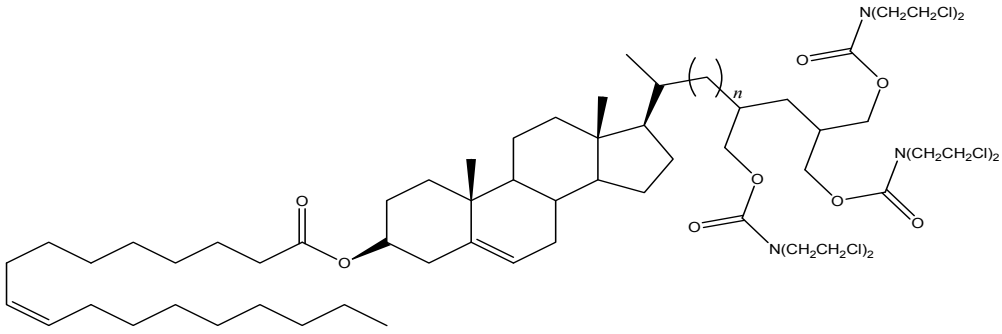
mustard derivatives of oleoyl-steroid carbamates (**26-28**), where the cytotoxic portion is attached away from the sterically congested steroid. The biological evaluation shows that the bis- nitrogen mustard (**27**) is twice as potent as mono- (**26**) and tris-mustard (**28**) [145].



(**26**)



(**27**) (n=1,3)



(**28**) (n=0,2)

Steroidal cyclophosphamide derivatives do not exhibit any cytotoxic effect [146, 147]. However, steroidal nitrogen mustard derivatives of phenol and aniline

mustards showed anticancer activity against several cancer cells [148]. Chlorambucil and 3-nitrochlorambucil esters of prasterone and pregnenolone were tested against several human cell lines [MCF-7 (ER+), MDA-MB-468 (ER−), MDA-NQ01, Widr, DAoy, H460, OVCA-3 and A375], out of which, the chlorambucil esters of prasterone displayed the highest anticancer activity [149]. Several estradiol-chlorambucil hybrids (the chlorambucil moiety was located at 16α position of the steroid) have been synthesized for site-directed chemotherapy in breast cancer [150]. Marquis *et al.* [151] designed a steroid-nitrogen mustard hybrid by linking nitrogen mustard moiety to a steroid.

Progesterone, when combined with chlorambucil, shows better biological results when tested among the rats. On the other hand, alkylating agent sensitive tumors do not respond well to the combination of melphalan and cyclophosphamide with prednisolone. Interestingly, a phase III trial assessing the ester of chlorambucil and prednisolone display better response rates on the treatment of patients with advanced breast cancer compared to the two independent drugs [152].

Arsenou *et al.* [153] synthesized a series of steroidal ester of para-[*N,N*-bis(2-chloroethyl)amino]phenylacetic acid (PHE) and investigated the influence of 7-carbonyl group in oxidized Δ^5-steroids on the antileukaemic activity. Their results showed that the steroidal part not only transports the nitrogen mustard moiety into the cells, but also participates directly in the mechanism of antileukaemic action in an unidentified fashion. They also synthesized a few steroidal derivatives of chlorambucil and evaluated their antileukaemic activity *in-vivo* (against P388 and L1210) and *in-vitro* on normal human lymphocytes [154]. Same results were observed as for the PHE derivatives which indicate that the lactam function on the B-steroidal ring led to potent hybrids but not as potent as the analogus 7-keto derivatives.

Steroidal hybrids possessing a lactam B-ring as well as 17β-amide have been synthesized; steroidal skeletons that carry a NHCO group are observed to be more powerful than the analogous unmodified steroids bearing the same nitrogen mustard moiety [155]. Steroidal ring A-lactam nitrogen mustard is found to be active against mouse leukaemia L1210 and mouse Sarcoma 180 [156].

Steroidal esters obtained by reducing the Δ5 double bond of steroidal ester of known hybrids containing the alkylating agent PHE or CHL (para-[N,N-bis(2-chloroethyl)amino]phenylbutyric acid) also show potential antineoplastic properties [157, 158]. Hecogenin and aza-homo-hecogenin steroids when linked to PHE show important anticancer activity against Lewis lung carcinoma [159, 160]. The activity of regioisomers of aza-homo-hecogenin has been observed to follow the order, ortho-> meta- > para- [161].

In summary, though earlier attempts to synthesize potent nitrogen mustards have been successful in overcoming the problem of drug resistance, the last few decades have witnessed a number of new nitrogen mustard derivatives synthesized by incorporating variety of groups in the N atom of mustine; incorporation of intercalating chromophore like 9-amino acridine has been particularly successful. Nitrogen mustard derivatives of distamycin, tallimustine are some other notably successful alternatives. CNS active chromophores have been successfully conjugated to nitrogen mustard to synthesize new candidates which can be used in cerebral tumour. In many cases conjugation of chlorambucil, melphalan and other nitrogen mustards to steroids have also resulted in new potent drugs. However, designing cell specific drugs still needs attention.

INTRODUCTION TO DFT AND DFRT

Density Functional Theory (DFT)

Computational chemistry is one of the most fascinating branches of chemistry that is useful in resolving many problems in chemistry. It comprises of a wide variety of methods developed over the last century. Density functional theory (DFT) is one of the most widely used computational chemistry methods and has become an increasingly popular tool for understanding many atomic-level intricate processes and its application is growing rapidly. It is a powerful, successful and promising approach to calculate molecular structures, total energy of the system, vibrational frequencies, atomization energies, ionization energies, electric and magnetic properties, reaction pathways, *etc* [162-165].

Formulation of the two most captivating theorems by Hohenburg and Kohn [166] in 1964 helped DFT to become a full-fledged theory. The first HK theorem [167]

states that the electron density, $\rho(\vec{r})$ determines the external potential (*i.e.* due to the nuclei), $v(\vec{r})$. $\rho(\vec{r})$ also determines *N*, the total number of electrons, *via* its normalization:

$$\int \rho(\vec{r})d\vec{r} = N \tag{1}$$

N and $v(\vec{r})$ determine the molecular Hamiltonian, $\hat{H}$, which in turn determines the energy of the system *via* the Schrodinger equation:

$$\hat{H}\psi = E\psi \tag{2}$$

ψ being the electronic wavefunction, $\rho(\vec{r})$ determines the system's energy and all other ground state electronic properties. The Hamiltonian operator, $\hat{H}$, involves the three energy factors: the kinetic energy, the energy due to interaction with the external potential (V_{ext}) and the electron-electron interaction energy (V_{ee}), *i.e.*:

$$\hat{H} = -\frac{1}{2}\sum_{i}^{N}\vec{\nabla}_i^2 + \hat{V}_{ext} + \sum_{i\rangle j}^{N}\frac{1}{\left|\vec{r}_i - \vec{r}_j\right|} \tag{3}$$

The second theorem restricts DFT to the studies of the ground states and establishes the variational principle. It states that for a fixed external potential $v(\vec{r})$, if there is any positive definite trial density, ρ_t coming from any *N*-electron wavefunction such that $\int \rho_t(\vec{r})d\vec{r} = N$, then it corresponds to a higher energy state compared to the ground state, *i.e.* $E[\rho_t] \geq E_0$.

The equation $E[\rho_t] = \langle \psi_t | \hat{H} | \psi_t \rangle \geq E_0$ follows immediately from the Variational theorem:

$$E[\psi] \geq E_0 \tag{4}$$

Trivial solution of the above equation and a good knowledge of the ground state wavefunction can specify the energy of the excited state, much orthogonal to the ground state.

The two theorems proposed by Hohenburg and Kohn led to the discovery of the fundamental equation of DFT:

$$\delta\left[E[\rho]-\mu\left(\int\rho(\vec{r})\,d\vec{r}-N\right)\right]=0 \tag{5}$$

The energy E does not change upon variation of optimal $\rho(\vec{r})$, provided that $\rho(\vec{r})$ integrates at all times to N (in equation 1). μ is the corresponding Lagrangian multiplier and is also known as the chemical potential of a system. μ can also be written as the partial derivative of the system's energy with respect to the number of electrons at fixed external potential $\nu(\vec{r})$:

$$\mu=\left(\frac{\partial E}{\partial N}\right)_{\nu(\mathrm{r})} \tag{6}$$

$E[\rho]$ in equation 5 is independent of the external potential for a particular system and can be inserted into the equation (only if its form is known) to obtain the exact energy and density of that particular system. Thus, in equation 2, we can define energy functional, $E[\rho]$ as the sum of three terms:

$$E[\rho]=\mathrm{T}[\rho]+V_{ext}[\rho]+V_{ee}[\rho] \tag{7}$$

where, $\mathrm{T}[\rho]$ is the kinetic energy, $V_{ext}[\rho]$ is the interaction with the external potential and $V_{ee}[\rho]$ is the electron-electron interaction.

The kinetic and electron-electron functionals are unknown and the interaction with the external potential is trivial:

$$V_{ext}[\rho]=\int\hat{V}_{ext}\rho(\vec{r})\,d\vec{r} \tag{8}$$

Kohn and Sham, [168] in 1965 have successfully come forward with their new formulation of replacing the kinetic energy functional $\mathrm{T}[\rho]$ of the interacting particles with that of non-interacting ones leading to the self-consistent Kohn-Sham (KS) equation which is described by a single determinant wavefunction in N orbitals, ϕ_i. KS formulation is structurally very much similar to that of Hartree-

Fock formulation; the only difference being that the non-local exchange potentials have been replaced by the local exchange-correlation potentials. Introduction of the variational orbitals has helped them to calculate the kinetic energy of the system with greater accuracy. Hence, we can determine the kinetic energy and electron density of a non-interacting system from the (variational) orbitals:

$$\mathrm{T}_s[\rho] = -\frac{1}{2}\sum_i^N \left\langle \phi_i \left| \vec{\nabla}^2 \right| \phi_i \right\rangle \tag{9}$$

where,
$$\rho(\vec{r}) = \sum_i^N |\phi_i| \tag{10}$$

$\rho(\vec{r})$ provides the solution to the exact ground state density of a system of a non-interacting electron. Taking classical Coulomb interaction into account, $V_H[\rho]$ in the electron-electron interaction, equation 7 can be rearranged to

$$E[\rho] = \mathrm{T}[\rho] + V_{ext}[\rho] + V_H[\rho] + E_{xc}[\rho] \tag{11}$$

where, $E_{xc}[\rho]$ is the exchange-correlation functional and is simply the sum of the error made in using a non-interacting kinetic energy and the error made in treating the electron-electron interaction classically.

Substituting the electron density of the energy functional (equation 11) in terms of ground state density of a system with non-interacting electrons (equation 10) and applying the variational theorem (equation 4), we find that the orbitals, which minimize the energy of a system, satisfy the following set of non-linear equation which describes the behaviour of non-interacting electrons in an effective local potential:

$$\left[-\frac{1}{2}\vec{\nabla}^2 + \nu(\vec{r}) + \int \frac{\rho(\vec{r})}{\left|\vec{r} - \vec{r}'\right|} d\vec{r}' + \nu_{xc}(\vec{r}) \right] \phi_i(\vec{r}) = \varepsilon_i \phi_i(\vec{r}) \tag{12}$$

If the exact energy functional is known, the orbitals yield the exact ground state density *via* equation 10 and exact ground state energy *via* equation 11.

The Local Density Approximation (LDA)

The exact form of the exchange and correlation energy represented by the fourth term of equation 11, $E_{xc}[\rho]$ for a system is very challenging to calculate and hence some approximations are necessary to solve it. The most trustworthy approximation is the local density approximation (LDA). This approximation helps in constructing approximate form of the exchange-correlation energy functional for an inhomogeneous electron gas from the knowledge of exchange-correlation energy of a homogeneous electron gas. As a result, the local approximations can also be considered synonymous with functionals.

The local exchange-correlation energy per electron may be approximated as a simple function of the local charge density (say $\varepsilon_{xc}(\rho)$), *i.e.* an approximation of the form:

$$E_{xc}[\rho] = \int \rho(\vec{r})\,\varepsilon_{xc}\left(\rho(\vec{r})\right)dr \tag{13}$$

where ε_{xc} is the exchange-correlation energy density and is a function of density alone. The E_{xc} term can be separated into E_x and E_c

$$E_{xc} = E_x + E_c \tag{14}$$

The solution to the first term of equation 14, *i.e.* E_x for a system with inhomogeneous electron density by applying the approximation results of a homogeneous electron gas gives us the expression: [169]

$$E_x^{LDA}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho(\vec{r})^{4/3}\,d\vec{r} \tag{15}$$

The exact correlation functional for a uniform electron gas is not known (except in high and low density limits) but the correlation energy of this system has been studied numerically and parameterized in the form of analytic functionals such as: [170]

$$E_c^{LDA}[\rho] = -A\int \rho(1+\alpha_1 \vec{r}_s)\ln\left[1+\frac{1}{A\left(\beta_1 \vec{r}_s^{1/2}+\beta_2 \vec{r}_s+\beta_3 \vec{r}_s^{3/2}+\beta_4 \vec{r}_s^{2}\right)}\right]d\vec{r} \qquad (16)$$

where, $\vec{r}_s = \left(\frac{3}{4\pi\rho}\right)^{\frac{1}{3}}$ and $A, \alpha_1, \beta_1, \beta_2, \beta_3$ and β_4 are fixed parameters.

The LDA predicts fairly accurate bond lengths and lattice constants, but severely overestimates atomization energies of molecules and solids. For comparison, the HF method, which is computationally more expensive than the LDA, predicts bond lengths much less accurately than LDA and overestimates atomization energies [171]. Various approaches, using different analytical forms for *E*c, have generated several LDAs for the correlation functional, including: Vosko-Wilk-Nusair (VWN) [172] Perdew-Zunger (PZ81) [173] Cole-Perdew (CP) [174] Perdew-Wang (PW92) [175] and many more.

The Generalised Gradient Approximation (GGA)

In foregoing discussion it is seen that LDA approximates the energy of a system by taking the energy of local constant density into account, but unfortunately it fails in some situations where the density undergoes rapid changes. The LDA uses the exchange-correlation energy for the uniform electron gas at every point in the system regardless of the homogeneity of the real charge density. But we encounter a very different case for non-uniform charge densities. In the later case, the exchange-correlation energy deviates significantly from that observed in the system with uniform charge density. To overcome this problem some better approximations that can express the deviation in charge density in terms of gradients and higher spatial derivatives necessitates. Emergence of generalized gradient approximation (GGA) [176, 177] has become helpful in solving the problem of slowly varying charge density as it uses the gradient of the charge density to estimate this deviation correctly. The typical form for a GGA functional is:

$$E_{xc} \approx \int \rho(\vec{r})\,\varepsilon_{xc}(\rho, \vec{\nabla}\rho)d\vec{r} \qquad (17)$$

The most popular and widely used GGA functional in solid state physics as well as in computational chemistry is the Perdew, Burke and Ernzerhof (PBE) [178, 179] and its modified form RPBE [180] and revPBE [181]. In one of their spectacular work, Perdew and co-workers [182, 183] have shown that it is impossible for a GGA to perform well for certain pairs of properties, *e.g.*, both for atomization energies and lattice constants of solids.

Meta-GGA Functionals

Meta-GGAs are the most sophisticated semi-local functionals, incorporating important exact conditions with almost the same computational cost as that of GGAs. A meta-GGA functional in its original form includes the second derivative of the electron density (the Laplacian); it depends explicitly on the semi-local information in the Laplacian of the spin density or on the Kohn-Sham orbitals through the kinetic energy density [184-186]. The functional is written in the form:

$$E_{xc} \approx \int \rho(\vec{r})\, \varepsilon_{xc}\left(\rho, \left|\vec{\nabla}\rho\right|, \vec{\nabla}^2\rho, \tau\right) d\vec{r} \tag{18}$$

where the kinetic energy density τ is:

$$\tau = \frac{1}{2}\sum_i \left|\vec{\nabla}\varphi_i\right|^2 \tag{19}$$

Some extensively used meta-GGA functionals are TPSSLYP1W, M06-L, M11-L, *etc*. Unfortunately, the meta-GGA functionals have not yet touched the level of accuracy of numerical performances when compared with the GGAs. As a matter of fact, they can be considered as a significant improvement for some properties only (*e.g*. thermochemistry) [187, 188], while they show very poor performances when some other molecular parameters are considered (*e.g*., geometries) [189].

Hybrid Exchange Functionals

Axel Becke, in 1993, [190] has came forward with one of his spectacular findings; the hybridization with Hartree-Fock (exact) exchange provides with a simple scheme for improving results of molecular properties such as atomization

energies, bond lengths and vibrational frequencies, which tend to be poorly described with simple '*ab-initio*' functionals [191].

The general principle lying behind this type of functionals is the mixing of fraction of '*exact exchange*' with GGA semi-local exchange. The exact exchange energy functional is expressed in terms of the Kohn-Sham orbitals rather than the density, so it is termed as an implicit density functional. A hybrid exchange-correlation functional is usually constructed as a linear combination of the Hartree-Fock exact exchange functional, E_x^{HF} :

$$E_x^{HF} = \frac{1}{2}\sum_{i,j}\int\int \psi_i^*(\vec{r}_1)\psi_j^*(\vec{r}_1)\frac{1}{|\vec{r}_{12}|}\psi_i(\vec{r}_2)\psi_j(\vec{r}_2)\,d\vec{r}_1 d\vec{r}_2 \tag{20}$$

One popular example of the hybrid exchange functional is the B3LYP (Becke, Three Parameter Hybrid Functional, Lee-Yang-Parr) [192, 193] having the form:

$$A * E_x^{Slater} + (1-A) * E_x^{HF} + B * \Delta E_x^{Becke} + E_c^{VWN} + C * \Delta E_c^{non-local} \tag{21}$$

where A, B and C are the constants determined by Becke *via* fitting the G1 molecule set. A number of different functionals have been proposed in the last few decades [194-198]. However, it has been observed that a single functional is not suitable for handling different situations, rather shows applicability to some specific cases only. In contrast, the hybrid functionals show wide applicability [199].

Density Functional Reactivity Theory (DFRT)

In the previous section we have discussed how density functional theory (DFT) proves itself as one of the most precious theories in determining the energy of a molecular system. But scientists are now interested to know some other chemical properties of systems (based on its electron density) such as chemical potential, chemical hardness, electrophilicity, nucleophilicity *etc*. that can well define the reactivity of a system. So, in order to quantify molecular parameters researchers have developed some quantum chemical models called *reactivity descriptors*. The foundation stone of these descriptors has been established by Robert G. Parr and

co-workers with the evolution of a second branch of DFT in 1970s and early 1980s, also called the *conceptual density functional theory* or *density functional reactivity theory* (DFRT) [200-203]. The last few decades have seen the evolution of many fundamental concepts, such as frontier molecular orbital (FMO) [204-206], electron localized function (ELF) [207, 208], molecular electrostatic potential (MEP) [209-213] and electronegativity equalization method (EEM) [214-218] that are extensively used to explain the stereo- and regio- selectivity of a wide variety of reactions. But these principles have remained empirical until the great theory by Parr came into play and provided the theoretical basis of these formal concepts.

Reactivity descriptors (DFT based) are some sort of mathematical parameters, which help us to quantify chemical properties of a molecular system as a whole or at some selective sites. Global reactivity descriptors (GRD) such as global hardness (η), electrophilicity (ω), chemical potential (μ) *etc*. are used to define the properties of a molecular system as a whole, whereas local reactivity descriptors (LRD) such as Fukui functions (*f*), local softness (*s*) *etc*. are helpful in understanding the behavior and reactivity of chemical species at a particular site.

Iczkowski and Margrave [219] expressed chemical potential (μ) as the negative element of electronegativity (χ). It is defined as the first derivative of energy with respect to the number of electron(s) *N* at constant external potential, $v(\vec{r})$.

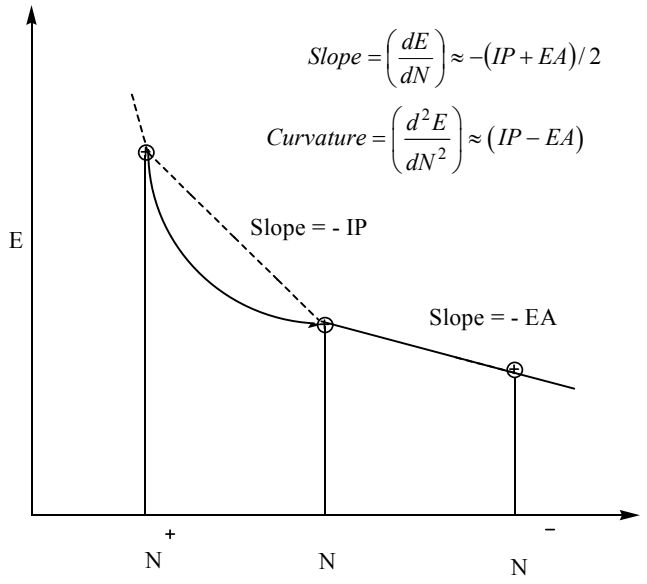
$$\mu = -\chi = \left(\frac{\partial E}{\partial N}\right)_{v(\vec{r})} \tag{22}$$

Global hardness (η) [220] of an electronic system is defined as the second derivative of energy (*E*) with respect to the number of electron(s) *N* at constant external potential $v(\vec{r})$.

$$\eta = \frac{1}{2}\left(\frac{\partial^2 E}{\partial N^2}\right)_{v(\vec{r})} = \frac{1}{2}\left(\frac{\partial \mu}{\partial N}\right)_{v(\vec{r})} \tag{23}$$

where μ is the chemical potential of the system.

A discontinuity in the value of energy E with respect to the variation of N makes it practically difficult to evaluate chemical potential (μ) and chemical hardness (η) [221-222]. In most of the numerical applications, μ and η are calculated using finite difference approximation [223] in terms of ionization potential (*IP*) and electron affinity (*EA*). A practical approach for evaluating *IP* and *EA* can be made from the plot of *E versus N* (**29**), generally these plots are not straight lines but are convex upwards. The curvature defines η whereas the slope defines μ.



(**29**)

Using finite difference approximation η and μ can be defined as:

$$\eta = \frac{IP - EA}{2} \tag{24}$$

$$\mu = -\left(\frac{IP + EA}{2}\right) \tag{25}$$

On the other hand, the ΔSCF method defines the terms, ionization potential (*IP*) and electron affinity (*EA*) of the system as:

$$IP = E_{N-1} - E_N \tag{26}$$

$$EA = E_N - E_{N+1} \tag{27}$$

where, E_{N-1}, E_N and E_{N+1} are the energies of *N-1, N* and *N+1* electron systems respectively.

Thus, once the energies of the neutral (*N*), cationic (N^+) and anionic (N^-) systems are known, μ and η values can easily be evaluated from the following formulae

$$\mu = \frac{(E_{N+1} - E_{N-1})}{2} \tag{28}$$

$$\eta = \frac{(E_{N-1} + E_{N+1} - 2E_N)}{2} \tag{29}$$

Global softness (*S*) is defined as the inverse of η.

$$S = \frac{1}{2\eta} \tag{30}$$

Using the finite difference approximation, *S* can be approximated as:

$$S = \frac{1}{IP - EA} \tag{31}$$

Koopmans' theorem [224] states that energy of HOMO (Highest Occupied Molecular Orbital) is equal to the negative of the first ionization energy (*IP*) of a molecular system. This theory is also formally useful for determination of electron affinities (*EA*), *i.e. EA* is equivalent to the negative of LUMO (Lowest Unoccupied Molecular Orbital). Unfortunately, the Koopmans' theorem does not hold good for DFT as it is not an MO method. But, HOMO energy refers to the eigenvalue of the highest occupied KS orbital in DFT. Hence, some interpretations of the Kohn-Sham orbital energies are possible. Using Koopmans'

theorem we can define *IP* and *EA* in terms of energies of HOMO (ε_{HOMO}) and LUMO (ε_{LUMO}) as:

$$IP = -\varepsilon_{HOMO} \tag{32}$$

$$EA = -\varepsilon_{LUMO} \tag{33}$$

and therefore μ and η can be expressed as:

$$\eta = \frac{\varepsilon_{LUMO} - \varepsilon_{HOMO}}{2} \tag{34}$$

$$\mu = \frac{\varepsilon_{LUMO} + \varepsilon_{HOMO}}{2} \tag{35}$$

Parr and his co-workers [225] have proposed electrophilicity index (ω) as a measure of electrophilicity of a system,

$$\omega = \frac{\mu^2}{2\eta} \tag{36}$$

It is a measure of the capacity of a species to accept an arbitrary number of electrons. Chattaraj *et al.* [226] have proposed a broader and very general local reactivity descriptor called philicity, which encompasses electrophilic, nucleophilic and radical reactions. Later, Roy *et al.* [227, 228] have outlined the limitations in applicability of this index and concluded that it may not always be logical to approximate the global reactivity to the local reactivity of the predominant site. Local electrophilicity index [229] is defined as:

$$\omega_k^{\pm} = \omega f_k^{\pm}, \tag{37}$$

where, f_k^+ is the electrophilic Fukui function and f_k^- is the nucleophilic Fukui function. The condensed–to–atom variants for the atomic site *k* can be written as:

$$\omega_k^{\alpha} = \omega f_k^{\alpha};\ \alpha = +, -, 0\ .$$

where, $\alpha = +, -$ and 0 refer to nucleophilic, electrophilic and radical attack respectively.

However, the condensed Fukui function (CFF, a local reactivity descriptor) bears a good significance in determining or in having a greater insight into a particular atom [230]. For an atom '*x*' in a molecule with *N* electrons in a constant external potential, $\nu(\vec{r})$, CFF can be obtained from finite different approximation as:

$$f_x^+ = [\rho_x(N_0 + 1) - \rho_x(N_0)] \text{ for nucleophilic attack} \tag{38}$$

$$f_x^- = [\rho_x(N_0) - \rho_x(N_0 - 1)] \text{ for electrophilic attack} \tag{39}$$

where, $\rho_x(N_0)$, $\rho_x(N_0 + 1)$ and $\rho_x(N_0 - 1)$ are electronic population on atom *x* in the molecule with *N0*, *N0 + 1* and $N_0 - 1$ electrons respectively.

Another important local descriptor is the local softness, defined as:

$$s^{\pm} = S f_k^{\pm} \tag{40}$$

Toro-Labbé *et al.* [231, 232] have proposed a dual descriptor (Δf), which is defined as the difference between the nucleophilic and electrophilic Fukui functions and is given by,

$$\Delta f = \left[f^+ - f^- \right] \tag{41}$$

For $\Delta f > 0$, the site is favored for nucleophilic attack, whereas for $\Delta f < 0$, the site could hardly be susceptible to undertake a nucleophilic attack but may be favored for an electrophilic attack. The associated dual local softness [233] is defined as:

$$\Delta s = \left[s^+ - s^- \right] \tag{42}$$

where, $s^+ = S f_k^+$ and $s^- = S f_k^-$

In recent past, these descriptors have been exploited to resolve a wide variety of structure and reactivity issues that includes global as well as local reactivity of

species, stability of systems, stability of conformations *etc*. and are well documented [234-248].

COMPUTATIONAL STUDIES ON NITROGEN MUSTARDS

Though lots of experimental studies have been devoted to understand the action of these drug molecules, proper understanding in molecular level has been possible only with the help of methods of computational chemistry. Although not much of computational work has so far been witnessed in this field, their importance in this regard cannot be ignored.

Hamza *et al.* [249] performed quantum mechanical study of S-methylated forms of sulfur mustard which can be considered as the pioneering computational work on nitrogen mustards as well. They performed HF calculations on episulfonium ion which shares a similar structure to that of Az^+ ion with an S-center. They also calculated the energy barrier of the reaction pathway for bis-(2-chloroethyl) methyl sulfonium dication ($MeHD^{2+}$). Shukla *et al.* [250] too have added some good piece of information based on quantum chemical studies on mustine. They have performed DFT and MP2 level of study on the reactions of mustine with different nucleophilic centers in DNA bases and observed some noteworthy results. The rate of the reaction is expected to be controlled by the magnitude of the free energy of activation. Gibb's free energy of activation of mustine when reacted to guanine N7 is observed to be minimum (17.48 kcal/mol) followed by adenine N3 (18.05 kcal/mol) and cytosine N3 (21.94 kcal/mol). This study has attested the assumption of earlier experimental studies that nitrogen mustards react preferentially at guanine N7. However, while studying a reaction, interaction energy between the species should also be taken into account. They have observed that the binding energy for alkylation at different sites is negative, confirming the formation of stable adducts. However, results obtained for alkylation at O2 and O4 of thymine, guanine O6 and cytosine O2 are not in favor of the formation of stable adducts. Being a positively charged species, binding energy of Az^+ ion with different nucleophilic sites in aqueous phase is found to be comparatively lower than that in the gas phase. The trend in aqueous phase (at MP2/6-31+g(d) level of theory) is observed to be: cytosine N3 > guanine N7 > guanine N3 > adenine N3 > adenine N1 > adenine N7. Thus, the calculations performed by Shukla *et al.*

prove that guanine N7 and adenine N3 in DNA are the most probable sites of alkylation by nitrogen mustards, which is in agreement with previous experimental observations [251].

Hydrolysis of chlorambucil analogs have been studied by Pineda *et al.* [252]. Both the SN1 and SN2 reaction pathways are studied and it has been confirmed that the reaction involves Az^+ ion formation *via* a first order reaction, subjected to an energy barrier of 24.8 kcal/mol (computed at M062X/6-31+g(d,p) level of theory).

Ab-initio calculations on isolated GC pair performed by Vasilescu *et al.* [253] have confirmed the preferential attack at guanine N7 over the other sites. Thermochemical studies also confirm that alkylation of isolated GC base pair by sulphur or nitrogen mustard is exothermic ($\Delta H < 0$) and spontaneous ($\Delta G < 0$). Computed chemical potential (μ) and the electrophilicity (ω) clearly show that an electronic charge flow from the nucleophilic GC base pair to the electrophilic species episulfonium or aziridinium ion is involved.

Mann [254] has performed an explicit solvent phase *ab-initio* molecular dynamics simulation to study the activation of nitrogen mustards: mustine and phosphoramide mustard. The simulations have predicted a concerted reaction occurring by means of neighbouring-group participation with the nearby nucleophilic tertiary nitrogen. The calculated free energy of activation for Az^+ ion formation for mustine has been observed to be 20.4 kcal/mol which is close to the experimental value of 22.5 kcal/mol. These simulations also indicate a dynamic transition state characterized by pronounced changes in the local water structure within the first hydration shell. The complete mechanism involving solvent reorganization, ionization of the C-Cl bond and internal cyclization of the Az^+ ion has been captured from elevated temperature simulations. Rate constants for Az^+ ion formation from both mustine and phosphoramide mustard have been calculated to be 26 s and 34.6 min respectively which are in agreement with the experimental values [255, 256].

Recently, Polavarapu *et al.* [257] have studied the mechanism of alkylation of guanine and adenine by mustine, melphalan and phenyl mustard using DFT. They

have concluded that the rate of formation of Az^+ ion of mustine is much preferred to the other two aromatic nitrogen mustards. The free energy of activation of Az^+ ion formation for mustine has been calculated to be 9.26 kcal/mol which is smaller as compared to those of melphalan and phenyl mustard (23.00 kcal/mol and 22.53 kcal/mol respectively). Thus the Az^+ ion formation by mustine is kinetically favored over the other two. They have also showed that imminium ion formation is thermodynamically favored but a huge amount of free energy of activation (46 kcal/mol) prevents its formation. The free energy barrier for guanine alkylation is comparable in all the cases (21.44 kcal/mol, 27.74 kcal/mol and 28.88 kcal/mol for mustine, phenyl mustard and melphalan respectively). However, in case of adenine alkylation these barriers are larger (27.79 kcal/mol, 35.33 kcal/mol and 33.01 kcal/mol for mustine, phenyl mustard and melphalan respectively). This study supports earlier experimental observation that guanine alkylation is preferred over adenine alkylation.

Our research group too have performed a number of studies on nitrogen mustards. Especially the alkylation reaction and different properties of the nitrogen mustards have been studied. We have made successful applications of DFT and DFRT to make an in-depth study on alkylation of DNA bases, reactivity of Az^+ ion, and stability of drug-guanine adducts *etc*. Few examples are discussed in section 1.4.

SOME APPLICATIONS OF DFT AND DFRT ON NITROGEN MUSTARDS

We have studied alkylation of DNA by nitrogen mustards. The kinetic as well as thermodynamic driving force involved in DNA alkylation by nitrogen mustards have been studied extensively, clarifying many doubts. Alkylation takes place in cellular environment and it is an uphill task for a computational chemist to mimic the cellular environments. Consideration of aqueous phase is a good approximation to some extent. Nevertheless simple model chemistry cannot represent the real situation. However, DFT studies provide a better understanding of the factors that influence the rate of the reaction, free energy of activation, conformation of the drug *etc*. Moreover, reactivity descriptors defined under DFRT are extensively used to explain the reactivity pattern of the drug molecules, intermediates and products formed during alkylation.

During the alkylation process, the Az^+ ion attacks guanine N7 and the positively charged carbon center in Az^+ ion accepts electron density from guanine N7. But this is possible only when the LUMO is associated with the carbon centers in the Az^+ ion. Therefore, the position as well as the energy of the LUMO of the Az^+ ion becomes important. It has been observed that, the LUMO of the Az^+ ion (*i.e.*, with ∠N3C2C1≈60°) is localized away from the tricyclic ring (**30**).

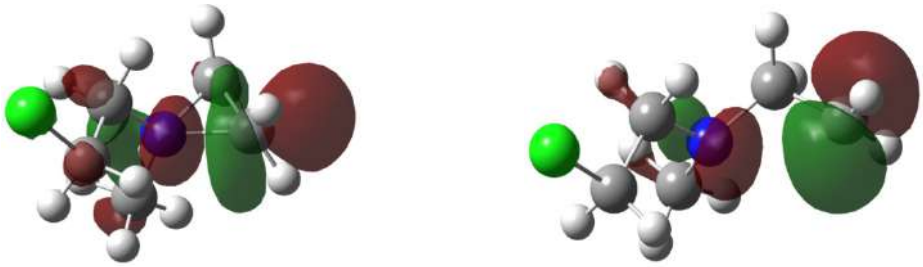
(**30**) Shape of the LUMO of the Az^+ ions

This position of the LUMO rules out the interaction between the Az^+ ion (C atoms of the tricyclic ring in Az^+ ion) and guanine (N7 center). Therefore, it becomes important to study the factors that facilitate the alkylation reaction. Let us discuss some of these factors based on our own research work. Most of our calculations have been carried out with 6-31+g(d) and 6-31++g(d,p) basis sets with Becke three parameter Lee, Yang and Parr correlation functionals (B3LYP) which is one of the most popularly used hybrid functionals.

Case Study I: Structural Variation in Aziridinium Ion Facilitates Alkylation

The optimized structure of the Az^+ ion shows that it possesses a perfect tricyclic ring with ∠N3C2C1≈60° (**30**) (Please note that the N3-atom described here is referred to as the N-atom of Az^+ ion with C1 and C2 as the other two carbon centers in the ring. The numbering has been done arbitrarily for the sake of simplicity and does not refer to any scientific nomenclature and this numbering will be used in the rest of the chapter). In this conformation the LUMO density is mostly confined within the chloroethyl side chain. During mono-adduct formation, the N3-C1 bond of the Az^+ ion rips apart and the ∠NCC bond angle becomes ~110° (in mono-adduct). We have considered the mustine molecule, and an attempt has been made to analyze what happens when the tricyclic ring rips apart. As the ∠NCC bond angle varies from 60° to 120°, the shape of the LUMO of the Az^+ ion changes and shifts from the chloroethyl side chain towards the ring carbon. A small variation (~5°) in ∠NCC

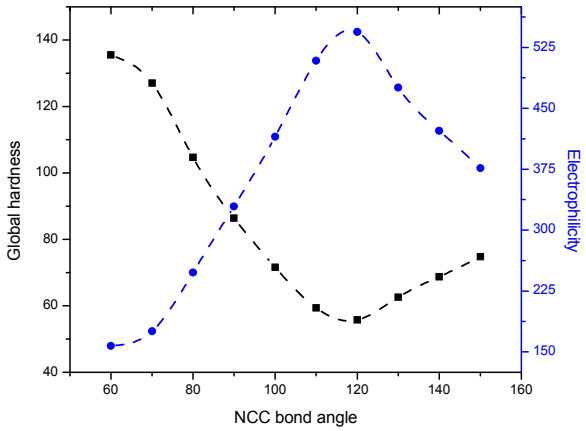
shifts the LUMO towards the ring carbon (**31a**) and at ∠NCC >110° most of the LUMO concentrate at the C1 center (**31b**) [258].



(**a**) ∠NCC=65° (**b**) ∠NCC=110°

(**31**) Variation of the shape of the LUMO with ∠NCC bond angle

This study verifies that the variation of the shape of the LUMO during alkylation of DNA by nitrogen mustards is an important factor for the alkylation reaction to occur. Moreover, the two important principles, maximum hardness principle (MHP, according to which, maximum hardness leads to maximum stability) and minimum electrophilicity principle (MEP, most stable configuration possess minimum electrophilicity) are also obeyed during structural variation (**32**) and maximum hardness is observed in case of Az^+ ion. However, as the ∠NCC bond angle increases, hardness decreases, implying destabilization of the species according to MHP.



(**32**) Variation in global hardness (■) and electrophilicity (●) (in kcal/mol) of the drug intermediate with variation in ∠NCC bond angle at B3LYP/6-311++g(d,p) level in gas phase.

Case Study II: Affinity of Aziridinium Ions Towards Different Nucleophiles

Apart from the DNA base pairs, there are plenty of nucleophilic centers present in cellular environment that compete with the base pairs to interact with the Az^+ ion and may interrupt DNA alkylation. So keeping this in mind, we have analyzed the affinity of Az^+ ion towards different nucleophiles using DFT and DFRT [259]. Interaction energy of the Az^+ ion towards different nucleophilic centers is one of the key factors in determining how strongly the Az^+ ion may get diverted from its target molecule (DNA bases). Two groups of nucleophiles are chosen for this purpose: one bearing negative charges, (group I) and the other with neutral nucleophilic centers (group II). BSSE (basis set superposition error) corrected interaction energies at different level of theories in gas and aqueous phases are shown in Table **1**.

Table 1: Interaction energy (in kcal/mol) of different nucleophiles with Az^+ ion at three different levels of theory in gas and aqueous phases

	In gas phase			In aqueous phase		
	B3LYP/6-31++g(d,p)	B3LYP/Aug-cc-pVDZ	B3LYP/6-311++g(d,p)	B3LYP/6-31++g(d,p)	B3LYP/Aug-cc-pVDZ	B3LYP/6-311++g(d,p)
Group I nucleophiles						
Cl^-	-125.89	-126.20	-126.65	-4.51	-4.75	-7.00
NH_2^-	-198.42	-197.84	-198.87	-78.42	-77.38	-73.08
$NHMe^-$	-200.60	-200.09	-200.83	-85.27	-82.85	-80.91
NMe_2^-	-197.18	-195.67	-197.11	-87.48	-85.68	-85.05
OH^-	-180.75	-180.65	-181.17	-52.61	-51.98	-49.81
OMe^-	-177.16	-176.98	-177.29	-58.30	-57.11	-56.82
SH^-	-146.63	-146.33	-147.40	-32.57	-31.96	-24.56
SMe^-	-155.65	-154.75	-156.13	-39.16	-38.44	-20.24
$MeCOO^-$	-141.20	-141.52	-141.33	-26.34	-26.40	-40.59
$HCOO^-$	-137.15	-137.43	-137.45	-34.30	-34.30	-30.81
Group II nucleophiles						
MeCOOMe	-0.45	-0.84	0.02	9.02	8.04	8.38
NH_3	-12.27	-12.37	-12.58	-23.09	-23.94	-19.39
$NHMe_2$	-24.56	-24.67	-24.80	-24.02	-24.10	-22.85
NMe_3	-26.28	-26.18	-26.43	-20.87	-20.60	-22.68
$OC(NH_2)_2$	-41.58	-44.56	-40.93	-36.31	-37.05	-27.03
$OCHNH_2$	-22.35	-21.98	-21.77	-22.75	-23.37	-15.34
$OCMe_2$	-2.02	-1.60	-1.57	6.80	6.78	1.57

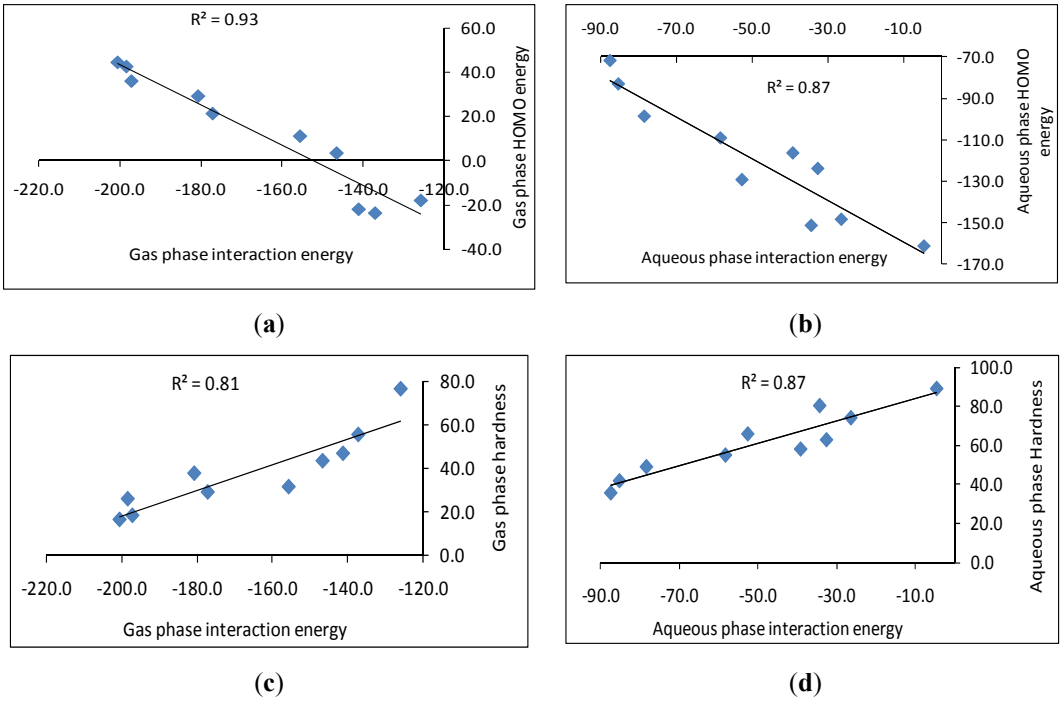
Table 1: contd…

$OCMeNH_2$	-26.34	-25.70	-24.69	-21.39	-21.37	-15.69
SHMe	-0.96	-0.49	-0.49	-1.71	-1.31	-1.70
SMe_2	-13.04	-12.35	-12.57	-7.24	-6.83	-8.39

In case of group I nucleophiles, interaction energies in gas phase are found to be comparatively higher than in aqueous phase. Presence of charges on the nucleophiles as well as on the Az^+ ion makes them stable in aqueous phase. Thus, it becomes easier for these species to remain in unreacted form in aqueous phase. In aqueous phase, the nucleophiles (group I) bearing N-centers exhibit the highest interaction energies whereas Cl^- ion shows the lowest. The order of interaction energies among the group I nucleophiles is found to be: N-center > O-center > S-center > Cl^-, with an exception in case of O-center in carboxylic group, which shows low interaction energy due to delocalization of the charge over the $–COO^-$ group. Higher affinity of the nucleophiles with N-centers suggests that Az^+ ion is more prone to attack at different N-centers (preferentially at tertiary N-centers) present in DNA, RNA and in different protein molecules. Exceptionally low affinity of Cl^- ions towards the Az^+ ion explains why nitrogen mustards form Az^+ ion by releasing Cl^- ion. The Az^+ ion exhibits a weak interaction with the group II nucleophiles, both in gas as well as in solvent phases. Some of them even show repulsive interaction (positive interaction energy). We have not observed any sharp variation in the values of the interaction energies on moving from gas to aqueous phase for group II nucleophiles. Nucleophiles with N-centers show strong interactions in both phases, whereas an exactly opposite case has been observed in case of nucleophiles with S-centers.

Interaction of the Az^+ ion and the nucleophiles resulted in a strong covalent bond formation; shorter (and hence stronger) bond formation has been observed in case of group I nucleophiles (especially with those having N-centers) compared to the group II nucleophiles.

In case of group I nucleophiles, some linear relationship between the reactivity descriptors and interaction energies are observed, (**33a-d**) but no such relationship has been observed in case of group II nucleophiles.



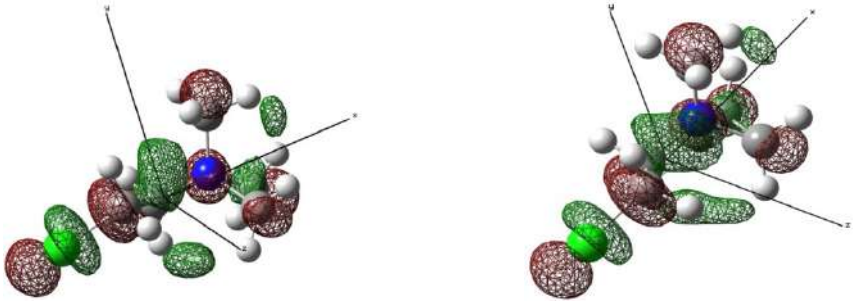
(**33**) Variation of interaction energy with HOMO energy, global hardness, chemical potential and electrophilicity of group I nucleophiles at B3LYP/6-31++g(d,p) level of theory in gas and aqueous phases (all parameters are in kcal/mol).

Case Study III: Effect of External Electric Field on the Reactivity of the Aziridinium Ion

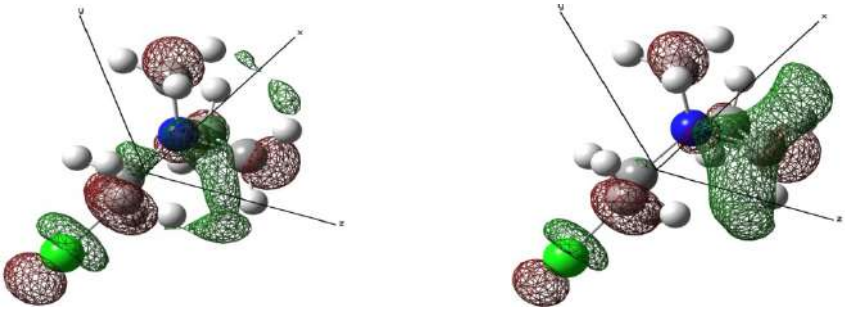
During its lifetime, the Az^+ ion has to pass through different environments ranging from non-polar (within the cell membrane) to polar environments (in extra- and intra- cellular fluids, blood *etc*.) before interacting with guanine (in DNA). These polar environments consist of charged particles and are expected to exert some electric fields of different magnitude on the species. Accordingly, we studied the effect of external electric fields on Az^+ ion of mustine and observed its behaviour using DFRT [260]. It has been noticed that in absence of external electric field the LUMO is mostly located at the chloroethyl side chain (**34a**). As the electric field is applied, the position of the LUMO shifts towards the direction of the applied field, and this shifting of LUMO is observed to depend on the magnitude of the applied field. For instance, when the applied field value is 0.10 V/Å, (**34b**), the LUMO starts shifting towards the direction of the external electric field and at a field value of 0.30 V/Å, a large portion of the LUMO gets shifted, (**34d**). At

higher field values, say greater than 0.50 V/Å, almost complete shifting of the LUMO is observed (**34e-f**).

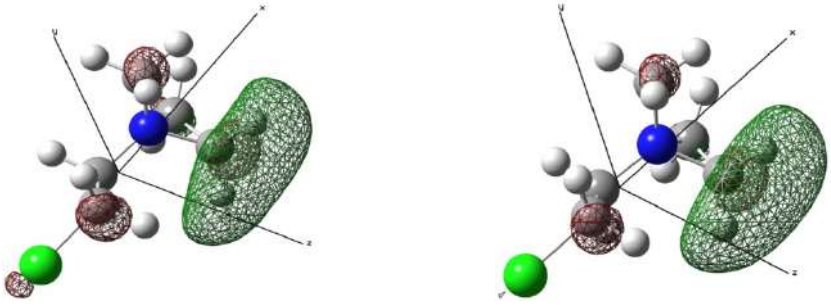
Thus, a cytoplasmic environment may shift the LUMO significantly towards the ring carbon which would facilitate alkylation. Hence external electric field is also an important factor that affects the shape and energy of the LUMO. Moreover, reactivity of the Az^{+} ion varies with the application of external electric field (**35a-b**) and MHP as well as MEP are obeyed.



(**a**) Field=0.0 V/Å (**b**) Field=0.10 V/Å

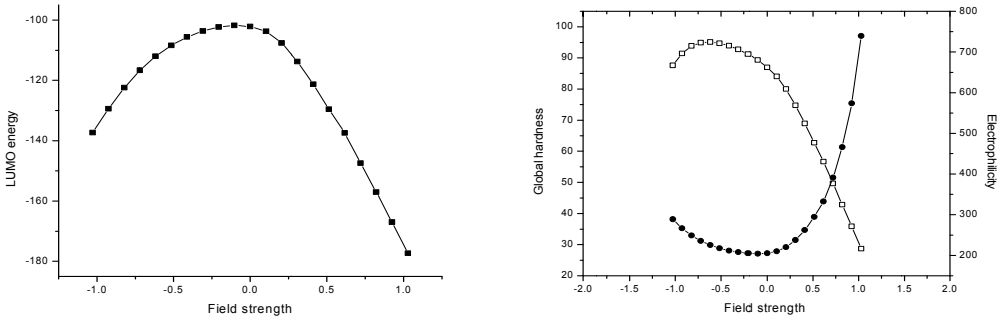


(**c**) Field=0.20 V/Å (**d**) Field= 0.30 V/Å



(e) Field= 0.50 V/Å (f) Field= 0.60 V/Å

(**34**) Shapes of LUMO in the presence of external electric field applied along z axis.



(**a**) (**b**)

(35) Variation of LUMO energy, global hardness (□) and electrophilicity (●) with external electric field along *x* axis (values of LUMO energy, hardness and electrophilicity are in kcal/mol and field strength is in V/Å)

Case Study IV: Variation of Reactivity During Alkylation

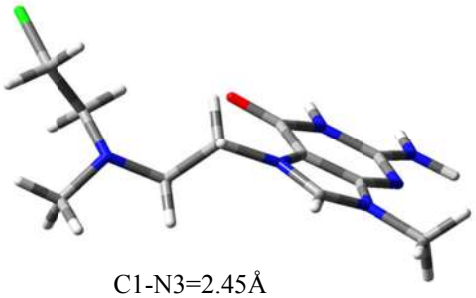
Another study comprises of the variation of reactivity/stability of the tricyclic Az^+ ion of mustine with the ∠NCC bond angle of the tricyclic ring during alkylation of guanine [261]. We followed the path of the reaction along the intrinsic reaction coordinates (IRC) and observed a sharp variation in the reactivity of the Az^+ ion around the transition state (TS). Along the reaction pathway, configuration of the Az^+ ion changes; at $d > 2.30$Å, (d is the distance between C1 center of Az^+ ion and guanine N7) the Az^+ ion retains its typical tricyclic structural parameters with ∠N3C2C1≈ 60.0° and both N-C (ring carbon) distances are equal to 1.59 Å (**36a**). As *'d'* gets shorter, the two species come within the interaction range and at $d \approx 2.30$ Å, the tricyclic ring starts to open up. The TS of the reaction is located at $d = 2.12$ Å, (**36b**) and in mono-adduct $d = 1.48$ Å.

C1-N3= 1.59 Å
C1-N7= 2.30 Å
∠C1C2N7=65.1°

(a) (Reactant)

C1-N3=1.89 Å
C1-N7=2.12 Å
∠C1C2N7=79.9°

(b) (TS)

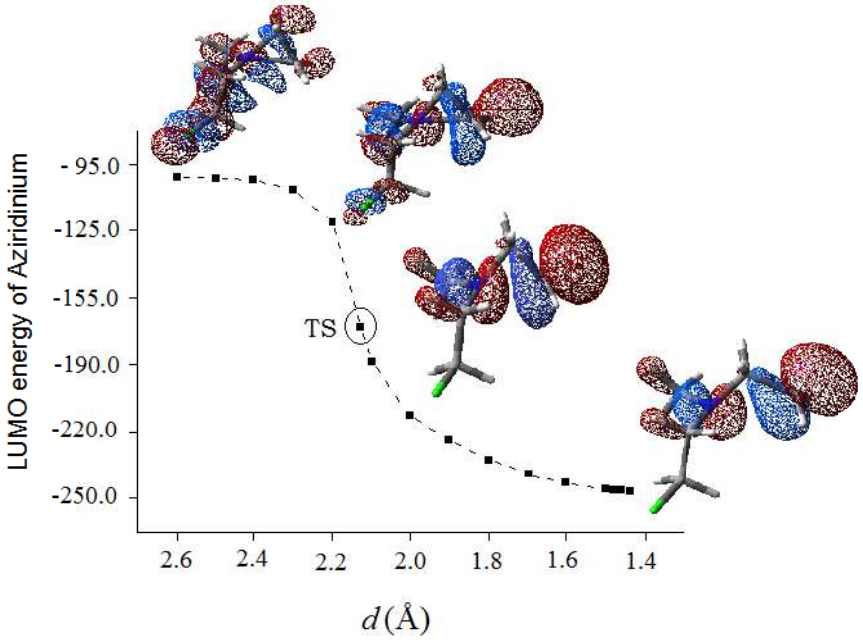


C1-N3=2.45Å
C1-N7=1.48 Å
∠C1C2N7=109.4°

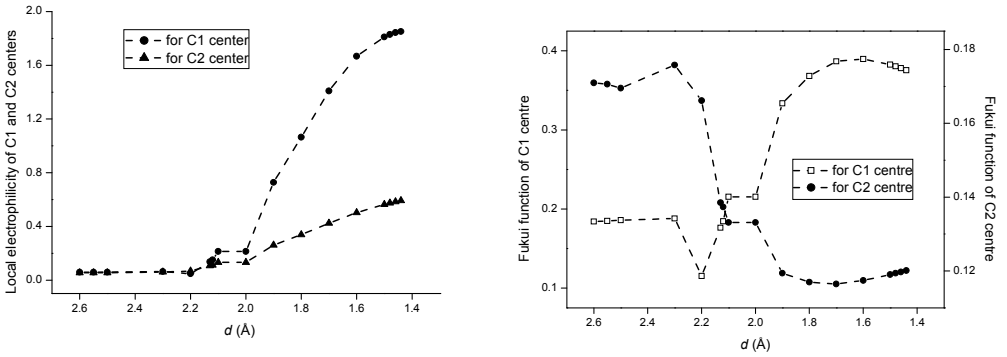
(c) (Product)

(**36**) Variation of the structure of the Az^+ ion during mono-adduct formation (at B3LYP/6-31+g(d) level of theory)

Shifting of the LUMO from the chloroethyl side chain towards the tricyclic ring with the variation of *'d'* is shown in **(37)**. Around the TS, larger part of the LUMO is shifted towards the ring carbon. Moreover, along the IRC, a drastic change in the local reactivity parameters of carbon centers of the Az^+ ion is observed (**38**).



(**37**) Variation of shape and energy of LUMO (in kcal/mol) of Az^+ ion along the IRC.



(**a**) Variation of ω^+ (**b**) Variation of f^+

(**38**) Variation of the local reactivity descriptors ω^+ and f^+

Case Study V: Mono- and Bis-Alkylation of DNA by Nitrogen Mustards

It has become cumbersome in some of the cases to define the reaction possibilities using only the global reactivity descriptors as it is a site specific reaction; it is thus advisable to study the reactivity patterns of those sites using the local reactivity descriptors. Local parameters such as Fukui function (f^+), local softness (s) and local philicity ($\omega^{\pm}$) are some tools handy for this purpose [262]. Here we have studied the formation of mono- as well as cross-linked adducts by a few nitrogen mustards and analyzed the local reactivity pattern. Of all the drugs chosen for the study *i.e.* mustine, melphalan, chlorambucil, phosphoramide mustard, bendamustine, uracil mustard and spiromustine, Az^+ ion of mustine exhibits the highest tendency for mono-adduct formation whereas second aziridinium ion, (Az^{2+}) (**Scheme 1**) of uracil mustard exhibits maximum tendency for cross-linked adduct formation. The extents of solvent effect measured in terms of free energy of solvation (ΔG_{sol}) suggest that the extent of the effect depends upon the charge present on the species; free energy of solvation of Az^{2+} is higher than that of Az^+. Stability of mono- and cross-linked adducts has been measured in terms of global hardness which predicts a different trend in both the cases. Interaction energy happens to be of cardinal importance for a drug molecule. Though in gas phase, interaction energy in case of cross-linked adducts are much higher than in mono-adducts; in aqueous phase they are very close to each other, Table **2**. Moreover,

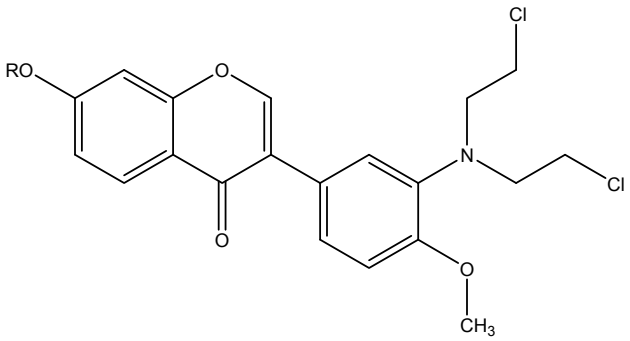
because of higher charge (+2), cross-linked adducts acquire more stability in aqueous phase compared to mono-adducts.

Table 2: Interaction energies (in kcal/mol) of mono- and cross-linked adducts in gas and aqueous phases (within bracket) at B3LYP/6-31+g(d) level of theory

Drug molecule	$\Delta E_{int\text{-}mono}$	ΔG_{sol}	$\Delta E_{int\text{-}cross}$	ΔG_{sol}
Mustine	-46.83 (-21.11)	-57.73	-62.39 (-21.43)	-138.32
Melphalan	-47.86 (-23.96)	-66.48	-64.67 (-27.12)	-140.58
Chlorambucil	-48.74 (-29.41)	-63.88	-61.07 (-27.49)	-141.90
Bendamustine	-43.54 (-24.37)	-69.49	-60.74 (-26.73)	-141.56
Phosphoramide mustard	-57.89 (-37.55)	-69.79	-79.17 (-38.99)	-141.76
Uracil mustard	-55.61 (-26.94)	-66.32	-71.61 (-30.14)	-146.29
Spiromustine	-51.27 (-25.72)	-59.12	-68.69 (-29.40)	-134.04

Case Study VI: Alkylation of DNA Base by Formononetin Derivatives of Nitrogen Mustard

Recently, we have analyzed the thermodynamic and kinetic aspects of guanine alkylation by nitrogen mustard derivatives of formononetin (**39**) using DFT [263].



(**39**)

Free energy of activation of the Az^+ ion and adduct formation by formononetin nitrogen mustard derivatives are comparable to those of chlorambucil and melphalan. Results confirm that Az^+ ion formation by the nitrogen mustard derivatives of formononetin is quite sluggish as compared to that of mustine. Higher energy barrier in the drug-guanine mono-adduct formation over Az^+ ion formation provides evidence of the later being kinetically favored.

Thermodynamic studies suggest that Az^+ ion formation is endothermic but drug-guanine adduct formation is exothermic. These drug molecules also exhibit significant interaction with guanine moiety (in DNA). It is worth mentioning that the results illustrate the applicability of MHP and MEP.

CONCLUDING REMARK

In summary, the nitrogen mustard derivatives are excellent DNA alkylator and have received extensive application in cancer chemotherapy over the last fifty years. This class of drug molecules inhibits cell growth by formation of interstrand cross-linking. Since its discovery, several hundreds of such molecules have been synthesized, and many of them are found to exhibit potent anticancer activity. Though the mechanism of action of these molecules has been properly explained, elucidation at molecular level is of utmost importance to make an exclusive in-depth understanding. In comparison to a large number of experimental studies, only a few computational studies have so far been performed on these drug molecules. In recent years DFT and DFRT have been successfully applied to understand the mechanism of DNA alkylation, effect of solvent on the alkylation process, stability of the chemical species involved in the reaction *etc*. In near future these kinds of studies may become helpful in developing new and more potent nitrogen mustards.

ACKNOWLEDGEMENTS

The authors acknowledge the funding from University Grant Commission, New Delhi, India (F. 41-201/2012(SR).

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

REFERENCES

[1] Robinson J.P. *The Problem of Chemical and Biological Warfare, The rise of CB weapons;* Stockholm: Almquist and Wikaell Humanities Pr.: New York, **1971**.

[2] Calabresi, P.; Parks, R.E. Antiproliferative agents and drugs used for immunosuppression. In: *The Pharmacological Basis of Therapeutics*. 6th ed.; Gilman, A.G.; Goodman, L.S.; Gilman, A., Eds.; Macmillan Publishing Co: New York, **1980**; pp. 1256-1233.

[3] Ross, W.C.J. *Biological alkylating agent*; Butterworths: London, **1962**.
[4] Chanutin, A.; Gjessing, E.C. The effect of nitrogen mustards upon the ultraviolet absorption spectrum of thymonucleate, uracil and purines. *Cancer Res*., **1946**, 6(11), 599-601.
[5] Elmore, D.T.; Gulland, J.M.; Jordan D.O.; Taylor, H.F.W. The reaction of nucleic acids with mustard gas. *Biochem. J.,* **1948**, 42(2), 308-316.
[6] Lawley, P.D. *Submolecular structure of the nucleic acid*; Proc. Chem. Soc.: London, **1957.**
[7] Lawley, P.D.; The relative reactivities of deoxyribonucleotides and of the bases of DNA towards alkylating agents. *Biochem. Biophys. Acta,* **1957**, 26(2), 450-451.
[8] Wheeler, G.P.; Morrow, J.S.; Skipper, H.E. Studies with mustards. II. Paper chromatography and radioautography of reaction mixtures of various compounds with S35-sulfur mustard. *Arch. Biochem. Biophys*., **1955**, 57(1), 133-139.
[9] Reiner, B.; Zamenhof, S. Studies on the chemically reactive groups of deoxyribonucleic acids. *J. Biol. Chem*., **1957**, 228(1), 475-486.
[10] Wheeler, G.P; Skipper, H.E. Studies with mustards. III. In vivo fixation of C14 from nitrogen mustard-C14H3 in nucleic acid fractions of animal tissues. *Arch. Biochem. Biophys.,* **1957**, 72(2), 465-475.
[11] Brookes, P; Lawley, P.W. The reaction of mustard gas with nucleic acids in vitro and in vivo. *J.Biochem.* **1960**, 77(3), 478-484.
[12] Balcome, S.; Soobong, P. Dorr, D.R.Q.; Hafner, L.; Philips, L.; Tretyakova, N. Adenine containing DNA-DNA cross-links of antitumor nitrogen mustard. *Chem. Res. Toxicol.,* **2004**, 17(7), 950-962.
[13] Rajski, S.R.; Williams, R.M. DNA Cross-linking agents as antitumor drugs. *Chem. Rev*., **1998**, 98, 2723-2795.
[14] Noll, D.M.; Manson, T.M; Miller, P.S. Formation and repair of interstrand cross-links in DNA. *Chem. Rev.,* **2006**, 106(2), 277-301.
[15] Singer, B. All oxygens in nucleic acids react with carcinogenic ethylating agents. *Nature,* **1976**, 264(5584), 333-339.
[16] Beranek, D.T; Weis, C.C; Swenson, D.H. A comprehensive quantitative analysis of methylated and ethylated DNA using high pressure liquid chromatography. *Carcinogenesis*, **1980**, 1(7), 595-606.
[17] Beranek, D.T. Distribution of methyl and ethyl adducts following alkylation with monofunctional alkylating agents. *Mutat. Res*., **1990**, 231(1), 11-30.
[18] Singer, B.; Grunberger, D. Molecular biology of mutagens and carcinogens. New York: Plenum press; 1983.
[19] Perrino, F.W.; Blans, P.; Harvey, S.; Gelhaus, S.; McGrath, C.; Akman, S.A.; Jenkins, S.; LaCourse, W.R.; Fishbein J.C. The N2-ethylguanine and the O6-ethyl- and O6-methylguanine lesions in DNA: contrasting responses from the bypass DNA polymerase eta and the replicative DNA polymerase alpha. *Chem. Res. Toxicol.,* **2003**, 16(12), 1616-1623.
[20] Guy, A.D.; Molko, L.; Teoul, W.R. Chemical synthesis of oligonucleotides containing N6-Methyladenine residues in the GATC Site. *Helv. Chem. Acta,* **1986**, 69(5), 1034-1040.
[21] Horn, T.; Urdea, M.S. Forks and combs and DNA: the synthesis of branched oligodeoxyribonucleotides. *Nucleic Acids Res.,* **1989**, 17(17), 6959-6967.
[22] Bernadou, J.; Blandin, M.; Meunier, B. A one step synthesis of O6-Methyl-2′deoxyguanosine. *Nucleosides & Nucleotides*, **1983**, 2(5), 459-464.

[23] Fischhaber, P.L.; Gall, A.S; Duncan, J.A.; Hopkins P.B. Direct demonstration in synthetic oligonucleotides that BCNU cross-links N1 of deoxyguanosine to N3 of deoxycytidine on opposite strands of duplex DNA. *Cancer Res.,* **1999**, 59(17), 4363-4368.

[24] Wilds, C.J.; Noronha, A.M.; Robidoux, S.; Miller, P.S. Mispair-aligned N3T-alkyl-N3T interstrand cross-linked DNA: synthesis and characterization of duplexes with interstrand cross-links of variable lengths. *J. Am. Chem. Soc.,* **2004**, 126(30), 9257-9265.

[25] Singer, B.; Kroger, M.; Carrano, M. O2- and O4-alkyl pyrimidine nucleosides: stability of the glycosyl bond and of the alkyl group as a function of pH. *Biochemistry*, **1978**, 17(7), 1246-1250.

[26] Spratt, T.E.; Campbell, C.R. Synthesis of oligodeoxyribonucleotide containing analogs of O6-Methylguanine and reaction with O6-Alkylguanine DNA alkyl transferase. *Biochemistry*, **1994,** 33(37), 11364-11371.

[27] Moschel, R.C.; Hudgins, W.R.; Dipple A. Selectivity in nucleoside alkylation and alkylation in relation to chemical carcinogenesis. *J. Org. Chem*., **1979**, 44(19), 3324-3328.

[28] Loechler, E.L. A violation of the Swain-Scott principle, and not SN1 versus SN2 reaction mechanisms, explains why carcinogenic alkylating agents can form different proportions of adducts at oxygen versus nitrogen in DNA. *Chem. Res. Toxicol.*, **1994**, 7(3), 277-280.

[29] Brookes, P.; Lawley, P.D. Effects of alkylating agents on T2 and T4 bacteriophages. *Biochem. J.,* **1963**, 89(1), 138-144.

[30] Lawley, P.D; Brookes, P. Molecular mechanism of cytotoxic action of difunctional alkylating agents and resistance of this action. *Nature,* **1965**, 206(983), 480-483.

[31] Mattes, W.B.; Hartley, J.A.; Kohn, K.W. DNA sequence selectivity of guanine-N7 alkylation by nitrogen mustards. *Nucleic Acids Res.,* **1986**, 14(7), 2971-2987.

[32] Silverman, R.B. The Organic Chemistry of Drug Design and Drug Action. 2nd ed. USA: Elsevier Academic Press; 2004.

[33] Hartley, J.A.; Bingham, J.P.; Souhami, R.B. DNA sequence selectivity of guanine-N7 alkylation by nitrogen mustards is preserved in intact cells. *Nucleic Acids Res.* **1992**, 20(12), 3175-3178.

[34] Pullman, A.; Pullman, B. Molecular electrostatic potential of the nucleic acids. *Quart. Rev. Biophys.* **1981**, 14, 289-380.

[35] Price, C.C.; Gaucher, G.M.; Koneru, P.; Shibakawa, R.; Sawa, J.R.; Yamaguchi, M. Relative reactivities for monofunctional nitrogen mustard alkylation of nucleic acid components. *Biochem. Biophys. Acts.* **1968**, 166(2), 327-359.

[36] McHugh, P.J.; Gill, R.D.; Waters, R.; Hartley, J.A. Excision repair of nitrogen mustard-DNA adducts in saccharomyces cerevisiae. *Nucleic Acid Res*., **1999**, 27(16), 3259-3266.

[37] Graham, A.; Fox, M. The role of suppression of DNA synthesis and inhibition of cell cycle progression in cellular sensitivity to alkylation damage. *Carcinogenesis*, **1983**, 4(3), 269-274.

[38] Carr, F.J.; Fox, B.W. The effects of bifunctional alkylating agents on DNA synthesis in sensitive and resistant Yoshida cells. *Mut. Res*., **1962**, 95(2-3), 441-456.

[39] Murray, D.; Meyn, R.E. Cell cycle-dependent cytotoxicity of alkylating agents: determination of nitrogen mustard-induced DNA cross-links and their repair in chinese hamster ovary cells synchronized by centrifugal elutriation. *Cancer Res.,* **1986**, 46(5), 2324-2329.

[40] Dean, S.W.; Gibson, N.W.; Tea, K.D. Selection of nitrogen mustard resistance in a rat tumour cell line result in loss of guanine-O6-alkyltransferase activity. *Mol. Pharmacol*., **1986**, 30(1), 77-80.

[41] Erickson, L.C.; Bradley, M.O.; Ducore, J.M.; Ewig, R.A.G.; Kohn, K.W. DNA crosslinking and cytotoxicity in normal and transformed human cells treated with antitumor nitrosoureas. *Proc. Natl. Acad. Sci. USA,* **1980**, 77: 467-471.

[42] Erickson, L.C.; Laurent, G.; Sharkey, N.A.; Kohn, K.W. DNA cross-linking and monoadduct repair in nitrosourea-treated human tumour cells. *Nature*, **1980**, 288(5792), 727-729.

[43] Kohn, K.W. DNA as target in cancer chemotherapy: measurement of macromolecular DNA damage produced in mammalian cells by anticancer agents and carcinogens. *Cancer Res.,* **1978**, 16, 291-345.

[44] Wheeler, G.P.; Bowdon, B.J.; Adamson, D.J.; Vail, M.H. Effects of certain nitrogen mustards upon the progression of cultured H.Ep. No. 2 cells through the cell cycle. *Cancer Res.,* **1970**, 30(1), 100-111.

[45] Clarkson, J.M.; Mitchell, D.L. The importance of DNA damage and repair in the cell cycle sensitivity of CHO cells to nitrogen mustard. *Radiation Res.,* **1981**, 88, 587-596.

[46] Hansson, J.; Lewensohn, R.; Ringborg, U.; Nilsson, B. Formation and removal of DNA cross-links induced by melphalan and nitrogen mustard in relation to drug-induced cytotoxicity in human melanoma cells. *Cancer Res*., **1987**, 47(10), 2631-2637.

[47] Oida, T.; Humphres, W.G.; Guengerich, P.F. Preparation and characterization of oligonucleotides containing S-[2-(N7-guanyl)ethyl]glutathione. *Biochemistry*, **1991**, 30(43), 10513-10522.

[48] Presmark, M.; Guengerich, F.P. Spectroscopic and thermodynamic characterization of the interaction of N7-guanyl thioether derivatives of d(TGCTG*CAAG) with potential complements. *Biochemistry*, **1994**, 33(29), 8662-8672.

[49] Gerchman, L.L.; Ludlum, D.B. The properties of O6 -methylguanine in templates for RNA polymerase. *Biochimica et Biophysica Acta*, **1973**, 308(2), 310-316.

[50] Lawley, P.D.; Martin, C.N. Molecular mechanisms in alkylation mutagenesis. Induced reversion of bacteriophage T4rII AP72 by ethyl methanesulphonate in relation to extent and mode of ethylation of purines in bacteriophage deoxyribonucleic acid. *J. Biochem.,* **1975**, 145(1), 85-91.

[51] Abbott, P.J.; Saffhill, R. DNA synthesis with methylated poly(dC-dG) templates. Evidence for a competitive nature to miscoding by O6-methylguanine. *Biochem. Biophys. Acta,* **1979**, 562(1), 51-61.

[52] Mehta, J.R.; Ludlum, D.B. Synthesis and properties of O6-methyldeoxyguanylic acid and its copolymers with deoxycytidylic acid. *Biochem. Biophys. Acta,* **1978**, 521(2), 770-778.

[53] Verley, W.G. Prereplicative error-free DNA repair. *Biochem. Pharmacol.,* **1980**, 29(7), 977-982.

[54] Gates, K.S.; Nooner, T.; Dutta, S. Biologically relevant chemical reactions of N7- alkylguanine residues in DNA. *Chem. Res. Toxicol.,* **2004**, 17(7), 839-856.

[55] Lindahl, T.; Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry,* **1972**, 11(19), 3610-3618.

[56] Lawley, P.D.; Brookes, P. Further studies on the alkylation of nucleic acids and their constituent nucleotides. *Biochem. J.,* **1963**, 89(1), 127-138.

[57] Sowers, L.C.; Sedwick, W.D.; Ramsay, S.B. Hydrolysis of N3-methyl-2'-deoxycytidine: model compound for reactivity of protonated cytosine residues in DNA. *Mut. Res.,* **1989**, 215(1), 131-138.

[58] Lawley, P.D.; Warren, W. Removal of minor methylation products 7-methyladenine and 3-methylguanine from DNA of Escherichia coli treated with dimethyl sulphate. *Chem. Biol. Int.,* **1976**, 12(2), 211-220.

[59] Moschel, R.C.; Hudgins, W.R.; Dipple, A.J. Substituent-induced effects on the stability of benzylated guanosines: model systems for the factors influencing the stability of carcinogen-modified nucleic acids. *J. Org. Chem.,* **1984**, 49(2), 363-372.

[60] Muller, N.; Eisenbrand, G. The influence of N7 substituents on the stability of N7-alkylated guanosines. *Chem. Biol. Int.,* **1985**; 53: 173-181.

[61] Haines, J.A.; Reese, C.B.; Todd, L.J. Alkaline opening of imidazole ring of 7-methylguanosine. 1. Analysis of the resulting pyrimidine derivatives. *Chem. Biol. Int.,* **1982**, 41, 217-233.

[62] Townsend, L.B.; Robins, R.K. Alkaline opening of imidazole ring of 7-methylguanosine. 2. Further studies on reaction mechanisms and products. *Chem. Biol. Int.,* **1982**, 41, 235-249.

[63] Hecht, S.M.; Adams, B.L.; Kozarich, J.W. Chemical transformations of 7,9-disubstituted purines and related heterocycles. Selective reduction of imines and immonium salts. *J. Org. Chem.,* **1976**, 41(13), 2303-2311.

[64] Humphreys, W.G.; Guengerich, F.P. Structure of formamidopyrimidine adducts as determined by NMR using specifically 15N-labeled guanosine. *Chem. Res. Toxicol*., **1991**, 4(6), 632-636.

[65] Fujii, T.; Itaya, T. The Dimroth rearrangement in the adenine series: A review updated. *Heterocycles,* **1998**, 48(2), 359-390.

[66] Fujii, T.; Saito, T.; Nakasaka, T. 3-Methyladenosine and 3-Methyl-2'-deoxyadenosine : Their synthesis, glycosidic hydrolysis, and ring fission. *Chem. Pharm. Bull*., **1989**, 37, 2601-2609.

[67] Hendler, S.; Furer, E.; Srinivasan, P.R. Synthesis and chemical properties of monomers and polymers containing 7-methylguanine and an investigation of their substrate or template properties for bacterial deoxyribonucleic acid or ribonucleic acid polymerases. *Biochemistry,* **1970**, 9(21), 4141-4153.

[68] Boger, D.L.; Yun, W. Reversibility of the duocarmycin A and SA DNA alkylation reaction. *J. Am. Chem. Soc.,* **1993**, 115(21), 9872-9873.

[69] Asai, A.; Nagamura, S.; Saito, H.; Takahashi, I.; Nakano, H. The reversible DNA-alkylating activity of duocarmycin and its analogues. *Nucleic Acids Res*., **1994**, 22(1), 88-93.

[70] Warpehoski, M.A.; Harper, D.E.; Mitchell, M.A.; Monroe, T.J. Reversibility of the covalent reaction of CC-1065 and analogues with DNA. *Biochemistry*, **1992**, 31(9), 2502-2508.

[71] Everett, J.L.; Roberts, J.J.; Ross, W.C. Aryl-2-halogenoalkylamines. Part III. Some carboxylic derivative of N,N-Di-2-chloroethylaniline. *J. Chem. Soc*., **1953**, 12, 2386-2392.

[72] Elderfield, R.C.; Covey, I.S.; Geiduschek, J.B.; Mayer, W.L.; Ross, A.B.; Ross, J.H. Synthesis of potential anticancer agents. I. Nitrogen mustards derived from p-[N,N-Bis(2-chloroethyl)amino]benzaldehyde. *J. Org. Chem.,* **1958**, 23, 1749-1753.

[73] Kagan, F.; Birkenmeyer, R.B.; Strube, R.E. Nitrogen mustard derivatives containing the phosphonate group. *J. Am. Chem. Soc.,* **1959**, 81, 3026-3031.

[74] Creech, H.J.; Breunlnger, E.; Hankwitz, R.F.; Poisky, JrG.; Wilson, M.L. Quantitative studies of the effects of nitrogen mustard analogs and other alkylating agents on ascites tumors in mice. *Cancer. Res*., **1960**, 20(7), 471-494.

[75] Bajusz, S.; Janaky, T.; Csernus, V.J.; Bokser, L.; Fekete, M.; Srkalovic, G.; Redding, T.W.; Schally, A.V. Highly potent analogues of luteinizing hormone-releasing hormone containing D-phenylalanine nitrogen mustard in position 6. *Proc. Nat. Acad. Sci. USA,* **1989**, 86(16), 6318-6322.

[76] Hitchings, G.H.; Rhodes, C.P. 6-Mercaptopurine. *Ann. NY Acad. Sci.,* **1954**, 185, 507-519.
[77] Robins, R.K. Antitumor activity and structural relationships of purine derivatives and related compounds against neoplasms in experimental animals. *J. Med. Chem*., **1964**, 7(2), 186-199.
[78] Montgomery. J.A. The relation of anticancer activity to chemical structure: a review. *Cancer Res*., **1959**, 19(5), 447-463.
[79] Skipper, H.E.; Bennett, L.L.Jr. Biochemistry of cancer. *Ann. Rev. Biochem*., **1958**, 27(1), 137-166.
[80] Lister, J.H.; Timmis, G.M. Potential anti-purines. Part III. Some 9-dialkylamino-alkyl- purines and -8-azapurines. *J. Chem. Soc*., **1960**, 327-334.
[81] Lin, H.; Price, C. Some 9-Bis(2-chloroethyl)aminoalkylpurines. *J. Org. Chem*., **1961**, 26(1), 266-267.
[82] Ellis, G.P.; West, G.B.; Eds. Progress in Medicinal Chemistry. London: Butterworth; 1971.
[83] Kennedy, B.J. Athanasios theologides, uracil mustard, a new alkylating agent for oral administration in the management of patients with leukemia and lymphoma. *N. Engl. J. Med.*, **1961**, 264, 790-793.
[84] Smith, S.J.; Busch, H. Effects of uracil mustard on in vivo incorporation of precursors into nucleic acids of the walker tumor. *Texas Rept. Biol. Med.,* **1964**, 22, 731-740.
[85] Abell, C.W.; Rosini, L.A.; DiPaolo, J.A. Effects of uracil mustard on DNA, RNA, and protein biosynthesis in tissues of A/J mice. *Cancer Res*., **1967,** 27(6), 1101-1108.
[86] Petering, H.G.; Buskirk, H.H; Musser, E.A; Evans, J.S. A comparison of the biological activity of uracil mustard and dopan. *Cancer Chemother. Rept*., **1963**, 27, 1-10.
[87] Ross, L.O; Lee, W.W.; Schelstraete, M.G.M.; Goodman, L.; Baker, B.R. Potential anticancer agents. LVI. Synthesis of 5-[Bis(2-chloroethyl) aminomethyl]uracil. *J. Org. Chem.,* **1961**, 26(8), 3021-3022.
[88] Elderfield, R.C.; Wood, J.R. Application of the Mannich reaction with β,β-Dichlorodiethylamine to derivatives of uracil. *J. Org. Chem*., **1961**, 26(8), 3042-3043.
[89] Jonsson, G.; Hogberg, B.; Scand, N.T. Treatment of advanced prostatic carcinoma with estramustine phosphate (Estracyt). *J. Urol. Nephrol*., **1977,** 11(3), 231-238.
[90] Prakash, A.S.; Denny, W.A.; Gourdie, T.A.; Valu, K.K.; Woodgate, P.D.; Wakelin, L.P.G. DNA-directed alkylating ligands as potential antitumor agents: sequence specificity of alkylation by intercalating aniline mustards. *Biochemistry,* **1990**, 29(42), 9799-9807.
[91] Kohn, K.W.; Orr, A.; O'Connor, P.M.; Guziec, L.J.; Guziec, F.S.Jr. Synthesis and DNA-sequence selectivity of a series of mono- and difunctional 9-aminoacridine nitrogen mustards. *J. Med. Chem*., **1994**, 37(1), 67-72.
[92] Fan, J.Y.; Valu, K.K.; Woodgate, P.D.; Baguley, B.C.; Denny, W.A. Aniline mustard analogues of the DNA-intercalating agent amsacrine: DNA interaction and biological activity. *Anti-Cancer Drugs Des*., **1997**, 12(3), 181-203.
[93] Bacherikov, V.A.; Chou, T.C.; Dong, H.J.; Chen, C.H.; Lin, Y.W.; Tsai, T.J.; Su, T.L. Potent antitumor N-mustard derivatives of 9-anilinoacridine, synthesis and antitumor evaluation. *Bioorg. & Med. Chem. Lett*., **2004**, 14(18), 4719-4722.
[94] Su, T.L.; Lin, Y.W.; Chou, T.C.; Zhang, X., Bacherikov, V.A.; Chen, C.H.; Liu, L.F.; Tsai, T.J. Potent antitumor 9-anilinoacridines and acridines bearing an alkylating N-mustard residue on the acridine chromophore: synthesis and biological activity. *J. Med. Chem*., **2006**, 49(12), 3710-3718.
[95] Kapuriya, N.; Kapuriya, K.; Zhang, X.; Chou, T.C.; Kakadiya, R.; Wu, Y.T.; Tsai, T.H; Chen, Y.T.; Lee, T.C; Shah, A; Naliapara, Y.; Su, T.L. Synthesis and biological activity of stable and

potent antitumor agents, aniline nitrogen mustards linked to 9-anilinoacridines via a urea linkage. *Bioorg. & Med. Chem*., **2008**, 16(10), 5413-5423.

[96] Bacherikov, V.A.; Chou, T.C.; Dong, H.J.; Zhang, X.; Chen, C.H.; Lin, Y.W.; Tsai, T.J.; Lee, R.Z.; Liu, L.F.; Su, T.L. Potent antitumor 9-anilinoacridines bearing an alkylating N-mustard residue on the anilino ring: synthesis and biological activity. *Bioorg. & Med. Chem.*, **2005**, 13(12), 3993-4006.

[97] Chen, C.H.; Lin, Y.W.; Zhang, X.; Chou, T.C.; Tsai, T.J.; Kapuriya, N.; Kakadiya, R.; Su,T.L. Synthesis and in vitro cytotoxicity of 9-anilinoacridines bearing N-mustard residue on both anilino and acridine rings. *Eur. J. Med. Chem.*, **2009**, 44(7), 3056-3059.

[98] Broggini, M.; Coley, H.M.; Mongelli, N.; Pesenti, E.; Wyatt, M.D.; Hartley, J.A.; D'Incalci, M. DNA sequence-specific adenine alkylation by the novel antitumor drug tallimustine (FCE 24517), a benzoyl nitrogen mustard derivative of distamycin. *Nucleic Acids Res*., **1995**, 23(1), 81-87.

[99] Wyatt, M.D.; Lee, M.; Garbiras, B.J.; Souhami, R.L.; Hartley, J.A. Sequence specificity of alkylation for a series of nitrogen mustard-containing analogues of distamycin of increasing binding site size: evidence for increased cytotoxicity with enhanced sequence specificity. *Biochemistry.,* **1995**, 34(40), 13034-13041.

[100] Baraldi, P.G.; Cozzi, P.; Geroni, C.; Mongelli, N.; Romagnolia, R.; Spalluto, G. Novel benzoyl nitrogen mustard derivatives of pyrazole analogues of distamycin A: synthesis and antileukemic activity. *Bioorg. & Med. Chem*., **1999**, 7(2), 251-262.

[101] Baraldi, P.G.; Romagnoli, R.; Pavani, M.G.; Nunez, MdelC.; Bingham, J.P.; Hartley, J.A. Benzoyl and cinnamoyl nitrogen mustard derivatives of benzoheterocyclic analogues of the tallimustine: synthesis and antitumour activity. *Bioorg. & Med. Chem*., **2002**, 10(5), 1611-1618.

[102] Baraldi, P.G.; Beria, I.; Cozzi, P.; Geroni, C.; Espinosa, A.; Gallo, M.A.;, Entrena, A.; Bingham, J.P.; Hartley, J.A.; Romagnoli, R. Cinnamoyl nitrogen mustard derivatives of pyrazole analogues of tallimustine modified at the amidino moiety: design, synthesis, molecular modeling and antitumor activity studies. *Bioorg. & Med. Chem*., **2004**, 12(14), 3911-3921.

[103] Cozzi, P.; Beria, I.; Biasoli, G.; Caldarelli, M.; Capolongo, L.; Geroni, C.; Mongelli, N. Novel phenyl nitrogen mustard and half-mustard derivatives of amidino-modified distamycin. *Bioorg. & Med. Chem. Lett*., **1997**, 7(15), 2979-2984.

[104] Cozzi, P.; Beria, I.; Biasoli, G.; Caldarelli, M.; Capolongo, L.; D'Alessio, R.; Geroni, C.; Ragg, S.; Mazzini, E.; Rossi, C.; Mongelli, N. Novel phenyl nitrogen mustard and half-mustard derivatives of distamycin A. *Bioorg. & Med. Chem. Lett.,* **1997**, 7(23), 2985-2990.

[105] Wang, Y.; Wright, S.C.; Larrick, J.W. Synthesis and preliminary cytotoxicity of nitrogen mustard derivatives of distamycin A. *Bioorg. & Med. Chem. Lett*., **2003**, 13(3), 459-461.

[106] Wang, Y.; Yang, Z.; Wright, S.C.; Larrick, J.W. Synthesis and preliminary antitumor activity of distamycin nitrogen mustards. *Lett. Drug. Desig. & Discovery*, **2007**, 4(1), 37-39.

[107] Singh, R.K.; Prasad, D.N.; Bhardwaj, T.R. Design, synthesis, chemical and biological evaluation of brain targeted alkylating agent using reversible redox prodrug approach. *Arab. J. Chem*., **2014** jan. Available from http://dx.doi.org/10.1016/j.arabjc.2013.12.008.

[108] Bodor, N.; Venkatraghavan, V.; Windwood, D.; Estes, K.; Brewster, E. Improved delivery through biological membranes. XLI. Brain-enhanced delivery of chlorambucil. *Int. J. Pharm.*, **1989**, 53(1), 195-208.

[109] Prokai, L.; Prokai-Tatrai, K.; Bodor, N. Targeting drug to the brain by redox chemical drug delivery system. *Med. Res. Rev*., **2000**, 20(5), 367-416.

[110] El-Sherbeny, M.A.; Al-Salem, H.S.; Sultan, M.A.; Radwan, M.A., Farag, H.A.; El-Subbagh, H.I. Synthesis in vitro & in vivo evaluation of a delivery system for targeting anticancer drug to brain. *Arch. Pharm*., **2003**, 336(10), 445-455.

[111] Singh, R.K.; Prasad, D.N.; Bhardwaj, T.R. Synthesis, in vitro/in vivo evaluation and in silico physicochemical study of prodrug approach for brain targeting of alkylating agent. Med. Chem. Res., **2013**, 22(11), 5324-5336.

[112] Singh, R.K.; Kumar, S.; Prasad, D.N.; Bhardwaj, T.R. Reversible redox delivery system based drug design, synthesis and evaluation of targeting anticancer agent across the brain. Med. Chem. Res., **2014**. Available from http://dx.doi.10.1007/s00044-013-0833-8.

[113] Singh, R.K.; Prasad, D.N.; Bhardwaj, T.R. Design, synthesis and evaluation of aminobenzophenone derivatives containing nitrogen mustard moiety as potential CNS antitumor agent. *Med. Chem. Res*., **2013**, 22(12), 5901-5911.

[114] Halmos, T.; Santarromana, M.; Antonakis, K.; Scherman, D. Synthesis of glucose-chlorambucil derivatives and their recognition by the human GLUT1 glucose transporter. Eur. J. Pharmacol., **1996**, 318(2/3), 477-484.

[115] Reux, B.; Weber, V.; Galmier, M.J.; Borel, M.; Madesclaire, M.; Madelmont, J.C.; Debiton, E.; Coudert, P. Synthesis and cytotoxic properties of new fluorodeoxyglucose-coupled chlorambucil derivatives. *Bioorg. & Med. Chem*., **2008**, 16(9), 5004-5020.

[116] Singh, R.K..; Prasad, D.N.; Bharadwaj, T.R. Synthesis, alkylation activity and physico-chemical evaluation of benzodiazepine linked nitrogen mustard agent to penetrate the blood-brain barrier. *Asian J. Chem*., **2012**, 24(12), 5605-5608.

[117] Giraud, I.; Rapp, M.; Maurizis, J.C.; Madelmont, J.C. Synthesis and in vitro evaluation of quaternary ammonium derivatives of chlorambucil and melphalan, anticancer drugs designed for the chemotherapy of chondrosarcoma. *J. Med. Chem*., **2002**, 45(10), 2116-2119.

[118] Pain, A.; Samanta, S.; Dutta, S; Saxena, A.K.; Shanmugavel, M.; Kampasi, H.; Qazi, G.N.; Sanyal, U.J. Evaluation of naphthalmustine, a nitrogen mustard derivative of naphthalimide as a rationally-designed anticancer agent. *J. Exp. Clin. Cancer Res*., **2003**, 22, 411-418.

[119] Faissat, L.; Martin, K.; Chavis, C.; Montero, J.L.; Lucas, M. New nitrogen mustards structurally related to (L)-Carnitine. *Biorg. & Med. Chem*., **2003**, 11(3), 325-334.

[120] Leiris, S.; Lucas, M.; d'Angeac, A.D.; Morère, A. Synthesis and biological evaluation of cyclic nitrogen mustards based on carnitine framework. *Eur. J. Med. Chem*., **2010**, 45(9), 4140-4148.

[121] Urbaniak, M.D.; Bingham, J.P.; Hartley, J.A.; Woolfson, D.N.; Caddick, S. Design and synthesis of a nitrogen mustard derivative stabilized by apo-neocarzinostatin. *J. Med. Chem.*, **2004**, 47(19), 4710-4715.

[122] Coggiola, B.; Pagliai, F.; Allegrone, G.; Genazzani, A.A.; Tron, G.C. Synthesis and biological activity of mustard derivatives of combretastatins. *Bioorg. & Med. Chem. Lett*., **2005**, 15(15), 3551-3554.

[123] Zheng, Q.Z.; Zahang, F.; Cheng, K.; Yang, Y.; Chen, Y.; Qian, Y.; Zhang, H.J.; Li, H.Q.; Zhou, C.F.; Jiao, S.Q.; An, Q.C.; Zhu, H.L. Synthesis, biological evaluation and molecular docking studies of amide-coupled benzoic nitrogen mustard derivatives as potential antitumor agents. *Bioorg. & Med. Chem*., **2010**, 18(2), 880-886.

[124] Zhao, H.; Meng, X.; Yuan, H.; Lan, M. Novel melphalan and chlorambucil derivatives of 2,2,6,6-Tetramethyl-1-piperidinyloxy radicals: synthesis, characterization, and biological evaluation in vitro. *Chem. Pharm. Bull*., **2010**, 58(3), 332-335.

[125] Marvania, B.; Lee, P.C.; Chaniyara, R.; Dong, H.; Suman, S.; Kakadiya, R.; Chou, T.C.; Lee, T.C.; Shah, A.; Su, T.L. Design, synthesis and antitumor evaluation of phenyl N-mustard-quinazoline conjugates. *Bioorg. & Med. Chem*., **2011**, 19(6), 1987-1998.

[126] Boens, B.; Teste, K.; Bouazza, A.H.; Ismaili, J.; Zerrouki, R.; New thymine-based derivative of nitrogen mustards. *Nucleosides, Nucleotides & Nucleic Acids,* **2012**, 31(3), 197-205.
[127] Amos, H.; Vollmayer, E. Effect of pentamidine on the growth of Escherichia coli. *J. Bacteriol*., **1957**, 73(2), 172-177.
[128] Libman, M.D.; Miller, M.A.; Richards, G.K. Antistaphylococcal activity of pentamidine. *Antimicrob. Agents Chemother*., **1990**, 34(9), 1795-1796.
[129] St-Germain, G. Effects of pentamidine alone in combination with ketaconazole or itraconazole on the growth of Canadida albicans. *Antimicrob. Agents Chemother*., **1990**, 34(12), 2304-2306.
[130] Puckowaska, A.; Bielawski, K.; Bielawska, A.; Nowaczek, K.M. Aromatic analogues of DNA minor groove binders-synthesis and biological evaluation. *Eur. J. Med. Chem*., **2004**, 39(1), 99-105.
[131] Drozdowaska, D.; Rusak, M.; Bielawski, T.; Bruzgo, I.; Nowaczek, K.M. Amino and chlorambucil analogues of pentamidine-synthesis and biological examinations. *Acta Poloniae Pharmaceutica Drug Res*., **2012**, 69(1), 63-73.
[132] Yang, Y.; Mao, W.J.; Li, H.Q.; Zhu, T.T.; Shi, L.; Lv, P.C.; Zhu, H.L. Synthesis and biological evaluation of 7-O-modified formononetin derivatives. Res. Lett. Org. Chem., **2008** (1-4) Available from http://dx.doi.10.1155/2008/209830.
[133] Ren, J.; Xu, H.J.; Cheng, H.; Xin, W.Q.; Chen, X.; Hu, K. Synthesis and antitumor activity of formononetin nitrogen mustard derivatives. *Eur. J. Med. Chem*., **2012**, 54(1), 1-13.
[134] Omoomi, F.D.; Siadat, S.D.; Nourmohammadi, Z.; Tabasi, M.A.; Pourhoseini, S.; Babaei, R.A.; Saffari, M.; Ardestani, M.S. Molecular chlorambucil-methionine conjugate: novel anticancer agent against Breast MCF-7 cell Model. *J. Cancer Sci. Theor*., **2013**, 5(1), 75-84.
[135] Scutaru, A.M.; Wenzel, M.; Gust, R. Bivalent bendamustine and melphalan derivatives as anticancer agents. *Eur. J. Med. Chem*., **2011**, 46(5), 1604-1615.
[136] Saha, P.; Debnath, C.; Bérubé, G. Steroid-linked nitrogen mustards as potential anticancer therapeutics: A review. *J. Steroid Biochem. & Mol. Biol.,* **2013**, 137, 271-300.
[137] Nogrady, T.; Vagi, K.M.; Adamkiewicz, V.W. N-mustrad derivatives of estrogens. *Canad. J. Chem*., **1962**, 40(11), 2126-2129.
[138] Jones, J.B.; Adam, D.J.; Leman, J.D. Steroids and steroidases: studies on some potentially antitumor active androstane compounds containing 2–17 nitrogen mustard functions. *J. Med. Chem*., **1971**, 14(9), 827-833.
[139] Duvaz, I.N.; Cambanis, A.; Tarnauceanu, E. Potential anticancer agents. II: Urethan-type nitrogen mustards of some natural sex hormones. *J. Med. Chem*., **1967**, 10(2), 172-174.
[140] Tew, K.D.; Hartley-Aps, B. Cytotoxic properties of estramustine unrelated to alkylating and steroid constituents. *J. Urol*., **1984**, 23(1), 28-33.
[141] Joensson, G.; Högberg, B.; Nilsson, T. Treatment of advanced prostatic car-cinoma with estramustine phosphate (Estracyt) Scand. *J. Urol. Nephrol*., **1977**, 11(3), 231-238.
[142] Hartley, B.; Gunnarsson, P.O. Growth and cell survival following treatment with estramustine, nor-nitrogen mustard, estradiol and testosterone of a human prostatic cell line. *J. Urol*., **1982**, 127(4), 818-822.
[143] Lidstroem, P.; Bonasera, T.A.; Marquez, M.M.; Nilsson, S.; Bergstroem, M.; Langstroem, B. Synthesis and in vitro evaluation of [carbonyl-11C] estramustine and phosphate. *Steroids*, **1988**, 63(4), 228-234.
[144] Firestone, R.A.; Pisano, J.M.; Falck, J.R.; McPhaul, M.M.; Krieger, M. Selective delivery of cytotoxic compounds to cells by the LDL pathway. *J. Med. Chem*., **1984**, 27(8), 1037-1043.

[145] Dubowchik, G.M.; Firestone, R.A. The synthesis of branched steroidal prodrugs of nitrogen mustard for antitumor targeting via reconstituted LDL. *Tetrahedron Lett*., **1994**, 35(26), 4523-4526.

[146] Foster, E.L.; Blickenstaff, R.T. Synthesis of steroidal cyclophosphamides. *Steroids,* **1976**, 27(3), 353-360.

[147] Kim, J.C.; Paek, H.D.; Moon, S.H.; Kim, S.H. Synthesis of steroidal cyclophosphamide. 2-Bis(2-chloroethyl)amino-2-oxo-6-(5-cholestanyl)-1,3,2-oxazaphosphorinane. *Bull. Kor Chem. Soc*., **1993**, 14(3), 318-319.

[148] Zhang, H.B.; Xue, J.J.; Zhao, X.L.; Liu, D.G.; Li, Y. Synthesis and biological evaluation of novel steroid-linked nitrogen mustards. *Chin. Chem. Lett*., **2009**, 20(6), 680-683.

[149] Shervington, L.A.; Smith, N.; Norman, E.; Ward, T.; Phillips, R.; Shervington, A. To determine the cytotoxicity of chlorambucil and one of its nitro-derivatives, conjugated to prasterone and pregnenolone, towards eight human cancercell-lines. *Eur. J. Med. Chem*., **2009**, 44(7), 2944-2951.

[150] Gupta, A.; Saha, P.; Descoteaux, C.; Leblanc, V.; Asselin, E.; Berube, G. Design, synthesis and biological evaluation of estradiol–chlorambucil hybrids as anticancer agents. *Bioorg. & Med. Chem. Lett*., **2010**, 20(5), 1614-1618.

[151] Marquis, J.C.; Hillier, S.M.; Dinaut, A.N.; Rodrigues, D.; Mitra, K.; Essigmann, J.M.; Croy, R.G. Disruption of gene expression and induction of apoptosis in prostate cancer cells by a DNA-damaging agent tethered to an androgen receptor ligand. *Chem. & Biol*., **2005**, 12(7), 779-787.

[152] Lober, J.; Mouridsen, T.H.; Christiansen, I.E.; Dombernowsky, P.; Mattsson, W.; Rorth, M. A phase III trial comparing prednimustine (LEO 1031) to chlorambucil plus prednisolone in advanced breast cancer. *Cancer*, **1983,** 52(9), 1570-1576.

[153] Arsenou, E.S.; Fousteris, M.A.; Koutsourea, A.I.; Papageorgiou, A.; Karayianni, V.; Mioglou, E.; Iakovidou, Z.; Mourelatos, D.; Nikolaropoulos, S.S. The allylic 7-ketoneat the steroidal skeleton is crucial for the antileukemic potency of chlorambucil's active metabolite steroidal esters. *Anti-Cancer Drugs*, **2004**, 15(10), 983-990.

[154] Fousteris, M.A.; Koutsourea, A.I.; Arsenou, E.S.; Papageorgiou, A.; Mourelatos, D.; Nikolaropoulos, S.S. Structure–anti-leukemic activity relationship study of B- and D-ring modified and non-modified steroidal esters of chlorambucil. *Anti-Cancer Drugs*, **2006**, 17(5), 511-519.

[155] Fousteris, M.A.; Koutsourea, A.I.; Lagonikakos, N.P.; Arsenou, E.S.; Spyri-donidou, C.; Mourelatos, D.; Nikolaropoulos, S.S. Rational design, synthesis and in vitro evaluation of three new alkylating steroidal esters. *Med. Chem*., **2006**, 2(6), 569-576.

[156] Dalmases, P.; Cervantes, G.; Quintana, J.; Bonet, J.J.; Giner-Sorolla, A.; Schmid, F.A. Antineoplastic agents. V: Nitrogen mustards of systematically modified steroidal ring A lactams. *Eur. J. Med. Chem*., **1984**, 19(5), 465-467.

[157] Koutsourea, A.I.; Fousteris, M.A.; Arsenou, E.S.; Papageorgiou, A.; Pairas, G.N.; Nikolaropoulos, S.S. Synthesis, in vivo antileukemic evaluation and comparative study of novel 5-7-keto steroidal esters of chlorambucil and its active metabolite. *In Vivo,* **2008**, 22(3), 345-352.

[158] Karapidaki, I.; Bakopoulou, A.; Papageorgiou, A.; Iakovidou, Z, Mioglou E. Nikolaropoulos S, Mourelatos D, Lialiaris T. Genotoxic, cytostatic, antineoplastic and apoptotic effects of newly synthesized antitumor steroidal esters. *Mutat. Res*., **2009**, 675(1), 51-59.

[159] Papageorgiou, A.; Tsavdaridis, D.; Geromichalos, G.D.; Camoutsis, C.; Karaberis, E.; Mourelatos, D.; Chrysogelou, E.; Houvartas, S.; Kotsis, A. A comparative study on cytogenetic and antineoplastic effects induced by two modified steroidal alkylating agents. *Cancer Detection & Prevention,* **2001**, 25(4), 369-374.

[160] Camoutsis, C.; Trafalis, D.; Pairas, G.; Papageorgiou, A. On the formation of 4-[N,N-bis(2-chloroethyl)amino]phenyl acetic acid esters of hecogenin andaza-homo-hecogenin and their antileukemic activity. *Farmaco*., **2005**, 60(10), 826-829.

[161] Camoutsis, C.; Mourelatos, D.; Pairas, G.; Mioglou, E.; Gasparinatou, C.; Iakovidou, Z. Synthesis and cytogenetic studies of structure–biological activity relationship of esters of hecogenin and aza-homo-hecogenin with N, N-bis(2-chloroethyl)aminocinnamic acid isomers. *Steroids,* **2005,** 70(9), 586-593.

[162] Wong, M.W. Vibrational frequency prediction using density functional theory. *Chem. Phys. Lett*., **1996,** 256(4-5), 391-399.

[163] Lieb, E.H. Density Functionals for Coulomb Systems. In: *Density functional methods in Physics;* Dreizler, R. M.; Providˆencia, J.da., Eds.; Plenum press: New York, **1985**; pp. 31-58.

[164] Jones, R.O.; Gunnarsson, O. The density functional formalism, its applications and prospects. *Rev. Mod. Phys*., **1989**, 61, 689-746.

[165] Koch, W.; Holthausen, M.C. *A chemist's guide to density functional theory*, 2nd ed.; Wiley-VCH: New York, **2001**.

[166] Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev. B*, **1964,** 136(3), B864-B871.

[167] Dreizler, R.M.; Gross, E.K.U. *Density Functional Theory*; Springer-Verlag: Berlin, **1990**.

[168] Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev. A.* **1965**, 140, 1133-1138.

[169] Dirac, P.A.M. Note on exchange phenomena in the Thomas-Fermi atom. *Proc. Cambridge Phil. Roy. Soc.* **1930**, 26(3), 376-385.

[170] Perdew, J.P. Unified Theory of Exchange and Correlation Beyond the Local Density Approximation. In: *Electronic structure of solids;* Ziesche, P.; Eschrig, H. Eds.; Akademie Verlag: Berlin, **1991**, pp. 11-20.

[171] Staroverov, V.N. Density-functional approximations for exchange and correlation. In: *A matter of density: Exploring the electron density concept in the chemical. Biological and material sciences*; Sukumar N. Ed.; Wiley: USA, **2012.**

[172] Vosko, S.H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Canad. J. Phys*., **1980**, 58(8), 1200-1211.

[173] Perdew, J.P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, **1981**, 23(10), 5048-5079.

[174] Cole, L.A.; Perdew, J.P. Calculated electron affinities of the elements. *Phys. Rev. A*, **1982**, 25(3), 1265-1271.

[175] Perdew, J.P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, **1992**, 45(23), 13244-13249.

[176] Perdew, J.P.; Wang, Y. Accurate and simple density functional for the electronic exchange energy. Generalized gradient approximation*. Phys. Rev. B*, **1986**, 33(12), 8800-8802.

[177] Perdew, J.P.; Chevary, J.A.; Vosko, S.H.; Jackson, K.A.; Pederson, M.R.; Singh, D.J. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B*, **1992**, 46(11), 6671-6687.

[178] Kohn, W. Nobel lecture: Electronic structure of matter-wave functions and density functional. *Rev. Mod. Phys*., **1999**, 71(5), 1253-1266.
[179] Perdew, J.P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple *Phys. Rev. Lett*., **1996**, 77(18), 3865-3868.
[180] Hammer, B.; Hansen, L.B.; Norskov, J.K. Improved adsorption energetics within density functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B*, **1999**, 59(11), 7413-7421.
[181] Zhang, Y.; Yang, W. Generalized gradient approximation made simple. *Phys. Rev. Lett*., **1998**, 80(4), 890-891.
[182] Perdew, J.P.; Ruzsinszky, A.; Csonka, G.I.; Vydrov, O.A.; Scuseria, G.E.; Constantin, L.A.; Zhou, X.; Burke, K. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett*., **2008**, 100(13), 136406-136410.
[183] Perdew, J.P.; Constantin, L.A.; Sagvolden, E.; Burke, K. Relevence of the slowly varying electron gas to atom, molecule and solids. *Phys. Rev. Lett*., **2006**, 97(2), 223002-223005.
[184] Tschinke, V.; Zieglar, T. On the shape of spherically averaged Fermi-hole correlation functions in density functional theory.1. Atomic system. *Canad. J. Chem*., **1989**, 67(3), 460-472.
[185] Neumann, R.; Handy, N.C. Higher-order gradient corrections for exchange functional. *Chem. Phys. Lett*., **1997**, 266(1-2), 16-22.
[186] Perdew, J.P.; Kurth, S.; Zupan, A.; Blaha, P. Accurate density functional with correct formal properties: A step beyond generalized gradient approximation. *Phys. Rev. Lett*., **1999**, 82(12), 2544-2547.
[187] Tao, J. An accurate MGGA-based hybrid exchange-correlation functional. *J. Chem. Phys*., **2002**, 116(6), 2335-2337.
[188] Kurth, S.; Perdew, J.P.; Blaha, P. Molecular and solid state test of density functional approximations: LSD, GGAs, and META-GGAs. *In. J. Quant. Chem*., **1999**, 75(4-5), 889-909.
[189] Adamo, C.; Ernzerhof, M.; Scuseria, G.E.J. The meta-GGA functional: Thermochemistry with a kinetic energy density dependent exchange-correlation functional. *J. Chem. Phys*., **2000**, 112(6), 2643-2649.
[190] Becke, A.D. A new mixing of Hartree-Fock and local density-functional theories. *J. Chem. Phys.*, **1993**, 98(2), 1372-1377.
[191] Perdew, J.P.; Ernzerhof, M.; Burke, K. Rational for mixing exact exchange with density functional approximation. *J. Chem. Phys*., **1996**, 105(22), 9982-9985.
[192] Becke, A.D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, **1993**, 98(7), 5648-5652.
[193] Lee, C.; Yang, W.; Parr, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev*. B, **1988**, 37(2), 785-789.
[194] Xu, X.; Goddard III, W.A. The X3LYP extended density functional for accurate descriptions of nonbond interactions, spin states, and thermochemical properties. *Proc. Natl. Acad. Sci. USA*, **2004**, 101(9), 2673-2677.
[195] Perdew, J.P.; Burke, K.; Ernzerhof, M. Errata: Generalized gradient approximation made simple. *Phys. Rev. Lett*., **1997**, 78(7), 1396-1396.
[196] Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys*., **1999**, 110(13), 6158-6169.
[197] Perdew, J.P.; Burke, K.; Yang, W. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Phys. Rev. B*, **1996**, 54(23), 16533-16539.

[198] Zhao, Y.; Truhlar, D.G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functional. *Theor. Chem. Acc*., **2008**, 120(1-3), 215-241.

[199] Cohen, A.J.; Anchez, P.M.; Yang, W. Challenges for density functional theory. *Chem. Rev*., **2012**, 112(1), 289-320.

[200] Parr, R.G.; Yang, W. Density-functional theory of the electronic structure of molecules. *Annu. Rev. Phys. Chem*., **1995**, 46, 701-728.

[201] Ayers, P.; Parr, R.G. Variational principles for describing chemical reaction: The fukui function and hardness revisited. *J. Am. Chem. Soc*., **2000**, 122(9), 2010-2018.

[202] Kohn, W.; Becke, A.D.; Parr, R.G. Density functional theory of electronic structure. *J. Phys. Chem.*, **1996**, 100(31), 12974-12980.

[203] De proft, F.; Geerlings, P. Conceptual density functional theory. *Chem. Rev*., **2003**, 103(5), 1793-1873.

[204] Fukui, K.; Yonezawa, T.; Shingu, H. A molecular orbital theory of reactivity in aromatic hydrocarbon. *J. Chem. Phys*., **1952**, 20(4), 722-725.

[205] Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H. Molecular orbital theory of orientation in aromatic, heteroaromatic and other conjugated molecule. *J. Chem. Phys*., **1954**, 22(8), 1433-1438.

[206] Fukui, K. *Theory of orientation and stereoselection*; Heidelberg: Berlin, Springer: New York, **1973.**

[207] Becke, A.D.; Edgecombe, K.E. A simple measure of electron localization in atomic and molecular system. *J. Chem. Phys*., **1990**, 92(9), 5397-5403.

[208] Savin, A.; Jepsen, O.; Flad, J.; Andersen, O.; Preuss, H.; Schnering, H.G.V. Electron localization in solid-state structures of the elements – the diamond structure. *Angew. Chem. In. Ed*., **1992**, 31(2), 187-188.

[209] Bonaccorsi, R.; Scrocco, E.; Tomasi, J. Molecular SCF calculations for the ground state of some three-membered ring molecule:$(CH_2)_3$, $(CH_2)_2NH$, $(CH_2)_2NH_2^+$, $(CH_2)_2O$, $(CH_2)_2S$, $(CH)_2CH_2$, and N_2CH_2. *J. Chem. Phys*., **1970**, 52(10), 5270-5284.

[210] Politzer, P. Electrostatic potential–electronic density relationships in atoms. *J. Chem. Phys*., **1980**, 72(5), 3027-3033.

[211] Truhlar, D.G.; Politzer, P. Introduction. In: *Chemical applications of atomic and molecular electrostatic potentials*; Politzer, P.; Truhlar, D.G., Eds.; Plenum press: New York, **1981.**

[212] Pullman, B. 30 years of Sanibel symposia: Structure and activity of biomolecules. *Int. J. Quant. Chem. Quant. Biol. Symp.,* **1990**, 17(1), 81-92.

[213] Tomasi, J.; Bonaccorsi, R.; Cammi, R., *Theoretical models of chemical bonding*; Maksic R. Eds.; Springer: Berlin, **1990**; pp. 230-268.

[214] Sanderson, R.T. An interpretation of bond lengths and a classification of bonds. *Science*, **1951**, 114(2973), 670-672.

[215] Mortier, W.J.; Ghosh, S.K.; Sahankar, S. Electronegativity-equalization method for the calculation of atomic charges in molecules. *J. Am. Chem. Soc*., **1986**, 108(15), 4315-4320.

[216] Baird, N.C.; Sichel, J.M.; Whitehead, M.A. Molecular orbital approach to electronegativity equalization. *Theor. Chim. Acta*, **1968**, 11(1), 38-50.

[217] Sanderson, R.T. *Chemical bonds and bond energy*, 2nd ed.; Academic Press: New York, 1976.

[218] Ray, N.K.; Samuels, L.; Parr, R.G. Studies of electronegativity equalization. *J. Chem. Phys*., **1979**, 70(7), 3680-3684.

[219] Iczkowski, R.P.; Margrave, J.L. Electronegativity. *J. Am. Chem. Soc*., **1961**, 83(17), 3547-3551.
[220] Parr, R.G.; Pearson, R.G. Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc*., **1983**, 105(26), 7512-7516.
[221] Perdew, J.P.; Parr, R.G.; Levy, M.; Balduz, J.L. Jr. Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Phys. Rev. Lett.,* **1982**, 49(23), 1691-1694.
[222] Zhang, Y.; Yang, W. Persppective on Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Theor. Chem. Acc*., **2000**, 103(3-4), 346-348.
[223] Parr, R.G.; Yang, W. *Density functional theory of atoms and molecules*; Oxford University Press: New York, **1989**.
[224] Koopmans, T., Uber die Zuordnung von Wellenfunktionen und Eigenwerten zu den einzelnenElektronen eines Atoms; *Physica*, **1933**, 1(1-6), 104-113.
[225] Parr, R.G.; Szentpaly, L.V.; Liu, S. Electrophilicity index. *J. Am. Chem. Soc*., **1999**, 121(9), 1922-1924.
[226] Chattaraj, P.K.; Maitti, B.; Sarkar, U. Philicity: An unified treatment of chemical reactivity and selectivity. *J. Phys. Chem. A*, **2003**, 107(25), 4973-4975.
[227] Roy, R.K. On the reliability of global and local electrophilicity descriptor. *J. Phys. Chem. A*, **2004**, 108(22), 4934-4939.
[228] Usha, R.K.; Paulovic, V.J.; Hirao, K. Are the local electrophilicity descriptor reliable indicators of global electrophilicity trends? *J. Phys. Chem. A,* **2005**, 109(20), 4601-4606.
[229] Chandra, A.K. Nguayen, M.T. Use of local softness for the interpretation of reaction mechanism. *Int. J. Mol. Sci*., **2002**, 3(4), 310-323.
[230] Yang, W. Mortier, W.J. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amine. *J. Am. Chem. Soc*., **1986**, 108(19), 5708-5711.
[231] Morell, C.; Grand, A.; Toro-Labbé, A. New dual descriptor for chemical reactivity. *J. Phys. Chem.* A, **2005**, 109(1), 205-212.
[232] Morell, C.; Grand, A.; Toro-Labbé, A. Theoretical support for using the Δf(r) descriptor**.** *Chem. Phys. Lett.,* **2006**, 425(4), 342-346.
[233] Bhakti, B.S.; Tanwar, A.; Pal, S. Reactivity descriptors and electron density analysis for ligand chemistry: A case study of 2,2-bipyridine and its analogues. *J. Chem. Sci*., **2007**, 119(5), 489-499.
[234] Chattaraj, P.K.; Nath, S.; Maiti, B. *Computational medicinal chemistry for drug discovery*. Tollenaere J.; Bultinck P, Winter HD, Langenaeker W. Eds.; Marcel Dekker: New York, **2003**.
[235] Chattaraj, P.K.; Sarkar, U.; Roy, D.R. Electrophilicity index. *Chem. Rev*., **2006**, 106(6), 2065-2091.
[236] Tanwar, A.; Pal, S. Separability of local reactivity descriptors. *J. Chem. Sci*., **2005**, 117(5), 497-505.
[237] Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Subramanian, V.; Chattaraj, P.K. Intermolecular reactivity through the generalized philicity concept. *Chem. Phys. Lett*., **2004**, 394(4-6), 225-230.
[238] Bulat, F.A.; Chamorro, E.; Fuentealba, P.; Toro-Labb´e, A. Condensation of frontier molecular orbital Fukui functions. *J. Phys. Chem. A*, **2004**, 108(2), 342-349.
[239] Chattaraj, P.K., Ed. *Chemical Reactivity Theory*; CRC press: Boca Raton, **2009**.
[240] De Proft, F.; Geerlings, P. Conceptual and computational DFT in the study of aromaticity. *Chem. Rev*., **2001**, 101(5), 1451-1464.

[241] Parr, R.G.; Yang, W. Density functional approach to the frontier-electron theory of chemical reactivity. *J. Am. Chem. Soc*., **1984**, 106(14), 4049-4050.

[242] Yang, W.; Parr, R.G. Hardness, softness, and the fukui function in the electronic theory of metal and catalysis. *Proc. Natl. Acad. Sci. USA*, **1985**, 82(20), 6723-6726.

[243] Fuentealba, P.; Perez, P.; Contreras, R. On the condensed Fukui function. *J. Chem. Phys*., **2000**, 113(7), 2544-2551.

[244] Russo, N.; Toscano, M.; Grand, A.; Mineva, T. Proton affinity and protonation sites of aniline. Energetic behaviour and density functional reactivity indices. *J. Phys. Chem. A*, **2000**, 104(17), 4017-4021.

[245] Nguyen, H.M.T.; Peeters, J.; Nguyen, M.T. Use of DFT-based reactivity descriptors for rationalizing radical reactions: A critical analysis, *J. Phys. Chem. A*, **2004**, 108(3), 484-489.

[246] Mendoza-Huizar, L.H.; Rios-Reyes, C.H. Chemical reactivity of atrazine employing the Fukui function, *J. Mex. Chem. Soc*., 2011, 55(3), 142-147.

[247] Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P.K. Group philicity and electrophilicity as possible descriptors for modeling ecotoxicity applied to chlorophenols. *Chem. Res. Toxicol*., **2006**, 19(3), 356-364.

[248] Chatterjee, A.; Iwasaki, T. A reactivity index study to choose the best template for a particular zeolite synthesis. *J. Phys. Chem*. *A*, **2001**, 105(25), 6187-6196.

[249] Hamza, A.; Borch, H.; Vasilescu, D. Quantum molecular study of S-methylated forms of sulphur mustard. *Chem. Phy*., **1996**, 204(2), 373-389.

[250] Shukla, P.K.; Mishra, P.C.; Suhai, S. Reactions of DNA bases with the anticancer nitrogen mustard mechlorethamine: A quantum chemical study. *Chem. Phy. Lett*. **2007**, 449(4-6), 323-328.

[251] Friedberg, E.C.; Walker, G.C.; Siede, W. *DNA Repair and Mutagenesis*; ASM Press: Washington, DC, USA, **1995.**

[252] Pineda, F.P.; Ortega-Castro, J.O.; Alvarez-Idaboy, J.R.; Frau, J.; Cabrera, B.M.; Ramrez, J.C.; Donoso, J.; Munoz, F. Hydrolysis of a chlorambucil analogue. A DFT study. *J. Phy. Chem. A*, **2011**, 115(11), 2359-2366.

[253] Vasilescu, D.; Adrian–Scotto, M.; Fadiel, A.; Hamza, A. Ab initio study of alkylation of guanine-cytosine base pair by sulfur and nitrogen mustards. *J. Biomol. Struct. & Dyn*., **2010**, 10(4), 465-476.

[254] Mann, D.J. Aziridinium ion ring formation from nitrogen mustards: Mechanistic insights from Ab initio dynamics. *J. Phys. Chem. A*, **2010**, 114(13), 4486-4493.

[255] Solimando, D.A. *Lexi-comp drug information handbook for oncology: A complete guide to combination chemotherapy regimens*, 6th Ed.; Lexi-Comp: Ohio, Hudson, **2007.**

[256] Goodman, G.; Limbird, L.; Molinoff, P.B.; Ruddon, R.W.; Gilman, A. *Goodman & Gilman's: The Pharmacological Basis of Therapeutics*, 10th ed.; McGraw-Hill: New York, **2001**.

[257] Polavarapu, A.; Stillabower, J.A.; Stubblefield, S.G.W.; Taylor, W.M.; Baik, M. The mechanism of guanine alkylation by nitrogen mustards: A computational study, *J. Org. Chem.*, **2012**, 77(14), 5914-5921.

[258] Bhattacharyya, P.K.; Kar, R. Does structural variation in the aziridinium ion facilitate alkylation? *Comput. Theor. Chem*., **2011**, 967(1), 5-11.

[259] Sarmah, N.; Neog, B.; Bhattacharyya, P.K. Affinity of aziridinium ion towards different nucleophiles: A density functional study. *Comput. Theor. Chem.,* **2011**, 976(1), 30-35.

[260] Neog, B.; Sarmah, N.; Kar, R.; Bhattacharyya, P.K. Effect of external electric field on aziridinium ion intermediate: A DFT study. *Comput. Theor. Chem.*, **2011**, 976(1), 60-67.

[261] Sinha, S.; Bhattacharyya, P.K. Variation of reactivity of aziridinium ion during alkylation. *Mol. Phy*., **2014**, 112(1), 14-21.

[262] Neog, B.; Sinha, S.; Bhattacharyya, P.K. Alkylation of DNA by nitrogen mustards: A DFT study. *Comput. Theor. Chem*., **2013**, 1018(1), 19-25.

[263] Sinha, S.; Bhattacharyya, P.K. Alkylation of guanine by formononetin nitrogen mustard derivatives: A DFT study. *Comput. Theor. Chem*., **2014**, 1027(1), 135-141.

FURTHER READINGS

Gately DP, Howell SB. Cellular accumulation of the anticancer agent cisplatin: A review. Br J Cancer 1993; 67: 1171–1176.

Ghanei M, Poursaleh Z, Harandi AA. Emadi SE, Emadi SN. Acute and chronic effects of sulfur mustard on the skin: A comprehensive review. Cutan Ocul Toxicol 2010; 29: 269-277.

Lie SO, Jonmundsson GK, Mellander L, Siimes MA, Yssing M, Gustafsson G. Chemotherapy of acute myelocytic leukemia in children. Ann N Y Acad Sci 1997; 824: 84-90.

Ghanei M, Harandi AA. Long term consequences from exposure to sulfur mustard: A review. Inhalation Toxico 2007; 19: 451-456.

Panahi Y, Sarayani A, Beiraghdar F, Amiri M, Davoudi SM, Sahebkar A. Management of sulfur mustard-induced chronic pruritus: A review of clinical trials. Cutaneous and Ocular Toxicol 2012; 31: 220-225.

Shakarjian MP, Heck DE, Gray JP, Sinko PJ, Gordon MK, Casillas RP, Heindel ND, Gerecke DR, Laskin DL, Laskin JD. Mechanisms mediating the vesicant actions of sulfur mustard after cutaneous exposure. Toxicol Sci 2010; 114: 5-19.

Ghanei M, Harandi AA. Molecular and cellular mechanism of lung injuries due to exposure to sulfur mustard: A review. Inhalation Toxico 2011; 23: 363-371.

Thomas DA, Cortes J, Kantarjian HM. New agents in the treatment of acute lymphocytic leukaemia. Best Pract Res Clin Haenatol 2002; 15: 771-790.

Laskin JD, Black AT, Jan YH, Sinko PJ, Heindel ND, Sunil V, Heck DE, Laskin DL. Oxidants and antioxidants in sulfur mustard-induced injury. Ann N Y Acad Sci 2010; 1203: 92-100.

Bayraktar UD, Bashir Q, Qazilbash M, Champlin RE, Ciurea SO. Fifty years of melphalan use in hematopoietic stem cell transplantation. Biology of Blood and Marrow Transplantation 2013; 19: 344-356.

Cheng-Ong K, Giaever G, Nislow C. DNA-Damaging Agents in Cancer Chemotherapy: Serendipity and chemical biology. Chemistry & biology 2013; 20: 648-659.

Catovsky D, Else M, Richards S. Chlorambucil still not bad: A reappraisal. clinical lymphoma myeloma and leukemia 2011; 11: S2-S6.

Trigg ME, Flanigan-Minnick A. Mechanisms of action of commonly used drugs to treat cancer. Community Oncology 2011; 8: 357-369.

Galmarini D, Galmarini CM, Galmarini FC. Cancer chemotherapy: A critical analysis of its 60 years of history. Critical Reviews in Oncology/Hematology 2012; 84: 181-199.

Baumann F, Preiss R. Cyclophosphamide and related anticancer drugs. Journal of Chromatography B: Biomedical Sciences and Applications 2001; 764: 173-192.

Khan GS, Shah A, Rehman Z, Barker D. Chemistry of DNA minor groove binding agents. Journal of Photochemistry and Photobiology B: Biology 2012; 115: 105-118.

Corrie PG. Cytotoxic chemotherapy: clinical aspects. Medicine 2011; 39: 717-722.

Nelson SM, Ferguson LR, Denny WA. Non-covalent ligand/DNA interactions: Minor groove binding agents. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis.

CHAPTER 5

From Conventional Prodrugs to Prodrugs Designed by Molecular Orbital Methods

Rafik Karaman*

Bioorganic Chemistry Department, Faculty of Pharmacy Al-Quds University, P.O. Box 20002, Jerusalem, Palestine

Abstract: In this chapter we attempt to present a novel prodrug approach which is based on enzyme models that have been advocated to understand the mechanism by which enzymes catalyze biochemical transformations. The tool exploited in the design of novel prodrugs is computational calculations using molecular orbital (MO) and molecular mechanics (MM) methods and correlations between experimental and calculated rate values for some intramolecular processes. In this approach, no enzyme is needed to catalyze the intraconversion of a prodrug to its active parent drug. The conversion rate is solely determined by the factors affecting the rate limiting step in the intramolecular (interconversion) process. Knowledge gained from unraveling the mechanisms of the studied enzyme models (cyclization of Bruice's dicarboxylic semi-esters and acid-catalyzed hydrolysis of Kirby's N-alkylmaleamic acids) was exploited in the design. It is believed that the use of this approach might eliminate all disadvantages related to prodrug interconversion by the metabolic approach (enzyme catalyzed process). By utilizing this approach we have succeeded to design novel prodrugs for a number of commonly used drugs such as the anti-bleeding agent, tranexamic acid, the antihypertensive agent, atenolol, the pain killer agent, paracetamol, and the antibacterial agents, amoxicillin, cephalexin and cefuroxime. *In vitro* studies have shown that in contrast to the active drugs (atenolol, paracetamol, amoxicillin and cephalexin) which possess bitter sensation, the corresponding prodrugs were bitterless. Hence, it is expected that patient compliance especially in the pediatric and geriatric population will be significantly increased.

Keywords: *Ab initio* calculations, DFT calculations, enzyme models, molecular mechanics calculations, prodrugs, prodrugs design.

INTRODUCTION

Drug discovery is considered as a lengthy interdisciplinary endeavor. It is a multi-

***Corresponding author Rafik Karaman:** Department of Bioorganic & Pharmaceutical Chemistry, College of Pharmacy, Al-Quds University, P. O. Box 20002, Jerusalem, Palestine; Tel & Fax: +972-2-2790413; E-mail: dr_karaman@yahoo.com

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

step process that commences with target and lead discovery, followed by lead optimization and *in vitro* and *in vivo* studies to evaluate if a chemical entity complies with a number of pre-set criteria to start clinical development. The number of years it takes to introduce a drug to the pharmaceutical market is estimated between 10-12 years with a cost of more than $0.9 billion dollars [1, 2].

In the past few decades the pharmaceutical industry has been subjected to considerable alterations in terms of improving drug drawbacks that are related to pharmacokinetic (pharmaceutical and biological) performance of existing drugs which may hinder drug development process [3-7]. Overcoming the undesirable physicochemical properties of a number of marketed drugs can be achieved through the development of new chemical entities with desirable efficacy and safety. However, this is an expensive and time consuming process that needs a screening of thousands of molecules for biological activity. Over the past two decades, an increased recognition that the discovery of potent therapeutics involving design of new entities, possess "drug-like" properties and high binding affinity for their biological targets, has been established. The drug-like properties consist of solubility, permeation across barriers and metabolic and excretory clearance [3-7].

An adequate balance of the physicochemical properties enables a drug moiety to attain and maintain the required systemic concentrations for achieving therapeutic effects *via* optimum absorption, distribution, metabolism, and excretion (ADME) processes. A drug moiety that is poorly absorbed, rapidly metabolized or rapidly excreted will not have the ability to attain and provide an efficient therapeutic potential. Such a drug will require much higher doses to achieve sufficiently high systemic concentrations for efficacy, which may not be beneficial in some cases or may cause side effects in others. Thus, good drug-like properties are often defined as physicochemical properties of a drug that enable it to circulate through physical, biochemical, and physiological barriers imposed in the physiological environments. The pharmaceutical properties of a drug candidate are optimized by *de novo* design which involves selections of appropriate physicochemical attributes into the drug moiety or through formulation of the drug candidate with pharmaceutical or biochemical agents that can improve the physicochemical properties. Another important approach that has been used to impart good

pharmaceutical properties is a design of a prodrug moiety that transiently modifies physicochemical properties of a drug to overcome a shortcoming. Hence, it becomes much more feasible to modify and improve the properties of existing drugs through exploring the prodrug approach for eliminating the undesirable properties and to increase the commercial life-cycle and patentability of the concerned drugs. The prodrug approach is a promising and well established strategy for the development of new entities that possess superior efficacy, selectivity and reduced toxicity. Hence an optimized therapeutic outcome can be accomplished using this approach. Approximately, about 10% of all worldwide marketed medicines can be categorized as prodrugs, and in 2008 alone, about 33% of all approved small molecular weight drugs were prodrugs, and this signifies the success of the prodrug approach [5-7].

A complete understanding of the physicochemical and biological behavior of a drug candidate is required when modifying the drug's absorption, distribution, metabolism and elimination (ADME) properties [8-12]. This approach consists of comprehensive evaluation of drug-likeness involving prediction of ADME properties which can be accomplished using *in vitro and in vivo* data obtained from tissue or recombinant material, from humans, and pre-clinical species. In addition, *in silico* or computational predictions of *in vitro* or *in vivo* data which involves an evaluation of various ADME properties, using computational methods such as quantitative structure activity relationship (QSAR) or molecular modeling, are required as well [1-7].

Studies have shown that high attrition rates in the drug development process are attributed to poor pharmacokinetics and toxicity, and researchers have reached to the conclusion that these issues should be heavily considered as early as possible in the drug discovery process in order to improve the efficiency and cost effectiveness of the drug candidate [13].

Therefore, the goal is to design drugs having an efficient dissolution and permeability to be transferred to the blood circulation (absorption) and efficiently reach their target (distribution) and to be sufficient stable to survive the physiological journey (metabolism) and to be eliminated in a reasonable time (elimination).

In order to achieve a drug's success in reaching the biological target the following drug's physicochemical properties are required: (1) chemical stability in aqueous solutions, such as stomach, intestine and blood circulation environments. (2) Metabolic stability; the drug must survive digestive and metabolic enzymes (liver) and any metabolites (product of drug metabolism) should not be toxic or ineffective. Metabolic enzyme activity (cyctochrome P450's) varies from individual to individual can be affected by other chemicals, such as grapefruit juice which inhibits activity; cigarette smoke and brussel sprouts which enhance it; other drugs may inhibit or promote P450 enzymes. For example, antibiotics can act as P450 inhibitors; slows the metabolism of other drugs by these enzymes. Phenobarbitone stimulates the P450 enzymes; accelerating the metabolism of warfarin (anticoagulant) and making it less effective. Cimetidine (antihistamine) inhibits the P450 enzymes; slowing the metabolism of warfarin (anticoagulant). St. John's wort (herbal medicine for mild to moderate depression) promotes P450 enzymes; decreasing the effectiveness of contraceptives and warfarin. Anticoagulants are bound by plasma protein in the blood, but aspirin displaces them, which can lead to a drug overdose, and (3) Successful absorption; diffusion across membrane (solubility and permeability; size, H-bonding). Too hydrophilic drugs can't cross membranes; more easily excreted by kidneys and too hydrophobic drugs have poor water solubility, poorly absorbed from GI tract because they coagulate in fatty globules [3-4].

PRODRUG OVERVIEW

The prodrug term involves chemically modified inert compound which upon administration releases the active parent drug to elicit its therapeutic activity within the body. Since few decades, prodrug strategy has increasingly being developed to overcome undesired drug physicochemical properties. Generally, prodrugs consist of a promoiety that is removed by enzymatic or chemical reactions, while other prodrugs release their active drugs after molecular modification such as an oxidation or reduction reactions. In some cases, two active drugs can be attached together in a single molecule called a codrug. In a codrug, each drug acts as a linker for the other. It is important to ensure that the prodrug should be pharmacologically inactive, rapidly converted to its active drug and a non-toxic moiety [14, 15]. Nearly 55 years ago, Albert introduced the

prodrug term for the first time in his book 'selective toxicity'. The first prodrug was not originally designed as a prodrug, but its nature was determined later. Earlier examples of compounds fulfill the classical criteria of prodrug were acetanilide and phenacetin, which exhibit their activities after being metabolized within the body [16]. Acetanilide is an antipyretic agent that entered the clinical use in 1886. It undergoes metabolism (aromatic hydroxylation) to paracetamol, in a similar manner to that of phenacetin which produces paracetamol *via* O-dealkylation [17].

In the late nineteenth century a chemist, Felix Hoffman at Bayar-Company, synthesized the antipyretic agent Aspirin (acetylsalicylic acid), which was introduced for the first time in clinical practice in 1899; it can be considered a less corrosive prodrug form of salicylic acid to minimize the gastric irritation and ulcerogenicity associated with salicylic acid. However, it remains a matter of debate whether aspirin is a true prodrug or not [18]. Since then many prodrugs were synthesized to overcome many pharmaceutical and pharmacokinetic problems such as, low bioavailability by increasing or decreasing lipophilicity of the parent drug, site selectivity for higher absorption and less toxicity, short duration of action to increase patient compliance, rapid metabolism to increase oral bioavailability and masking bitter sensation of commonly used drugs, which is crucial for geriatric and pediatric patient compliance.

Prodrug design is an efficient approach used to overcome these problems. The lipophilicity of poorly permeable drugs can be increased by linking the drug to a lipophilic linker such that it can be used for oral, ocular or local drug delivery. Prodrugs can be also used to increase aqueous solubility by linking the drug to polar or ionizable groups. In addition, prodrugs use has succeeded to overcome site selectivity problems, which can be achieved by targeting a specific enzyme or receptor, such as targeting an enzyme that is over expressed in tumor cells. Further, mAbs have been used as ligands to transport prodrugs to tumor cells. They are designed as drug-antibody conjugate or antibody enzyme conjugate [19], targeting membrane transporters is utilized in order to increase absorption such as in the case of valacyclovir prodrug. Several prodrugs have been used to prolong duration of action, such as buprenorphine decanoate and fluphenazine decanoate ester prodrugs [20]. Prodrugs of naltrexone, nalbuphine, estradiol and dopamine

are another example of prodrugs to increase bioavailability of drugs that are susceptible to presystemic metabolism, by protecting the groups involved in metabolism [21]. An example for dopamine prodrug to increase bioavailability is shown in Fig. **1**. Prodrugs also are applied to decrease pain at injection site by making drugs more water soluble. Masking bitter taste and improvement of odor are important applications of prodrugs to increase patient compliance. Taste masking is achieved by blocking chemical groups that are involved in the drug interaction with bitter taste receptors.

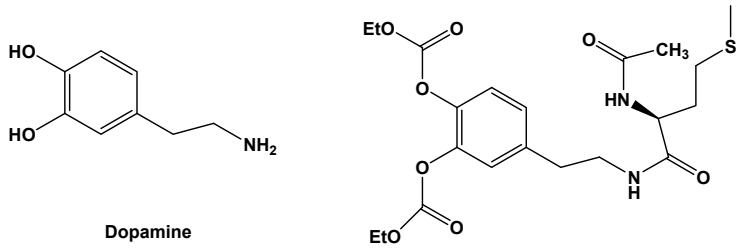


Figure 1: Chemical structures of docarpamine [*N*-(*N*-acetyl-L-methionyl)-*O*, *O*-*bis* (ethoxycarbonyl) dopamine), a pseudopeptide prodrug of dopamine and dopamine.

INTRAMOLECULAR PROCESSES USED FOR THE DESIGN OF POTENTIAL PRODRUGS

Most of the prodrugs that are in clinical use require enzymatic catalysis in order to interconvert into their corresponding parent drugs; they are typically esters of drugs containing carboxyl or hydroxyl groups, which are readily interconverted by esterase catalyzed hydrolysis [22]. However, applying enzymatic activation as mentioned before suffers many disadvantages such as high chemical reactivity that precludes either liquid or solid formulation of the prodrug or low chemical reactivity, resulting in low *in vivo* concentration level of the active drug. Therefore, the development of prodrugs through non-enzymatic pathways has emerged as an alternative approach in which prodrug activation is not affected by inter- and intra-individual variability that has consequences on the enzymatic activity. In particular, cyclization-activated prodrugs have been capturing the attention of medicinal chemists since more than two decades.

Many different strategies have been exploited in recent years for the development of intramoleculary-activated prodrugs using the cyclization pathways that control

the rate of the drug release from its corresponding prodrug. These strategies include: (1) cyclization reactions that release the active drug as the cyclization product, (2) cyclization reactions involving elimination of the parent drug and (3) cyclization reactions which preced by an enzymatic reaction that generates the internal nucleophile (also called two-step activation). A number of reviews addressing these approaches have previously been published by Shan *et al.* [23], Testa & Mayer [24] Wang *et al.* [25], Vinšová and Imramovský [26] and Stella [27]. Another interesting review on anticancer prodrugs selectively activated by elimination and cyclization reactions was published by Papot *et al.* [28] in 2002.

Additionally, a chapter on cyclization-elimination strategies for prodrug activation has recently published [29].

Another example of intramolecular process that has been utilized for prodrug development is the "conformation lock" (trimethyl lock system). Borchardt and co-workers developed two-step activation prodrugs by carriing out covalent attachment of model drugs to the carboxyl group of the hydrocinnamic acid moiety while masking the *o*-hydroxyl substituent as a precursor structure sensitive to either reductases [30-32], esterases [33-35] or phosphatases [36]. Consequently, the *o*-hydroxyl group could be released in a first enzymatically-promoted transformation, after which fast lactonization would lead to drug release (Fig. **2**).

In 1985, Bundgaard and co-workers proposed pilocarpine prodrugs based on pilocarpic acid double esters [37].The latter were shown to work as prodrugs of pilocarpine both *in vitro* and *in vivo* and, in aqueous solution, to undergo a quantitative and apparently specific-base-catalyzed lactonization to pilocarpine. This process was based on an initial ester hydrolysis step that leaves a hydroxyl nucleophile free to attack the benzyl ester moiety, thus promoting the final cyclization-elimination reaction [37].

Molecular modeling studies of a series of pilocarpic acid mono- and diester prodrugs by Konschim *et al*. have been done to gain an understanding of their general physicochemical properties. Molecular structures and conformers have been determined with molecular mechanics and quantum chemical AM1

calculations [38]. However, no rate calculation has been done to calculate (predict) the effect of the structural features on the drug's release.

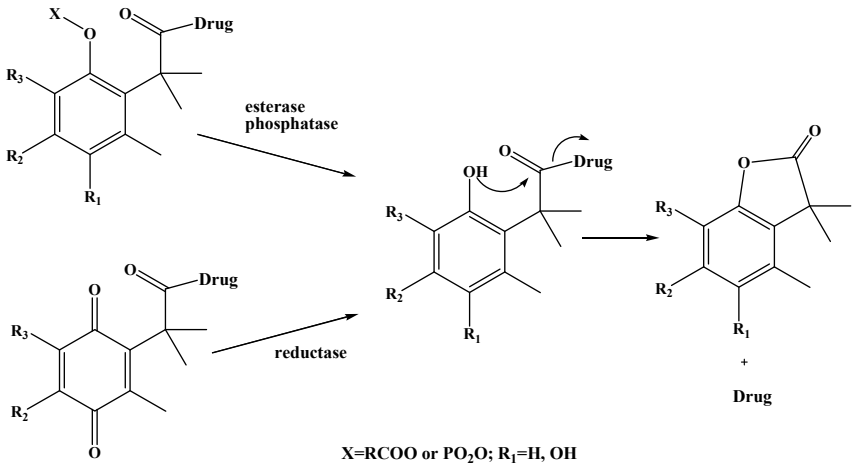


Figure 2: An illustration of the "trimethyl lock" concept used for the design of two-step prodrugs.

In this chapter, a novel prodrug approach, by which the intraconversion of a prodrug to its parent drug is determined only on the structure features of the linker (promoiety), discussed implies prodrugs design based on enzyme models that have been advocated to understand how enzymes work. The tool used in the design is computational approach consisting of calculations using a variety of different molecular orbital such as DFT and *ab initio*, and molecular mechanics methods and correlations between experimental and calculated rate values (activation energy) for some intramolecular processes that were utilized to understand the mechanism/s by which enzymes exert their high catalytic effeciency. In this approach, there is no need to enzyme catalysis of the intra-conversion of a prodrug to its active drug. The release rate of the active drug is solely determined by the factors affecting the rate limiting step of the prodrug intraconversion process. Knowledge gained from the mechanisms of the previously studied enzyme models was used in the design.

It should be worth noting that the use of this approach might eliminate all disadvantages that are associated with prodrug interconversion by enzymes. The bioconversion of prodrugs is perhaps the most vulnerable link in the chain, because there are many intrinsic and extrinsic factors that can affect the process.

For example, the activity of many prodrug activating enzymes may be varied due to genetic polymorphisms, age-related physiological changes, or drug interactions, leading to adverse pharmacokinetic, pharmacodynamic, and clinical effects. In addition, there are wide interspecies variations in both the expression and function of the major enzyme systems activating prodrugs, and these can pose some obstacles in the preclinical optimization phase.

ENZYME MODELS USED IN THE PRODRUG DESIGN

Over the past few decades, pioneer studies by several organic chemists and biochemists, such as Bruice, Bender, Jencks, Menger, Kirby and Walesh, have contributed much for understanding how enzymes catalyze biochemical reactions [39-52].

The consensus today is that the principal of enzyme's catalysis is based on both, the catalysis by functional groups and the ability to reroute intermolecular reactions through alternative pathways by which substrates can bind to preorganized active sites. Moreover, studies have demonstrated that rates enhancement by enzymes can be driven by a number of factors: (1) covalently enforced proximity, such as in the case of chymotrypsin, (2) non-covalently enforced proximity, as seen in the catalysis of metalloenzymes, (3) covalently enforced strain, and (4) non-covalently enforced strain, which has been researched on models that mimic the lysozyme enzyme.

In general, rates of enzymatic reactions are more than 10^{10}-10^{18} -fold the corresponding non-enzymatic bimolecular counterparts. For example, catalyzed reactions by orotidine monophosphate decarboxylase are enhanced by 10^{17}-fold whereas that catalyzed by cyclophilin are accelerated by10^{5}-fold. The huge rates enhancement brought about by enzymes is as a result of the substrate binding within the confines of the enzyme's active site. The binding energy of the resulting enzyme substrate complex is the main driving force to catalysis. It is assumed that in all enzyme catalyzed biotransformation, binding energy is used to overcome physical and thermodynamic factors imposing barriers to the reaction (free energy) such as: (i) the change in entropy, in the form of the freedom of motions of the reactants in solution; (ii) the hydrogen bonding around

biomolecules in aqueous solution; (iii) a proper alignment of catalytic functional groups on the enzyme and (iv) the distortion of a substrate that must occur before the reaction commences [39-53].

The extraordinary high efficiency of enzymes depends on a combination of few factors that have been recognized but none of them was fully understood. Despite the growing progress reached in understanding enzyme catalysis a number of important factors still to be investigated. The high rates of intramolecular reactions are fascinating for chemists because they are reminiscent of the efficiency of enzyme catalysis and it is widely believed that a common source is, at least for a significant part, responsible both effects. The similarity between intramolecularity and enzymes has encouraged a number of chemists and biochemists to design chemical models based on intramolecular reactions involving two reactive centers in order to reveal the mode and mechanism of enzymes catalysis. Over the past 50 years suggestions have been proposed from attempts to interpret changes in reactivity *versus* structural variations in intramolecular systems. Among those proposals: (1) Koshland ‘‘orbital steering’’ which suggests a rapid intramolecularity arises from a severe angular dependence of organic reactions as has been shown in the lactonization of hydroxy acids [44]; (2) ‘‘proximity’’ in intramolecular processes (near attack conformation) model as advocated by Bruice and demonstrated in the lactonization of di-carboxylic acids semi-esters[45-47]; (3) ‘‘stereopopulation control’’ based on the concept of freezing a molecule into a productive rotamer as presented by Cohen [48-50], (4) Menger’s ‘‘spatiotemporal hypothesis’’ which postulates that the rate of reaction between two reactive centers is proportional to the time that the two centers reside within a critical distance [51-55] and (5) Kirby’s proton transfer models on the acid-catalyzed hydrolysis of acetals and N-alkylmaleamic acids which demonstrated the importance of hydrogen bonding formation in the products and transition states leading to them [56-64].

Studies on intramolecularity have played a fundamental role in elucidating the chemistry of the groups involved in enzyme catalysis as well as in unraveling the mechanisms available for particular processes. Thus, it is believed that these studies have the potential to provide an adequate understanding of how efficiency depends on structure in intramolecular catalysis which in turns could shed light on

related problems in enzyme catalysis, and could be potentially used as prodrugs linkers.

COMPUTATIONAL METHODS BACKGROUND

The use of computational chemistry for geometries and energies of ground and transition states calculations has become a daily tool for organic, bioorganic and medicinal chemists alike. Computer science principles are utilized to aid in solving chemical problems by incorporating theoretical chemistry results into efficient computer programs to calculate the geometries and physical and chemical properties of molecules [65].

Kinetics and thermodynamics calculations for biological systems having pharmaceutical and medicinal interests are considered a challenge to the health community. Computational calculations utilizing quantum mechanics (QM) such as *ab initio*, semi-empirical and density functional theory (DFT) methods, and molecular mechanics (MM) methods are increasingly being used and broadly accepted as reliable tools for the prediction of potential drugs and prodrugs alike [65].

The above mentioned computational methods can handle both static and dynamic situations. In all cases the computer time, memory and disk space increase drastically with the studied system's size. *Ab initio* methods generally are useful only for small systems. They are based entirely on theory from first principles. The *ab initio* molecular orbital methods (QM) such as HF, G1, G2, G2MP2, MP2, MP3 and MP4 are based on rigorous use of the Schrödinger equation with a number of approximations. *Ab initio* electronic structure methods have the advantage that they can be made to converge to the exact solution, when all approximations are sufficiently small in magnitude and when the finite set of basis functions tends toward the limit of a complete set. The convergence, however, is usually not monotonic, and sometimes the smallest calculation gives the best result for some properties. The disadvantage of *ab initio* methods is their time-consuming cost [66-67].

Other less accurate methods are the semi-empirical because they make many approximations and obtain some parameters from empirical data. The semi-

empirical quantum chemistry methods are based on the Hartree-Fock formalism and they are very important in computational chemistry for treating large molecules where the full Hartree-Fock method without the approximations is too expensive. Within the framework of Hartree-Fock calculations, some pieces of information (such as two-electron integrals) are sometimes approximated or completely omitted. In order to correct for the loss of two-electron integrals, semi-empirical methods are parameterized, that is their results are fitted by a set of parameters, normally in such a way as to produce results that best agree with experimental data, but sometimes to agree with *ab initio* results. Semi-empirical calculations are much faster than their *ab initio* counterparts. Their results, however, can be very wrong if the molecule being computed is not close enough to the molecules in the database used to parameterize the method. The most used semiempirical methods are MINDO, MNDO, MINDO/3, AM1, PM3 and SAM1 [68-71]. Calculations of molecules exceeding 70 atoms can be done using such methods (the number of atoms to be calculated is dependent on the computer efficiency used).

Another commonly used quantum mechanical method in chemistry and physics to study the electronic structure, especially the ground state of many- systems, in particular atoms, molecules, and the condensed phases is the density functional theory (DFT). With this theory, the properties of many systems can be predicted by using functionals, *i.e.* functions of another function, which in this case is the spatially dependent electron density. Hence the name density functional theory comes from the use of functionals of the electron density. DFT is among the most common methods available in condensed-matter physics, computational physics, and computational chemistry. The DFT method is used to calculate geometries and energies for medium-sized systems (up to 60 atoms depending on the basis set used) of biological and pharmaceutical interest and is not restricted to the second row of the periodic table [72-74].

On the other hand, molecular mechanics is a mathematical approach used for the computation of structures, energy, dipole moment, and other physical properties. It is widely used in calculating many diverse biological and chemical systems such as proteins, large crystal structures, and relatively large solvated systems. However, this method is limited by the determination of parameters such as the large number of unique torsion angles present in structurally diverse molecules [75].

Quantum mechanics methods are important tool to investigate functional mechanisms of biological macromolecules based on their 3D and electronic structures. The system size which *ab initio* calculations can handle is relatively small despite the large sizes of biomacromolecules surrounding solvent water molecules. Accordingly, isolated models of areas of proteins such as active sites have been studied in *ab initio* calculations. However, the disregarded proteins and solvent surrounding the catalytic centers have also been shown to contribute to the regulation of electronic structures and geometries of the regions of interest.

To overcome these discrepancies combined quantum-mechanics/molecular-mechanics (QM/MM) approaches have become the method of choice for modeling reactions in biomolecular systems. Quantum-mechanical (QM) methods are required for calculating chemical and other electronic processes, such as charge transfer or electronic excitation. However, QM methods are restricted to systems of up to a limited number of atoms. However, the size and conformational complexity of biopolymers needs methods capable of treating up to several 100,000 atoms and allowing for simulations over time scales of tens of nanoseconds. This is achieved by highly efficient, force-field-based molecular mechanics (MM) methods. Thus to model large biomolecules the ideal approach to be used is a combination of the two methods and to use a QM method for the chemically active site and MM treatment for the surroundings. The resulting techniques are called combined or hybrid QM/MM methods. They enable the modeling of reactive biomolecular systems at a reasonable computational time while providing a reasonable accuracy. The pioneer work of the QM/MM method was accomplished by Warshel and Levitt, and since then, there has been much progress on the development of a QM/MM algorithm and applications to biological systems [76-78].

CALCULATION METHODS USED FOR EXPLORING THE ENZYME MODELS MECHANISMS AND FOR PRODRUGS DESIGN

The MM2 molecular mechanics strain energy calculations were done using Allinger's MM2 program [75]. The Becke three-parameter, hybrid functional combined with the Lee, Yang, and Parr correlation functional, denoted B3LYP, were employed in the calculations using density functional theory (DFT). All

calculations were carried out using the quantum chemical package Gaussian 03 and Gaussian 09 [79]. Calculations were carried out based on the restricted Hartree-Fock method. The starting geometries of all calculated molecules were obtained using the Argus Lab program [80] and were initially optimized at the HF/6-31G level of theory, followed by optimization at the B3LYP/6-31G(d,p) and B3LYP/6-311+G(d,p) levels. Total geometry optimizations included all internal rotations. Second derivatives were estimated for all 3N-6 geometrical parameters during optimization. An energy minimum (a stable compound or a reactive intermediate) has no negative vibrational force constant. A transition state is a saddle point which has only one negative vibrational force constant [81]. The "reaction coordinate method" [82] was used to calculate the activation energy in the studied processes. Transition states were located first by the normal reaction coordinate method [82] where the enthalpy change was monitored by stepwise changing the interatomic distance between two specific atoms (0.1 Å). The geometry at the highest point on the energy profile was re-optimized by using the energy gradient method at the B3LYP/6-31G(d,p) or B3LYP/6-311+G(d,p) level of theory. The activation energy values were calculated from the difference in energies of the global minimum structures (GM) and the derived transition states. The transition state structures were verified by their only one negative frequency. The activation energies obtained from the DFT for all molecules were calculated with and without the inclusion of water. The calculations with the incorporation of water were performed using the integral equation formalism model of the Polarizable Continuum Model (PCM) [83-86]. In this model the cavity is created *via* a series of overlapping spheres. The radii type employed was the United Atom Topological Model on radii optimized for the PBE0/6-31G(d) level of theory. In this chapter, the mechanisms of some enzyme models that have been advocated to understand enzyme catalysis were computationally investigated. The tool used in the study is computational approach consisting of calculations using a variety of different molecular orbital and molecular mechanics methods and correlations between experimental and calculated reactions rates [87-105].

BRUICE'S ENZYME MODEL BASED ON RING-CLOSING OF DI-CARBOXYLIC SEMI-ESTERS [95]

Bruice and Pandit have studied the hydrolysis of di-carboxylic semi-esters **1-6** shown in Fig. **3** and found that the relative rate (k_{rel}) **6**>**5**>**4**>**3**>**2**>**1**. They

attributed the enhancement in rates to proximity orientation. Using the observation that alkyl substitution on succinic acid influences rotamer distributions, the ratio between the reactive gauche and the unreactive anti-conformations, they proposed that *gem*-dialkyl substitution increases the probability of the resultant rotamer adopting the more reactive conformation. Therefore, for cyclization to occur, the two reacting centers must be in the gauche conformation. In the unsubstituted reactant, the reactive centers are almost completely in the anti-conformation for minimizing steric interactions [45-47].

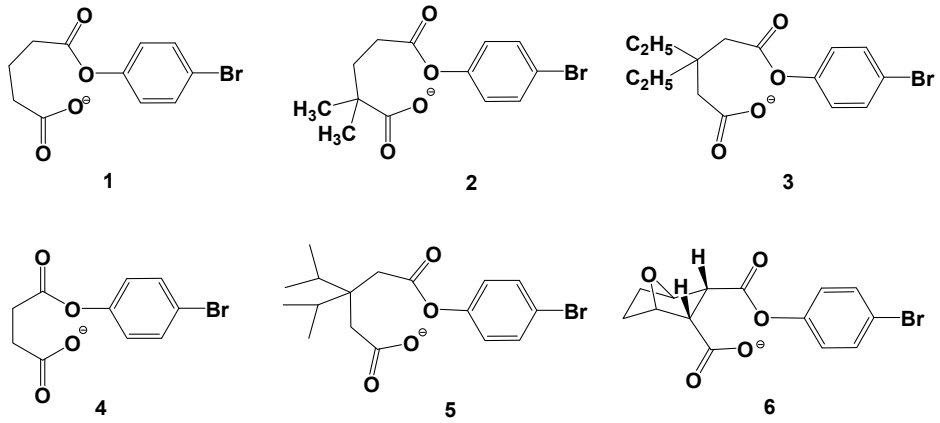


Figure 3: Chemical structures for di-carboxylic semi-esters 1-6.

Menger and Bruice ascribed the phenomenon of rate accelerations in some intramolecular processes to the importance of the proximity of the nucleophile to the electrophile of the ground state molecules [45-47, 51-55]. Menger in his "spatiotemporal" hypothesis developed an equation relating activation energy to distance and based on this equation, he concluded that enormous rate enhancements in reactions catalyzed by enzymes are achieved by imposing short distances between the reactive centers of the substrate and enzyme [51-55]. On the other hand, Bruice attributed the catalysis by enzymes to favorable 'near attack conformations'. According to Bruice's hypothesis, systems that have a high quota of near attack conformers will have a higher reaction rate and *vice versa*. [45-47]. In contrast to the proximity proposal, others believe that high rate acceleration in intramolecular processes is due to relief of the reactants strain energy [106]. To examine whether the discrepancy in ring-closing rates of di-carboxylic semi-esters **1-6** is due to proximity orientation or to strain effects, we

have calculated, using *DFT* method at B3LYP/6-31G(d,p) and HF/6-31G(d,p) levels, the ground state, intermediate and transition state structures as well as the activation energy values for processes **1**-**6**. The DFT optimized geometries for the global minimum (GM) and transition state structures (TS) for systems **1**-**4** and **6** are shown in Figs. **4** and **5**, respectively. In accordance with Bruice's results [45-47] we have found that the cyclization reaction proceeds by one mechanism, by which the rate-limiting step is the tetrahedral intermediate collapse and not its formation (Fig. **6**). However, contrarily to Bruice's conclusion we have found that the acceleration in rate is due to strain effect rather than to proximity orientation stemming from the "rotamer effect" (for further information, see Table **1**.

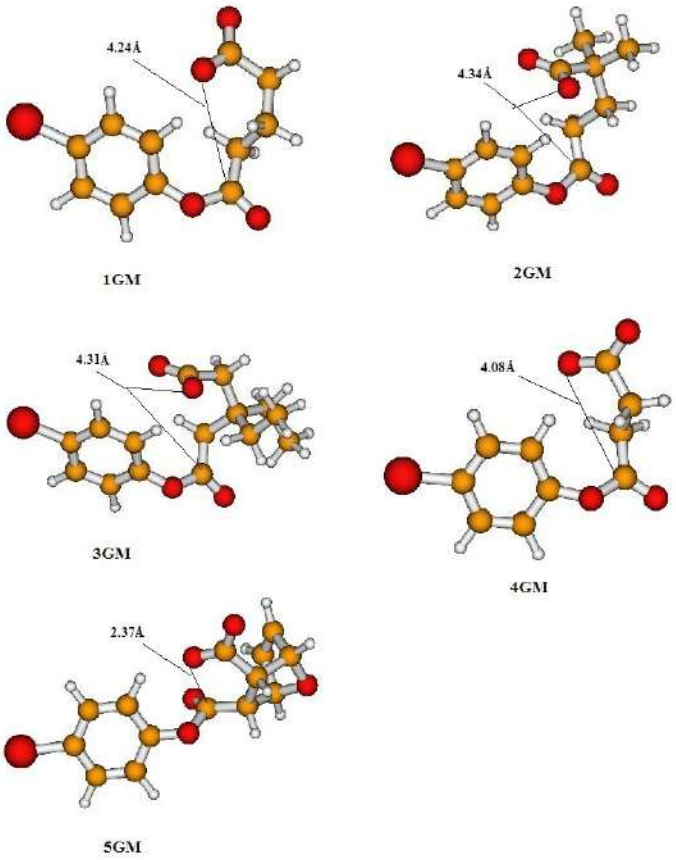


Figure 4: DFT at B3LYP/6-31G(d,p) level optimized structures for the global minimum (GM) in di-carboxylic semi-esters 1-4 and 6.

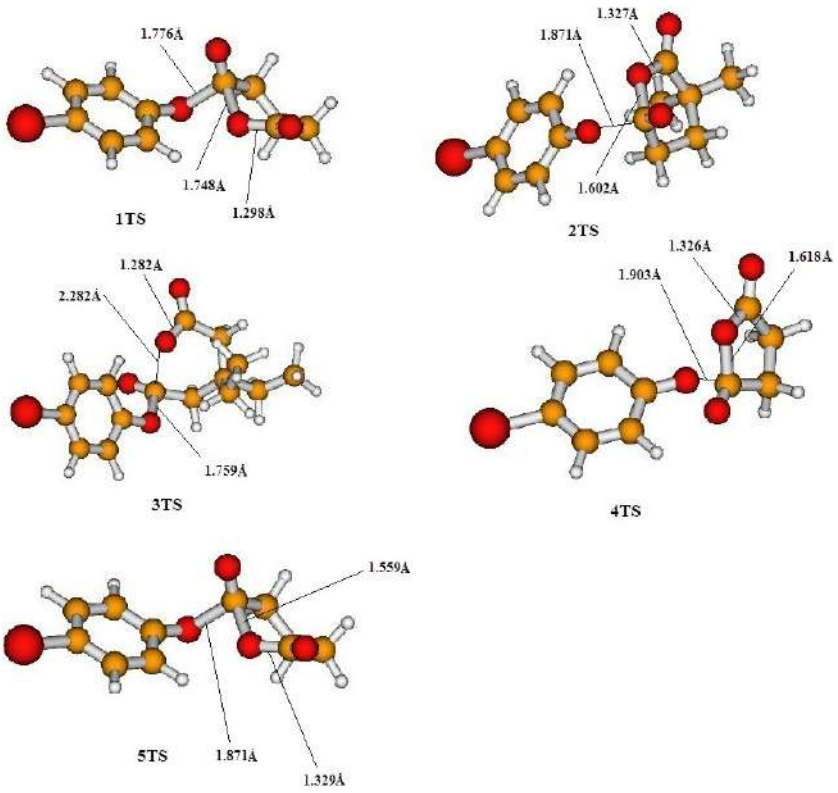


Figure 5: DFT at B3LYP/6-31G(d,p) level optimized structures for the transition state (TS) in di-carboxylic semi-esters 1-4 and 6.

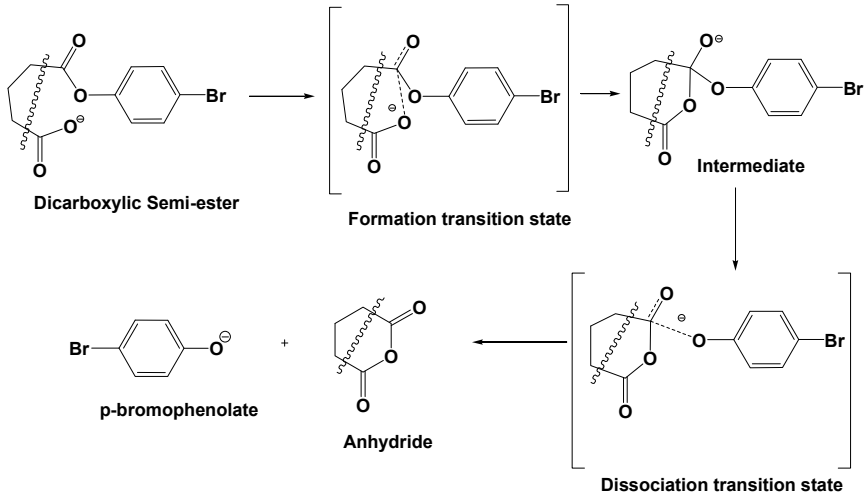


Figure 6: Proposed mechanism for the cyclization of di-carboxylic semi-esters 1-6.

To test whether the discrepancy in the reactions rates of 1-6 (Fig. **3**) stems from proximity orientation or due to steric effects, Allinger's MM2 strain energy values [75] for the reactants and intermediates in systems 1-4 and 6 were calculated and their values were correlated with the experimental relative rate (log k_{rel}) values [45-47] (see Table **1**). The correlation results demonstrated strong correlation between the two parameters. Attempts to correlate the distance between the two reacting centers (r_{GM}) and log k_{rel} failed to give any linearity between the two parameters. This suggests that the driving force for the acceleration in the ring-closing processes is driven by strain energy and not by Bruice's "near attack rotamer" [45-47]. Further support to this conclusion was obtained by a strong correlation found between the activation energy values ($\Delta G^{\ddagger}_{H2O}$and $\Delta G^{\ddagger}_{GP}$) for 1-4 and 6 with both log k_{rel} and the MM2 strain energy values, ΔEs (TS - AN).

Table 1: DFT calculated properties for the cyclization reactions of 1-4 and 6

System	log k_{rel} [45-47] (exp.)	ES_{INT-GM} (MM2 calc.)	$\Delta G^{\ddagger}$ (GP) B3L	$\Delta G^{\ddagger}$ (H_2O) B3L	r_{GM} B3L	$\Delta G^{\ddagger}$ (GP) B3L311	$\Delta G^{\ddagger}$ (H_2O) B3L311
1	3.00	8.70	19.09	29.37	4.24	9.26	20.33
2	3.30	9.30	12.22	21.10	4.34	13.13	22.03
3	5.26	8.07	12.83	16.13	4.31	10.27	13.98
4	5.36	4.24	1.43	9.03	4.08	2.76	12.54
6	7.90	2.31	10.48	16.51	2.37	------	------

log k_{rel} is the experimental relative rate [45-47]. $\Delta G^{\ddagger}$ is the activation free energy (kcal/mol). r_{GM} is the distance between the nucleophile (O1) and the electrophile (C6) in the reactant. B3L and B3L311 refer to calculated by B3LYP/6-31G(d,p) and B3LYP/6-311+G(d,p), respectively. GP and H_2O refer to calculated in the gas phase and water, respectively.

The salient points emerged from this computational study are (1) the activation energy in **1-6** is dependent on the difference in the strain energies of the transition states and the reactants, and there is no relationship between the cyclization rate and the distance between the nucleophile (O-) and the electrophile (C). (2) The observation of opening the cyclic ring during the reaction rate limiting step supports the notion that the difference in the strain energies of the reactant and the transition states plays a crucial role in the discrepancy in the rates of cyclization of the di-carboxylic semi-esters **1-6**. (3) Strained reactants such as **6** are more reactive than the less strained reactants, and the reactivity extent is linearly correlated with the strain energy difference between the transition state and the reactant (ΔEs). (4) The energy needed to provide a stable transition state for a

strained system is less than that for the unstrained system, since the conformational change from the reactant to the transition state in the former is smaller [105].

COMPUTATIONALYY DESIGNED PRODRUGS BASED ON BRUICE'S ENZYME MODEL - THE ANTIMALARIAL AGENT ATOVAQUONE (ATQ)

Malaria-like febrile illnesses have been described since Hippocrates as fevers that were periodic and associated with marshes and swamps. The word "malaria" comes from the Italian "mal'aria" for "bad airs." It was not until 1880s that scientists were able to identify the malaria parasite and link the transmission of malaria to mosquitoes. Malaria is transmitted to humans *via* the bite of infected female mosquito of anopheles species [107]. Malaria can exist, in a mild form that most commonly associated with flu-like symptoms; fever, vomiting, and general malaise. While in the sever form caused by *P. falciparum*, a nervous, respiratory and renal complications frequently coexist due to serious organ failure [108]. In severe cases it can progress to coma or death. Malaria commences with a bite from an infected female *Anopheles* mosquito, which brings the protists *via* saliva into the circulatory system. In the blood circulation, the protists are transferred to the liver to be matured and reproduced.

The disease is a global public health problem, affecting 40% of the population and causes about 2 million deaths per year [109]. Most of disease cases are found in the poorest countries; tropical Africa, Latin America, Southern Asia and Oceania [110]. A more concern is being given now to malaria even in countries where there is a low risk of infection; this is due to the phenomena of global warming which is significantly increasing [111]. World Health Organization (WHO) assesses that 81% of cases and 91% of deaths are found in African regions. Children under 5-years old and pregnant women are the most severely affected. This protozoan disease is caused by 5 parasites species of the genus *Plasmodium* that affect humans (*P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi*) [107]. The only one among these parasites that can cause life threatening complications is *P. falciparum* [110], which is dominated in Africa and to which most drug-resistant cases are attributed.

Several medications, alone or in combination such as chloroquine, antifolates, artemisinins and others show effectiveness and were considered as being the corner stone in malaria treatment. However, drug or multi-drug resistance to these agents has been escalated and constitutes a major challenge in malaria treatment [108]. Accordingly, the need for new antimalarial drugs is now widely recognized, particularly those that are structurally different from existing antimalarial drugs and possess a novel mechanism of action. Atovaquone, a hydroxynaphthoquinone, is relatively new treatment option, active against *Plasmodium spp*. It has a novel mechanism of action, acts by inhibition of the electron transport system at the level of cytochrome bc1 complex. In malaria parasites, the mitochondria act as a sink for the electrons generated from dihydrorotate dehydrogenase; an essential enzyme for pyrimidine biosynthesis; Inhibition of electron transport by ATQ leads to dihydrorotate dehydrogenase inhibition resulting in reduced pyrimidine biosynthesis and thus parasite replication inhibition. It is well established that atovaquone has an excellent safety profile and long half-life, besides it can be administered *via* oral route. However, atovaquone has poor oral bioavailability (less than 10% under fasted condition) and variable oral absorption due to its poor solubility that results from its lipophilic structure. Consequently, this results in low and variable plasma and intracellular levels of the drug which is an important determinant of therapeutic outcome [112-115].

The prodrug approach has the potential to be the most successful among other approaches to overcome this shortcoming. Continuing our study on design and synthesis of atovaquone prodrugs [112], the study herein was to design atovaquone prodrugs through linking atovaquone to a di-carboxylic semiester linker (Bruice's enzyme model) to produce a system that is more hydrophilic than its parent drug and is able to release the active drug in a chemically driven controlled manner without any activation by enzyme. Thus, introducing novel atovaquone prodrugs that fulfill the following requirements: (1) enhanced water solubility; (2) improved oral bioavailability; (3) controlled release rate; (4) predicted plasma levels and (5) improved antiparasitic activity. Based on our calculations that enabled us to unravel the mechanism for the ring-closing reaction of Bruice's dicarboxylic semiesters [45-47] and to assign the factors determining

the reaction rate we have designed five different atovaquone prodrugs with the potential to have better water solubility than their parent drug and to release the active parent drug in a controlled manner (Fig. **7**).

Calculation of the Effective Molarity (EM) and Relative Rate, Log k_{rel}, for Atovaquone Prodrugs, ATQ ProD 1-ProD 5

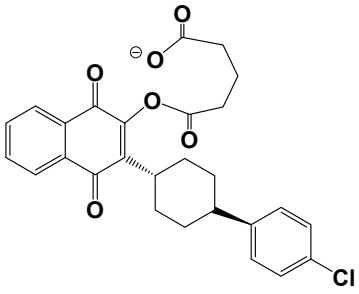
The measure generally used for intramolecular efficiency is the effective molarity (EM). The EM parameter is defined as k_{intra}/k_{inter} for corresponding intramolecular and intermolecular processes driven by identical mechanisms. The factors determining the EM value are ring size, solvent and reaction type. Cyclization reactions *via* intramolecular nucleophilic addition are much more efficient than intramolecular proton transfers. Values in the order of 10^9-10^{13} M have been measured for the effective molarity in intramolecular processes occurring through nucleophilic addition. Whereas for proton transfer processes values of less than 10 M were obtained. The effective molarity parameter is considered an excellent tool to describe the efficiency of a certain intramolecular Process [56-64, 116].

The experimental relative rates for the intramolecular cyclization of **1-6** (Fig. **3**) were obtained from the division of the intramolecular rate and the corresponding intermolecular reaction. For obtaining the relative rates (effective molarity, EM) for processes ATQ **ProD 1- ProD 5** we assume that their corresponding intermolecular process is similar to that for systems **1-6**.

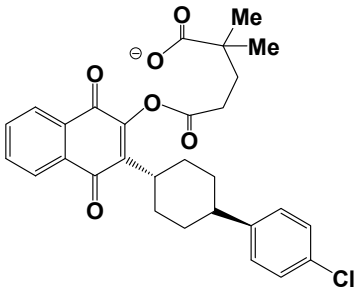
Since an excellent correlation was obtained between the activation free energy values ($\Delta G^{\ddagger}$) for **1-6** and ATQ **ProD 1- ProD 5** and the difference in the strain energy values of the reactants and intermediates, $\Delta Es_{(INT-GM)}$, the calculated values of $\Delta Es_{(INT-GM)}$ for ATQ **ProD 1-ProD 5** were used to calculate their corresponding relative rates (EM and log k_{rel}). The calculated EM values for ATQ **ProD 1-ProD 5** were 6.96, 6.47, 3.78, 6.50 and -12.82, respectively. These values demonstrate that ATQ **ProD 1** and ATQ **ProD 4** are the most efficient processes among all systems studied and the least efficient are ATQ **ProD 3** and ATQ **ProD 5**.

Using the experimental $t_{1/2}$ (the time needed for the conversion of 50% of the reactants to products) value for the cyclization reaction of di-carboxylic semiester **1** and the calculated log k_{rel} values for prodrugs ATQ **ProD 1-ProD 5** we have

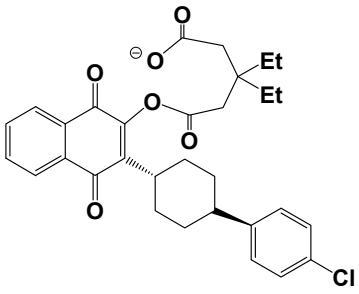
calculated the $t_{1/2}$ values for the conversion of ATQ **ProD 1**- **ProD 5** to their parent drug. The calculated $t_{1/2}$ values were: ATQ **ProD 3**, 22.44 hours; ATQ **ProD 1**, ATQ **ProD 2** and ATQ **ProD 4**, few seconds and ATQ **ProD 5** few years. Therefore, the intraconversion rates of atovaquone prodrugs to atovaquone can be programmed according to the nature of the prodrug linker.



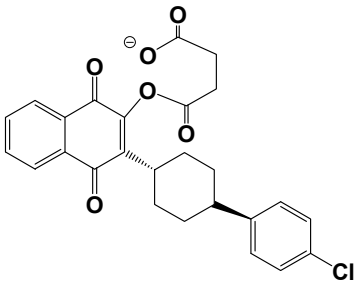
ATQ ProD 1



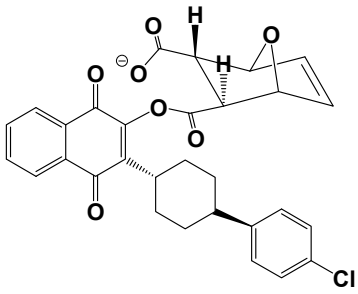
ATQ ProD 2



ATQ ProD 3



ATQ ProD 4



ATQ ProD 5

Figure 7: Chemical structures for atovaquone prodrugs ATQ ProD 1- ProD 5.

BITTERLESS PARACETAMOL PRODRUGS BASED ON BRUICE'S ENZYME MODEL

The palatability of active drugs is a significant obstacle in developing a patient convenient dosage form. Organoleptic properties, such as taste, are an important factor when selecting a drug from the generic products available in the market having the same active ingredient. It is a key issue for doctors and pharmacists to assure an adequate drug's taste upon administration of drugs particularly to the pediatric and geriatric populations [117].

Organic and inorganic molecules dissolve in saliva and bind to taste receptors on the tongue to give a bitter, sweet, salty, sour, or umami sensation. Bitter taste is sensed by the receptors on the posterior part of the tongue. The sensation is a result of signal transduction from taste receptors located in areas known as taste buds. The taste buds contain very sensitive nerve endings, which are responsible for the production and transmission of electrical impulses *via* cranial nerves VII, IX, and X to certain areas in the brain that are devoted to the perception of taste [118]. Molecules with bitter taste [119-123] are very diverse in their chemical structure and physicochemical properties [124, 125]. In humans, bitter taste perception is mediated by 25 G-protein coupled receptors of the hTAS2R gene family. Drugs such as macrolide antibiotics, non-steroidal anti-inflammatory and penicillin derivatives have a pronounced bitter taste [126]. Masking the taste of water soluble bitter drugs, especially those given in high doses, is difficult to achieve by using sweeteners alone. Consequently, several approaches have been studied and have resulted in the development of more efficient techniques for masking the bitter taste of molecules. There are various techniques available which are commonly used for masking drug's bitterness: (1) taste masking with flavors, sweeteners, and amino acids [127]; (2) taste masking with lipophilic vehicles such as lipids, lecithin, and lecithin- like substances[128]; (3) coating which is classified based on the type of coating material, coating solvent system, and the number of coating layers [129]; (4) microencapsulation based on the principle of solvent extraction or evaporation [130]; (5) sweeteners are generally used in combination with other taste masking technologies [131]; (6) taste suppressants and potentiators, such as Linguagen's bitter blockers (*e.g.* adenosine monophosphate), are used for masking the bitter taste of various compounds by competing with binding to the G-protein coupled receptor

sites (GPCR) [132]; (7) resins are utilized to mask pharmaceuticals bitterness by forming insoluble resonates [133, 134]; (8) inclusion complex by which the drug molecule fits into the cavity of a complexing agent and forms a stable complex that masks the drug bitterness by decreasing its oral solubility [135]; (9) pH modifiers are capable of generating a specific pH microenvironment in aqueous media that has the ability to facilitate *in situ* precipitation of the bitter drug compound in saliva thus reducing the overall taste sensation for liquid dosage forms [136]; (10) adsorbates; the compound may be adsorbed or entrapped in the matrix of the adsorbate pore, which may result in a delayed release of the bitter tastant during passage through the oral cavity and thus masking the taste [137] and (11) the prodrug approach by which a functional group/s binds to the bitter taste receptor is blocked by a promoiety. All of the developed techniques are based on the physical modification of the formulation containing the bitter tastant. Although these approaches have helped to improve the taste of some drugs formulations, the problem of drug bitterness in pediatric and geriatric formulations still creates a serious challenge to the health community. Thus, different strategies should be developed in order to overcome this serious problem. Bitter tastant molecules interact with taste receptors on the tongue to give bitter sensation. Altering the ability of the drug to interact with its bitter taste receptors could reduce or eliminate its bitterness. This could be achieved by an appropriate modification of the structure and the size of the bitter compound. Bitter molecules bind to the G-protein coupled receptor-type T2R on the apical membrane of the taste receptor cells located in the taste buds. In humans, about 25 different T2R's are described. Additionally, several alleles are known and about 1000 different bitter phenotypes exist in human beings [120-126]. Due to the large variation of structural features of bitter taste molecules; it is difficult to generalize the molecular requirements for bitterness. Nevertheless, it was reported that a bitter tastant molecule requires a polar group and a hydrophobic moiety. A QSAR model was developed and has been established for the prediction of bitterness of several tastant analogues. For example, it was reported that the addition of a pyridinium moiety to an amino acid chain of a variety of bitter amino acid compounds decreases bitterness, such as in the case of glycine. Other structural modifications, such as an increase in the number of amino groups/residues to more than 3 and a reduction in the polyhydroxyl group/ COOH, have been proven to decrease bitter sensation. Moreover, changing the configuration of a bitter tastant molecule by making isomer

analogues was found to be important for binding affinity to enhance bitterness agonist activity (*e.g.* L-tryptophan is bitter while D-tryptophan is sweet) [138].

Paracetamol is an odorless, bitter crystalline compound used as an over the counter analgesic and antipyretic drug. Paracetamol is used to relief minor aches. It is used as pain killer by decreasing the synthesis of prostaglandin due to inhibition of cyclooxygenases (COX-1 and COX-2) [139]. Paracetamol is favored over aspirin as pain killer in patients who have excessive gastric secretion or prolonged bleeding [139]. It was approved to be used as fever reducer in all ages. Pharmacokinetic studies have shown that urine of patients who had taken phenacetin contained paracetamol. Later was demonstrated that paracetamol was a urinary metabolite of acetanilide [139]. Phenacetin is known historically to be one of the first non-opioid analgesics lacking or has a very slight bitter taste [139]. Comparison of the structures of paracetamol and phenacetin shows close similarity between both analgesics except of the nature of the group on the *para* position of the benzene ring. While in paracetamol the group is hydroxyl, in phenacetin it is ethoxy. On the other hand, acetanilide has a chemical structure similar to that of paracetamol and phenacetin but it lacks any group at the *para* position of the benzene ring. Acetanilide lacks the bitter taste characteristic for paracetamol [139]. The comparisons of the three compounds suggest that the presence of hydroxy group on the *para* position of the benzene ring plays a major role in the bitter sensation resulted from administering paracetamol. Therefore, it is expected that masking the hydroxyl group in paracetamol with a suitable linker might inhibit the binding of paracetamol to its bitter taste receptor/s and hence masking its bitterness. It is likely that paracetamol binds to the active site of its bitter taste receptor *via* hydrogen bonding interactions by which its phenolic hydroxyl group is engaged. It is worth noting, that linking paracetamol with Bruice's enzyme model linker *via* its phenolic hydroxyl group might hinder paracetamol bitter taste.

Based on the DFT calculations on the cyclization of Bruice's **1**-**6** (Fig. **3**), two paracetamol prodrugs were proposed (Fig. **8**). As shown in Fig. **8**, the paracetamol prodrugs, **ProD 1-2**, have a carboxylic acid group as a hydrophilic moiety and the rest of the prodrug, as a lipophilic moiety, where the combination of both groups

provides a moderate HLB. It should be noted that the HLB value is determined upon the pH of the physiologic environment by which the prodrug is dissolved. For example, in the stomach, the paracetamol prodrugs will primarily exist in the carboxylic acid form whereas in the blood circulation the carboxylate form will be dominant. Since Bruice's cyclization reaction occurs in basic medium paracetamol **ProD 1-2** were obtained as carboxylic free acid form, since this form is expected to be stable in acidic medium such as the stomach.

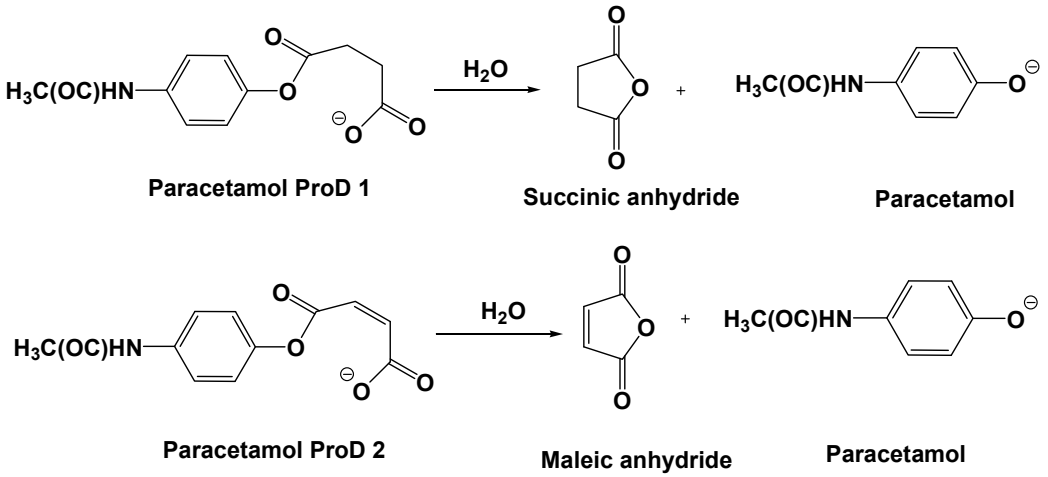


Figure 8: Hydrolysis of paracetamol ProD 1 and paracetamol ProD 2.

IN VITRO INTRA-CONVERSION OF PARACETAMOL PROD 1-PROD 2 TO THEIR ACTIVE DRUG, PARACETAMOL

The hydrolysis of paracetamol **ProD 1-ProD 2** was studied in four different media; 1N HCl and buffers pH 3, pH 6.6 and pH 7.4. The prodrug hydrolysis was monitored using HPLC analysis. At constant pH and temperature the release of paracetamol from its prodrug was followed and showed a first order kinetics. k_{obs} (h^{-1}) and $t_{1/2}$ values for the intraconversion of paracetamol **ProD 1-ProD 2** was calculated from regression equation obtained from the plot of log concentration of residual of paracetamol **ProD 1** *vs*. time. The kinetics results in the different media are summarized in Table **2** and Fig. **9**.

Table 2: The observed *k* value and $t_{1/2}$ of paracetamol ProD 1-ProD 2 In 1N HCl and buffers pH 3 and 7.4

Medium	ProD1 k_{obs} (h^{-1})	ProD2 k_{obs} (h^{-1})	ProD1 $t_{½}$ (h)	ProD2 $t_{½}$ (h)
1N HCl	No reaction	No reaction	No reaction	No reaction
Buffer pH 3	6.3×10^{-5}	-------	3	Very fast
Buffer pH 7.4	6.1×10^{-4}	----------	0.3	Very fast

As shown in Table **2** the hydrolysis rate of paracetamol **ProD 1** at pH 7.4 was the fastest among all media, followed by pH 3 medium. In 1N HCl no conversion of the prodrug to the parent dug was observed. The discrepancy in the behavior between the two prodrugs is due to the fact that the strain energy of maleic anhydride is higher than that of succinic anhydride. It should be emphasized that the reaction rate in these processes is determined on the strain energy of the system.

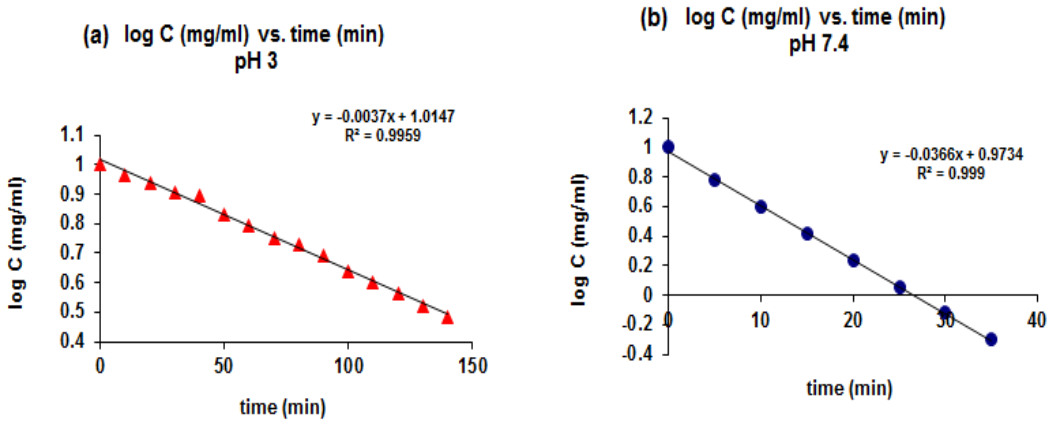


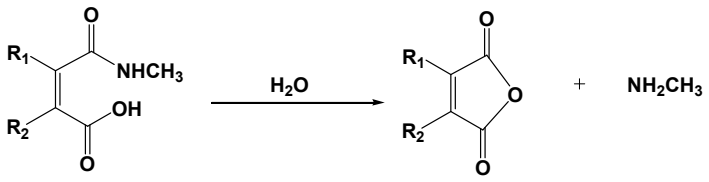
Figure 9: First order hydrolysis plot of paracetamol ProD 1 in (a) buffer pH 3and (b) buffer pH 7.4.

At pH 7.4, paracetamol **ProD 1-ProD 2** are mainly exist as the carboxylate anion form which is expected to undergo fast hydrolysis according to Bruice's mechanism shown in Fig. **3**. At pH 3, the prodrug exists in both form, the carboxylate anion and the carboxylic free acid forms since the pKa of the prodrug is about 3. In 1N HCl, the prodrug is entirely exists as the carboxylic free acid form and since only the carboxylate anion form undergoes Bruice's cyclization. The hydrolysis rate in 1N HCl is almost negligible or zero.

KIRBY'S ENZYME MODEL BASED ON THE ACID-CATALYZED HYDROLYSIS OF N-ALKYLMALEAMIC ACIDS [58, 93, 140-145].

Proton transfer reactions which are typically acid- or base-catalyzed reactions are the most common reactions that enzymes catalyze. Examples of biotransformation processes for such catalysis are the proton transfers catalyzed by triose phosphate isomerase (k_{cat} = 53,000 s^{-1}) and Δ5-3-ketosteroid isomerase (k_{cat} = 8300 s^{-1}) which involve weakly basic and acid groups to achieve such magnificent rates. The fact that reactions of substrate bound in an enzyme active site are between functional groups held in a close proximity encouraged scientists to utilize intramolecularity in modeling the extremely high efficiency of enzymes. Both, enzymes and intramolecularity are similar in that the reacting centers are held together, non-covalently with the former, and covalently with the latter. The significant high efficiency of enzymes catalysis depends on a combination of some factors that most of them have been recognized but none of them was fully understood. Although the devoted research to the chemistry of enzyme catalysis is growing rapidly a number of crucial factors remain to be investigated [151-164]. Kirby *et al.* have researched the mechanism of the acid catalyzed hydrolysis of N-alkylmaleamic acids **7-15** to their corresponding maleamic acids and amines (Fig. **10**). The study found that the reaction is remarkably sensitive to the pattern of substitution on the carbon-carbon double bond. In addition, it revealed that the rates of hydrolysis of the studied dialkyl-N-methylmaleamic acids range over more than ten powers of ten, and the "effective concentration" of the carboxy-group of the most reactive amide, dimethyl-N-n-propylmaleamic acid, is greater than 10^{10} M. This acid amide was found to be converted into the more stable dimethylmaleic anhydride with a half-life of less than one second at 39 ˚C below pH 3 [58]. In addition, Kirby's study demonstrated that the amide bond cleavage is due to intramolecular nucleophilic catalysis by the adjacent carboxylic acid group. Furthermore, based on the fact that the tetrahedral intermediate isomaleimide was converted quantitatively into N-methylmaleamic acid (Fig. **11**), Kirby suggested that the rate-limiting step is the dissociation of the tetrahedral intermediate [58]. Later on Kluger and Chin researched the intramolecular hydrolysis mechanism of a series of N-alkylmaleamic acids derived from aliphatic amines having a wide range of basicity [146]. Their study revealed that the

identity of the rate-limiting step is a function of both the basicity of the leaving group and the acidity of the solution. In 1990, Katagi has computationally studied the reaction mechanism using AM1 semiempirical method and based on his study he concluded that the rate- limiting step is the formation of the tetrahedral intermediate and not its dissociation [147]. For further exploring the factor playing dominant role in proton transfer processes we have computationally studied Kirby's intramolecular acid catalyzed hydrolysis of (4-amino- 4-oxo-2-butenoic) acids (N- alkylmaleamic acids) **7**-**15**.



7; $R_1=R_2=H$
8; $R_1=R_2=Me$
9 ; $R_1=H$; $R_2=Me$
10; R_1,R_2 Cyclopent-l-ene-1,2-diyl
11; R_1, R_2 Cyclohex-l-ene-1,2-diyl
12; $R_1=H$; $R_2=Et$
13; $R_1=H$; R_2=n-Propyl
14; $R_1=H$; R_2=Trifluoromethyl
15; $R_1=R_2$= Trifluoromethyl

Figure 10: Chemical structures for N-alkylmalic acids 7-15.

The aims of our computational work were to: (i) investigate whether the rate-limiting step in **7**-**15** is the formation or the collapse of the tetrahedral intermediate, and to unravel the nature of the driving force(s) responsible for the extremely high rates determined for the acid catalyzed hydrolysis of **8** and **11**, (ii) assign the structural factors associated with high reactivity in the hydrolysis, in the expectation that similar factor will be operative in enzyme catalysis.

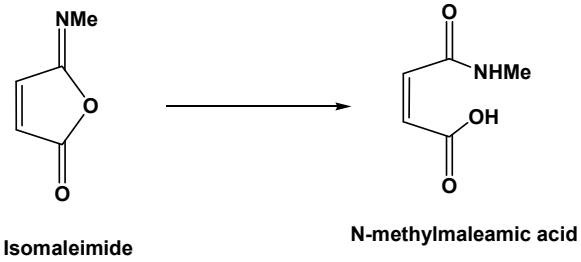


Figure 11: Conversion of isomaleimide to N-methylmaleamic acid.

Using DFT calculation methods at B3LYP/6-31G(d,p), B3LYP/6-311+G(d,p) levels and hybrid GGA (MPW1k) we have computed the acid catalyzed hydrolysis of maleamic (4-amino-4-oxo-2-butenoic) acids (Kirby's N-alkylmaleamic acids) **7**-**15** (Fig. **10**) and the calculation results confirmed that the reaction proceeds in three steps: (a) a proton transfer from the carboxylic group to the adjacent amide carbonyl carbon followed by, (b) nucleophilic attack of the carboxylate anion onto the protonated carbonyl carbon and (c) dissociation of the tetrahedral intermediate to provide products (Fig. **12**). In addition, the calculation results demonstrate that the rate-limiting step is dependent on the reaction medium. When the calculations were run in the gas phase the rate-limiting step was the formation of the tetrahedral intermediate, whereas when they were conducted in the presence of water the dissociation of the tetrahedral intermediate was the rate-limiting step (see Tables **3** and **4**). Further, when the leaving group ($CH3NH_2$) in **7**-**15** was replaced with a group having a low pKa value the rate-limiting step was the formation of the tetrahedral intermediate, such as in the case where CH3NH2 was replaced with CF3NH2 (see Fig. **12** and Tables **3** and **4**).

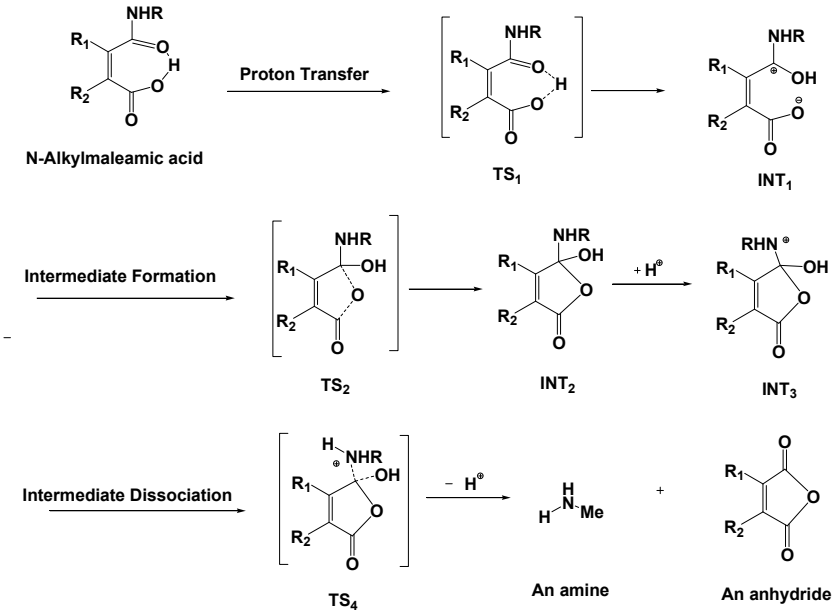


Figure 12: Proposed mechanism for the hydrolysis of N-alkylmaleamic acids 7-15.

Table 3: DFT (B3LYP) calculated kinetic and thermodynamic properties in the gas phase for the acid catalyzed hydrolysis of 7-15

System	$\Delta H_{BGP}^{\ddagger}$ (kcal/mol)	$T\Delta S_{DGP}^{\ddagger}$ (kcal/mol)	$\Delta G_{DGP}^{\ddagger}$ (kcal/mol)	$\Delta H_{FGP}^{\ddagger}$ (kcal/mol)	$T\Delta S_{FGP}^{\ddagger}$ (kcal/mol)	$\Delta G_{FGP}^{\ddagger}$ (kcal/mol)	$\Delta G_{BEth}^{\ddagger}$ (kcal/mol)	$\Delta G_{FEth}^{\ddagger}$ (kcal/mol)	$\Delta G_{BGP}^{\ddagger}$ (kcal/mol) In high pH
7	27.31	-0.77	28.08	32.46	-1.09	33.53	34.27	31.35	45.55
8	13.93	-2.49	16.42	25.67	-1.41	27.08	16.32	21.29	33.59
9	24.41	-0.49	24.90	30.68	-1.89	32.57	29.53	28.23	44.68
10	34.42	-2.35	36.77	41.88	-3.49	45.37	40.23	39.27	53.5
11	13.25	-4.16	17.41	24.55	-2.32	26.87	18.51	22.34	34.52
12	23.83	-0.09	23.92	30.11	-2.01	32.12	29.53	27.50	45.72
13	24.86	-0.17	25.03	30.76	-1.54	32.30	28.35	27.76	------
14	24.08	-0.89	24.87	29.79	-2.58	32.37	29.76	28.20	------
15	17.88	0.64	17.24	24.17	-1.77	25.94	24.46	18.69	------

B3LYP refers to values calculated by B3LYP/6-31G(d, p) method. $\Delta H^{\ddagger}$ is the calculated activation enthalpic energy (kcal/mol). $T\Delta S^{\ddagger}$ is the calculated activation entropic energy (kcal/mol). $\Delta G^{\ddagger}$ is the calculated activation free energy (kcal/mol). D and F refer to tetrahedral intermediate dissociation and tetrahedral intermediate formation. GP and Eth refer to calculated in the gas phase and in ether, respectively.

Table 4: DFT (B3LYP/6-31G(d,p)) calculated kinetic and thermodynamic properties for the acid catalyzed hydrolysis of 7-15

System	$\Delta H_{DW}^{\ddagger}$ (kcal/mol)	$\Delta G_{DW}^{\ddagger}$ (kcal/mol)	log k_{rel} [58]	log EM (Exp) [56]	log EM (Calc)	Es (INT_2) (kcal/mol)	Es (P) (kcal/mol)	Es (GM) (kcal/mol)	$\Delta G_{FW}^{\ddagger}$ (kcal/mol)	Exp $\Delta G_{[58]}^{\ddagger}$ (kcal/mol)
1	32.29	33.06	0	7.724	8.52	20.55	25.08	10.16	26.10	23.70
2	17.56	20.05	4.371	15.86	18.08	16.16	18.93	10.82	17.90	17.30
3	27.93	28.42	1.494	7.742	11.93	17.32	21.70	9.40	24.80	21.14
4	35.76	38.11	-4.377	1.255	4.81	27.89	32.75	12.30	32.16	30.70
5	18.96	23.12	2.732	15.190	15.82	19.25	23.13	9.18	17.89	19.75
6	27.19	27.28	1.516	6.962	12.76	17.59	22.95	5.12	23.87	------
7	27.38	27.55	1.648	8.568	12.57	18.55	24.00	6.20	24.40	------
8	29.23	30.12	------	------	6.36	22.34	27.77	12.86	23.66	
9	15.79	15.15	------	------	21.68	26.92	35.64	28.29	11.97	

B3LYP refers to values calculated by B3LYP/6-31G(d, p) method. $\Delta H^{\ddagger}$ is the calculated activation enthalpic energy (kcal/mol). $T\Delta S^{\ddagger}$ is the calculated activation entropic energy (kcal/mol). $\Delta G^{\ddagger}$ is the calculated activation free energy (kcal/mol). Es refers to strain energy calculated by Allinger's MM2 method [75]. INT_2 and P refer to intermediate 2 and product, respectively. EM = $e^{-(\Delta G\ddagger inter - \Delta G\ddagger intra)/RT}$. DW and FW refer to tetrahedral intermediate dissociation and tetrahedral intermediate formation calculated in water, respectively. Exp Refers to experimental value. Calc refers to calculated DFT values.

The calculations demonstrated that the efficiency of the intramolecular acid-catalyzed hydrolysis by the carboxyl group is remarkably sensitive to the pattern

of substitution on the carbon-carbon double bond. The rate of hydrolysis was found to be linearly correlated with the strain energy of the tetrahedral intermediate or the product. Systems having strained tetrahedral intermediates or products were found to be with low rates and *vice versa* [58, 93,140-145].

COMPUTATIONALLY DESIGNED PRODRUGS BASED ON INTRAMOLECULAR ACID-CATALYZED HYDROLYSIS OF KIRBY'S N-ALKYLMALEAMIC ACIDS - TRANEXAMIC ACID PRODRUGS

Tranexamic acid is a synthetic lysine amino acid derivative. It was originally developed to prevent and reduce excessive hemorrhage in hemophilia patients and reduce the need for replacement therapy during and following tooth extraction. It is often prescribed for excessive bleeding. The mechanism by which tranexamic acid exerts its antifibrinolytic activity is by competitively inhibits the activation of plasminogen to plasmin, a molecule responsible for the degradation of fibrin. Tranexamic acid has roughly 8 times the antifibrinolytic activity of an older analogue, ε-aminocaproic acid. Over the past few years, the use of tranexamic acid has been expanding beyond the small number of hemophilia patients; it is an important agent in decreasing mortality rate due to bleeding in trauma patients; this can be seen from CRASH-2 study which concludes that all causes to mortality, relative risk and relative death due to bleeding were reduced with tranexamic acid group more than placebo group. It can be used safely in women whom undergo lower segment cesarean section, in this operation it was found that tranexamic acid reduces the blood loss during and after surgery, and it is pharmacologically active in reducing intra-operative using of blood heart surgery, hip and knee replacement surgery and liver transplant surgery. Recently, a new oral formulation of tranexamic acid was shown to be safe and effective for treatment of heavy menstrual bleeding. Oral administration of tranexamic acid results in a 45% oral bioavailability. The total oral dose recommended in women with heavy menstrual bleeding was two 650 mg tablets three times daily for 5 days. Accumulation following multiple dosing was reported to be minimal. Post-partum hemorrhage is a leading cause of maternal mortality, accounting for about 100000 maternal deaths every year. Medications used to control postpartum hemorrhage (PPH) are in the category of uterotonic drugs. These drugs stimulate contraction of the uterine muscle, helping to control PPH. The two medications

most commonly used for treatment include oxytocin or misoprostol. In addition, patients are commonly given an IV blood transfusion in cases of severe hemorrhage. In third world countries, availability of blood and fluid replacement may be an issue. One approach to decrease the risk of maternal hemorrhage may be to improve the availability of blood and fluid replacement. An alternative approach is to decrease the likelihood of maternal hemorrhage. Furthermore, all the treatment options mentioned above are intended for intravenous administration; this may not be a viable option in under-developed countries. Therefore, a cheaper oral alternative may be better suited for such circumstances. Tranexamic acid can be used safely and effectively to reduce bleeding resulting from caesarian section (CS). After the withdrawing of aprotinin from worldwide market in November 2009, tranexamic acid is the only marketed antifibrinolytic agent available in the market. Further, it was found that tranexamic acid is also effective in inhibiting the activity of urokinase in urine and it is safe and effective for treating severe hematuria in patient with chronic renal impairment that poorly respond to conventional therapy.

Recent studies have showed that tranexamic acid inhibits the ultraviolet radiation induced pigmentation activity, thus it can be used as bleaching agents. Oral tranexamic acid dosage form was found to be effective and safe in treating malesma, a hypermelanosis disease that occurs in Asian women. Since tranexamic acid is an amino acid derivative and undergoes ionization in physiologic environments its oral bioavailability is expected to be low due to inefficient absorption through membranes. Note the log P (partition coefficient) for tranexamic acid is -1.6. Hence, there is a necessity to design and synthesis relatively more lipophilic tranexamic acid prodrugs that can provide the parent drug in a sustained release manner which might result in better clinical outcome, more convenient dosing regimens and potentially fewer side effects than the original medication. Pharmacologically inactive chemical derivatives that could be used to alter the physicochemical properties of tranexamic acid, in a temporary manner and to increase its usefulness should be lipophilic linkers that are covalently linked to the parent drug and can be converted *in vivo* to the active drug molecule, enzymatically or no enzymatically, to exert a therapeutic effect. Ideally, the prodrugs should be converted to the original drug as soon as the goal is achieved, followed by the subsequent rapid elimination of the released

derivatizing group. For example, tranexamic acid is given by continuous IV infusion resulting in peak plasma concentration following administration. If a slow release prodrug can be prepared, then C_{max} related side effects may be avoided and longer duration exposure may be achieved resulting in potentially better maintenance paradigm. Improvement of tranexamic acid pharmacokinetic properties and hence its effectiveness may increase the absorption of the drug *via* a variety of administration routes, especially the oral and SC injection routes [148-153].

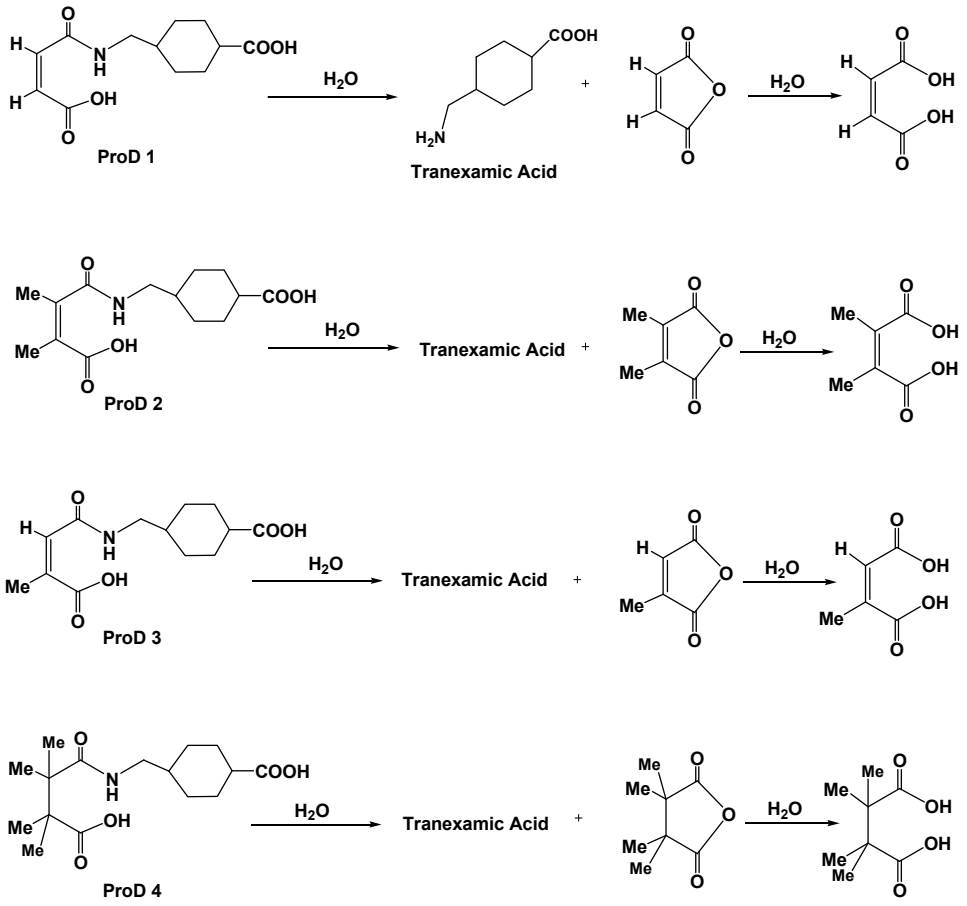


Figure 13: Acid-catalyzed hydrolysis of tranexamic acid prodrugs ProD 1 -ProD 4.

Based on DFT calculations for the acid-catalyzed hydrolysis of several N-alkylmaleamic acid derivatives (Fig. **10**) four tranexamic acid prodrugs were designed (Fig. **13**). The DFT results on the acid catalyzed hydrolysis revealed that

the reaction rate-limiting step is determined on the nature of the amine leaving group. When the amine leaving group was a primary amine or tranexamic acid moiety, the tetrahedral intermediate collapse was the rate-limiting step, whereas in the cases by which the amine leaving group was aciclovir or cefuroxime the rate-limiting step was the tetrahedral intermediate formation. The DFT optimized global minimum, intermediate and transition state structures are illustrated in Figs. **14, 15** and **16**, respectively. Based on the DFT calculated rates the predicted $t_{1/2}$ (a time needed for 50% of the prodrug to be converted into drug) values for tranexamic acid prodrugs **ProD 1- ProD 4** (Fig. **13**) at pH 2 were 556 hours, 253 hours, 70 seconds and 1.7 hours, respectively (for correlation of experimental *vs*. calculated values, see Fig. **17**).

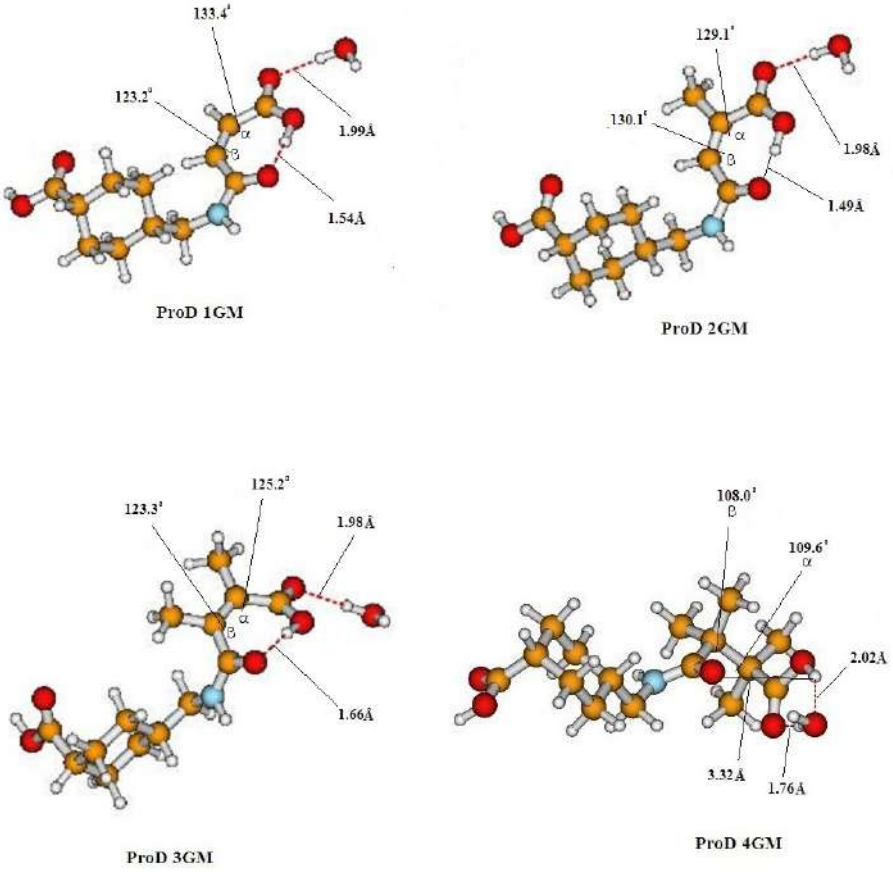


Figure 14: DFT optimized structures for the global minimum (GM) in tranexamic acid ProD 1-ProD 4.

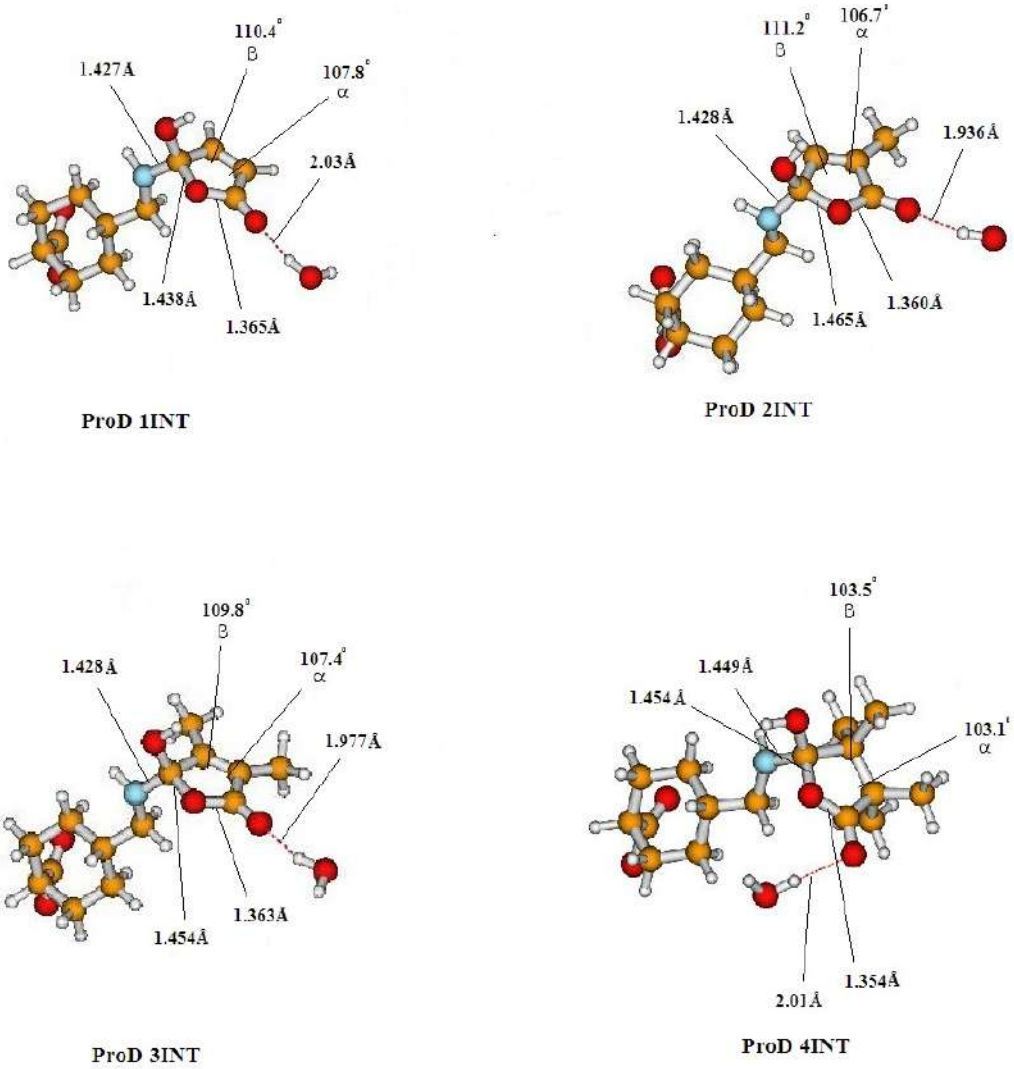


Figure 15: DFT optimized structures for the tetrahedral intermediate (INT) in tranexamic acid ProD 1- ProD 4.

The kinetic study for the acid-catalyzed hydrolysis of tranexamic acid **ProD 1** was carried out in aqueous buffer in the same manner as that done by Kirby on Kirby's enzyme model **7**-**15**. This is in order to explore whether the prodrug hydrolyzes in aqueous medium and to what extent or not, suggesting the fate of the prodrug in the system. Acid-catalyzed hydrolysis kinetics for the synthesized tranexamic acid **ProD**

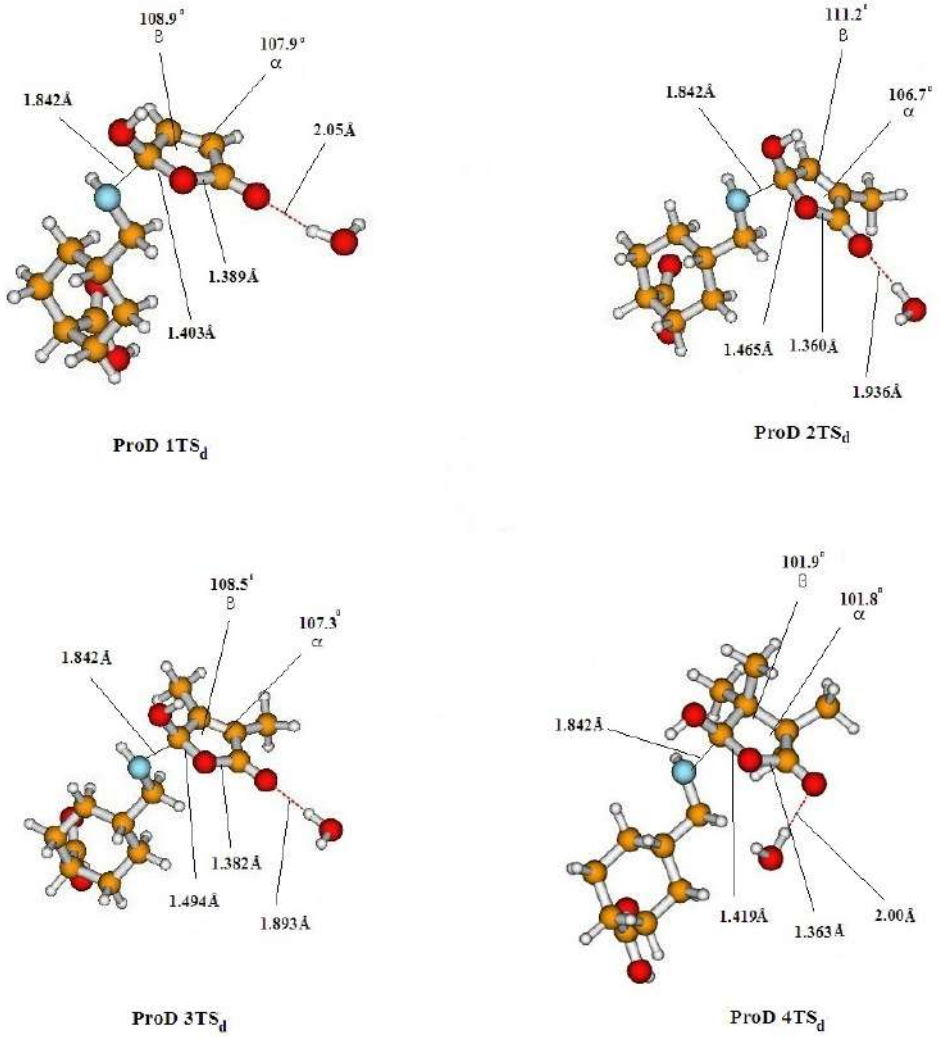


Figure 16: DFT optimized structures for the tetrahedral intermediate dissociation step (TS_d) in tranexamic acid ProD 1- ProD 4.

1 was studied in four different aqueous media: 1 N HCl, buffer pH 2, buffer pH 5 and buffer pH 7.4. Under the experimental conditions the target compounds hydrolyzed to release the parent drug (Fig. **18**) as evident by HPLC analysis. At constant pH and temperature the reaction displayed strict first order kinetics as the k_{obs} was fairly constant and a straight plot was obtained on plotting log concentration of residual prodrug verves time. The rate constant (k_{obs}) and the corresponding half-

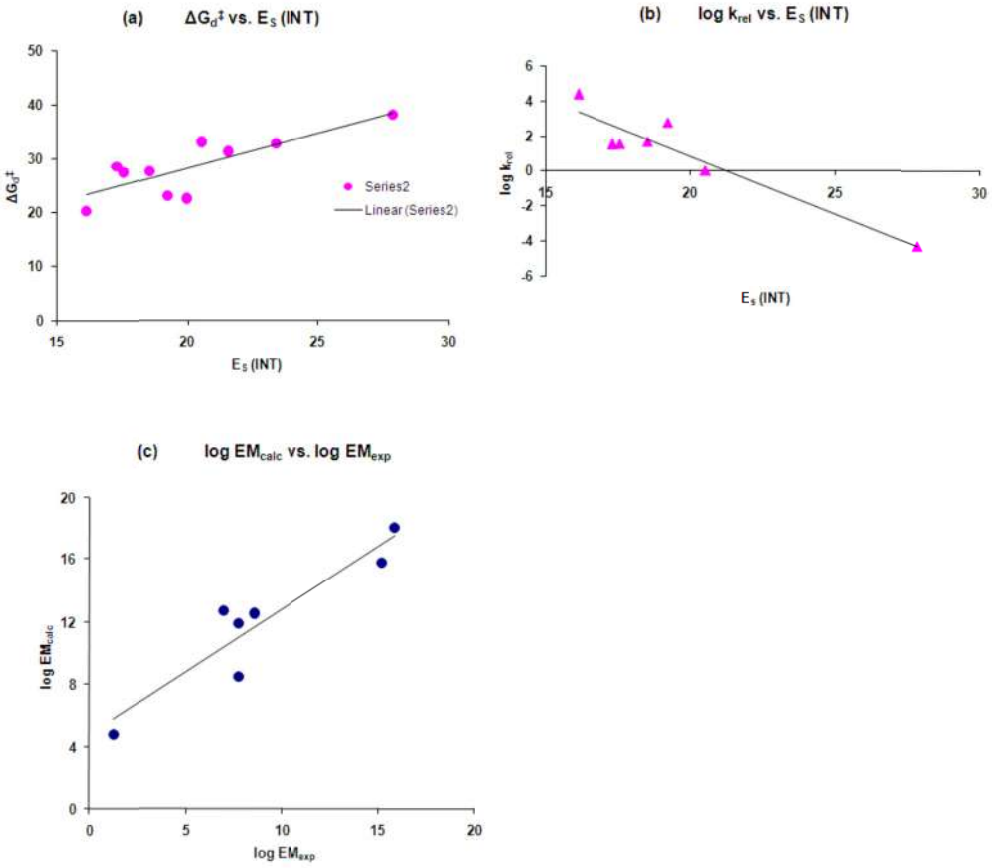


Figure 17: (a) Plot of activation energy for tetrahedral intermediate dissociation $\Delta G_d^{\ddagger}$ *vs*. strain energy of the tetrahedral intermediate (Es_{INT}) in systems 7-13 and tranexamic acid ProD 1- ProD 3. (b) Plot of acid-catalyzed hydrolysis experimental rate (log k_{rel}) *vs*. intermediate strain energy (Es_{INT}) for processes 7-13. (c) Plot of the DFT calculated effective molarity (EM_{calc}) *vs*. the experimental effective molarity (EM_{exp}) for systems 7-13.

lives ($t_{1/2}$) for tranexamic acid prodrug **ProD 1** in the different media were calculated from the linear regression equation correlating the log concentration of the residual prodrug verses time. The kinetic data are listed in Table **5**. The 1N HCl, pH 2 and pH 5 were selected to examine the interconversion of the tranexamic acid prodrug in pH as of stomach, because the mean fasting stomach pH of adult is approximately 1-2 and increases up to 5 following ingestion of food. In addition, buffer pH 5 mimics the beginning small intestine pathway. Finally, pH 7.4 was selected to examine the interconversion of the tested prodrug in blood circulation system. Acid-catalyzed hydrolysis of the tranexamic acid **ProD 1** was found to be higher in 1N HCl than at

pH 2 and 5 (Fig. **18**). At 1N HCl the prodrug was hydrolyzed to release the parent drug in less than one hour. On the other hand, at pH 7.4, the prodrug was entirely stable and no release of the parent drug was observed. Since the pK_a of tranexamic acid **ProD 1** is in the range of 3-4, it is expected at pH 5 the anionic form of the prodrug will be dominant and the percentage of the free acidic form that undergoes the acid-catalyzed hydrolysis will be relatively low. At 1N HCl and pH 2 most of the prodrug will exist as the free acid form and at pH 7.4 most of the prodrug will be in the anionic form. Thus, the difference in rates at the different pH buffers.

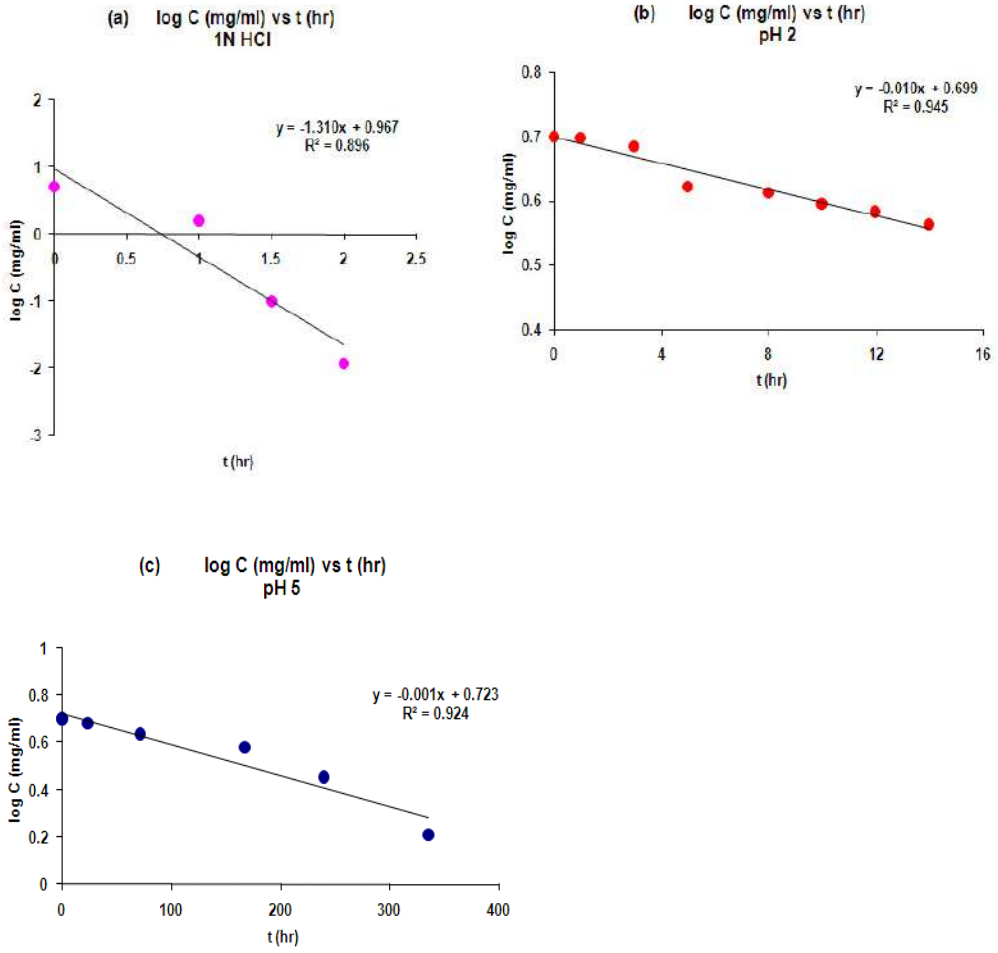


Figure 18: First order hydrolysis plot of tranexamic acid ProD 1in (a) 1N HCl, (b) buffer of pH 2 and (c) buffer pH 5.

Table 5: The observed *k* value and $t_{1/2}$ of tranexamic acid prodrug (ProD 1) In 1N HCl and at pH 2, 5 and 7.4

Medium	k_{obs} (h^{-1})	$t_{½}$ (h)
1N HCl	5.13×10^{-3}	0.9
Buffer pH 2	3.92×10^{-5}	23.9
Buffer pH 5	3.92×10^{-6}	270
Buffer pH 7.4	No reaction	No reaction

Comparison between the calculated $t_{1/2}$ values (556 h) for tranexamic acid **ProD 1** to the experimental value (23.9 h) indicates that the calculated value is about 23 times larger than the experimental. This discrepancy between the calculated and the experimental values might be attributed to the fact that the PCM model (calculations in presence of water) is not capable for handling calculations in acidic aqueous solvent (medium) since the dielectric constant for pH 2 aqueous solutions is not known. In the study calculations the value of 78.39 (dielectric constant for pure water) was used instead. The $t_{1/2}$ experimental value at pH 5 was 270 hours and at pH 7.4 no interconversion was observed. The lack of the reaction at the latter pH might be due to the fact that at this pH tranexamic acid **ProD 1** exists solely in the ionized form (pK_a about 4). As mentioned before the free acid form is a mandatory requirement for the reaction to proceed. On the other hand, tranexamic acid **ProD 4** has a higher pK_a than tranexamic acid **ProD 1** (about 6 *vs*. 4). Therefore, it is expected that the interconversion rate of tranexamic acid **ProD 4** to its parent drug, tranexamic acid, at all pHs studied will be higher (log EM for **ProD 4** is 14.33 *vs*. 9.53 for **ProD 1**).

Future strategy to achieve desirable tranexamic acid prodrugs capable of releasing tranexamic acid in a controlled manner and enhancing the parent drug bioavailability is: (i) synthesis of tranexamic acid **ProD 4**; (ii) kinetic studies (*in vitro*) of **ProD 4** will be performed in at pH 6.5 (intestine) and pH 7.4 (blood circulation system) (iii) *in vivo* pharmacokinetic studies will be done in order to determine the bioavailability and the duration of action of the tested prodrug. Furthermore, based on the *in vivo* pharmacokinetics characteristics of tranexamic acid **ProD 4** new prodrugs may be design and synthesized.

BITTERLESS ATENOLOL PRODRUGS

Atenolol, 4-[2-hydroxy-3-[(1-methylethyl) amino] propoxy] benzene acetamide is a relatively polar hydrophilic compound with a log partition coefficient (octanol/ water) of 0.23. Lipid insoluble hydrophilic compounds such as atenolol, sotalol, nadolol are excreted only by the kidneys and have low brain penetration. Atenolol is a selective $ß_1$-adrenoceptor antagonist, applied in the treatment of numerous cardiovascular disorders including: hypertension, angina, acute myocardial infarction, supraventricular tachycardia, ventricular tachycardia, and the symptoms of alcohol withdrawal *via* restricting certain nerve impulses, thereby controlling the rate and force of contraction and consequently reducing blood pressure in addition to its treatment of Angina Pectoris. Atenolol is marketed as tablets and an injectable formulation [154, 155]. Atenolol has a *pKa* of 9.6; it undergoes ionization in the stomach and intestine thus its oral bioavailability is low due to inefficient absorption through membranes. The bioavailability of atenolol is between 45% and 55% of the given dose and is not increased by administration of the drug in a solution form [156-158]. About 50% of administered atenolol is absorbed; however, most of the absorbed quantity reaches the systemic circulation. Atenolol peak blood levels are reached within two to four hours after ingestion. Differently from propranolol or metoprolol, atenolol is resistant to metabolism by the liver and the absorbed dose is eliminated by renal excretion. More than 85% of I.V. dose is excreted in urine within 24 hours compared to 50% for an oral dose. Only 6-16% is protein-bound resulting in relatively consistent plasma drug levels with about a four-fold inter-patient variation. The elimination half-life of atenolol is between 6 to 7 hours and there is no alteration of kinetic profile of a drug by chronic administration. Atenolol is one of the most important medicines used for prevention of several types of arrhythmias in childhood, but unfortunately it is still unlicensed [159]. On the other hand, atenolol is indicated as a first-step therapy for hypertension in elderly patients, who have difficulty in swallowing and, thus, tablets and capsules are frequently avoided. The ease of administration makes a liquid formulation an ideal dosage form for such patients [160]. Therefore, extemporaneous compounding (off label), involves preparation of an oral liquid from a pure drug powder is required. However, formulations compounded from tablets and pure

active drug suffer instability and are only stable for less than one week [142, 161, 162]. Furthermore, atenolol bitterness is considered as a great challenge to health sector when used among children and geriatrics [163]. The main problem in oral administration of bitter drugs such as atenolol is incompliance by the patients [164] and this can be overcome by masking the bitterness of the drug either by decreasing its oral solubility on ingestion or eliminating the interaction of drug particles to taste buds [165].

Previous studies on stability of atenolol ester prodrugs for the use in transdermal preparations have shown that these ester derivatives are much more stable than the corresponding alcohol, atenolol, when they are formulated in aqueous solutions [166-167]. On the other hand, the only atenolol prodrug intended for oral dosage form use was atenolol aspirinate prodrug; it is described for antihypertensive therapy to reduce cardiovascular death, stroke, and myocardial infarction (MI), however, recent studies showed that coupling of atenolol with acetyl salicylic acid by means of an ester linkage did not produce efficient pharmacological profile, neither *in vitro* nor *in vivo* [168]. Continuing our study on the design and synthesis of novel prodrugs for drugs with bitter sensation, and medicines having low bioavailability we sought to: (1) design atenolol prodrugs that can be (i) formulated in aqueous solutions and maintain stability over a long period of time, (ii) lack bitterless and have the capability to undergo intraconversion in physiological environment to provide the parent active drug, atenolol, in a programmable manner and (2) synthesize, characterize and *in vitro* study the kinetic of the interconversion of the designed prodrugs in different media: 1 N HCl and at buffers of pH 2, pH 5 and pH 7.4. Our proposed atenolol prodrugs that were designed based on the acid-catalyzed hydrolysis reactions of N-alkyl maleamic acids **7**-**15** (Fig. **10**) are depicted in Fig. **19**. As illustrated in Fig. **19**, the only difference between the proposed atenolol prodrugs and the parent drug, atenolol, is that the amine group in atenolol was replaced with an amide moiety. This chemical change is expected to increase the stability of the alcohol derivative (prodrug) compared to the corresponding amine alcohol, atenolol, due to general chemical stability for tertiary alcohols over amine alcohols. In addition, stability studies on atenolol ester derivatives showed the ester derivatives to be much more stable than their corresponding alcohols upon

formulating in aqueous solutions. On the other hand, kinetic study of atenolol and propranolol revealed that increasing the lipophilicity of the drug leads to an increase in the stability of its aqueous solutions. Based on that, it is expected that atenolol prodrugs (atenolol amide derivatives) shown in Fig. **19** will be more resistant to heat or/oxidation when standing in aqueous solutions.

In similar to paracetamol, it is expected that blocking the amine group in atenolol with a suitable linker might inhibit the interaction between the amine group in atenolol and its bitter taste receptor and hence masks its bitterness. The nature of the bitter taste receptors with either paracetamol (*via* the phenolic group) or atenolol (*via* the amine group) is likely to be as a result of hydrogen or ionic bonding between the substrate and the receptor.

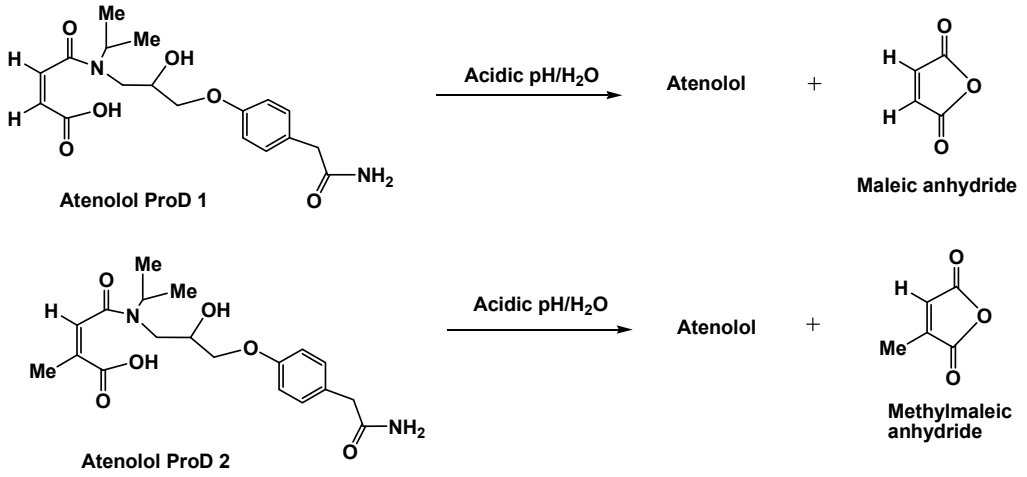


Figure 19: Acid-catalyzed hydrolysis of atenolol ProD 1-ProD 2.

As shown in Fig. **19**, the proposed atenolol prodrugs, atenolol **ProD 1** and atenolol **ProD 2**, have a hydroxyl and carboxylic acid groups (hydrophilic moiety) and the rest of the prodrug molecule is a lipophilic moiety, where the combination of both groups ensures a molecule with a moderate hydrophilic lipophilic balance (HLB).

It is worthy to note that the HLB value of atenolol prodrug moiety will be largely determined on the medium (physiologic environment) by which the prodrug is dissolved. For instance, in the stomach (pH 1-2), atenolol prodrugs will exist in the

free carboxylic acid form whereas in the blood circulation (pH 7.4) the carboxylate anion form will be dominant. It is planned that atenolol **ProD 1- ProD 2** (Fig. **19**) will be formulated as sodium carboxylate salts since this form is expected to be stable in neutral aqueous medium. However, upon exposure to stomach (pH less than 3) the prodrugs will exist mainly as free carboxylic acid forms thus enabling the acid-catalyzed hydrolysis to proceed. The DFT calculations for the acid-catalyzed hydrolysis of atenolol **ProD 1-ProD 2** demonstrated (the optimized structures for the global minimum and transition state structures are shown in Fig. **20**) that the reaction rate is linearly correlated with the following: (a) the strain energy of the tetrahedral intermediate and product and the strain energy difference between the intermediate and the reactant. (b) The distance between the hydroxyl oxygen of the carboxyl group and the amide carbonyl carbon, and (c) the attack angle by which the approach step commences. Based on the experimental $t_{1/2}$ (the time needed for the conversion of 50% of the reactants to products) and EM (effective molarity) values for processes **7-15** we have calculated the $t_{1/2}$ values for the conversion of the two atenolol prodrugs to the parent drug, atenolol. The calculated $t_{1/2}$ values for atenolol **ProD 1-2** are predicted to be 65.3 hours and 11.8 minutes, respectively. Thus, the rate by which atenolol prodrug undergoes cleavage to release atenolol can be determined according to the nature of the linker of the prodrug (Kirby's N-alkylmaleamic acids **7-15**, Fig. **10**).

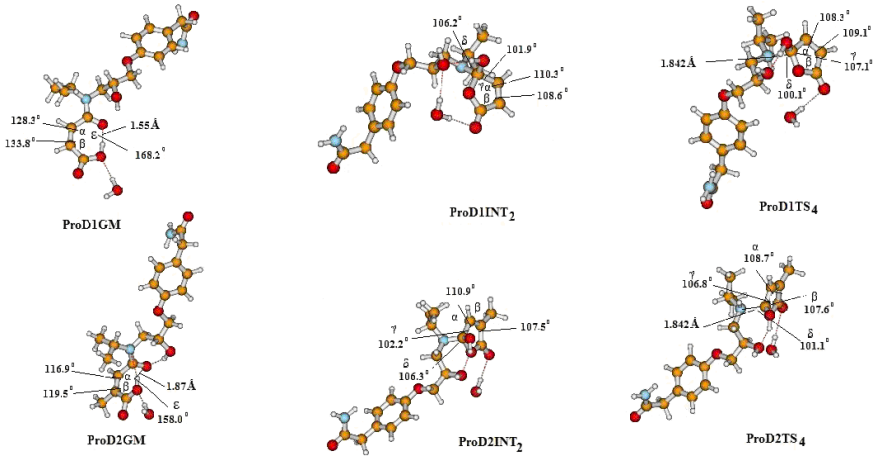


Figure 20: DFT optimized structures for the global minimum (GM), intermediate (INT_2) and transition state (TS_4) in atenolol ProD 1- 2.

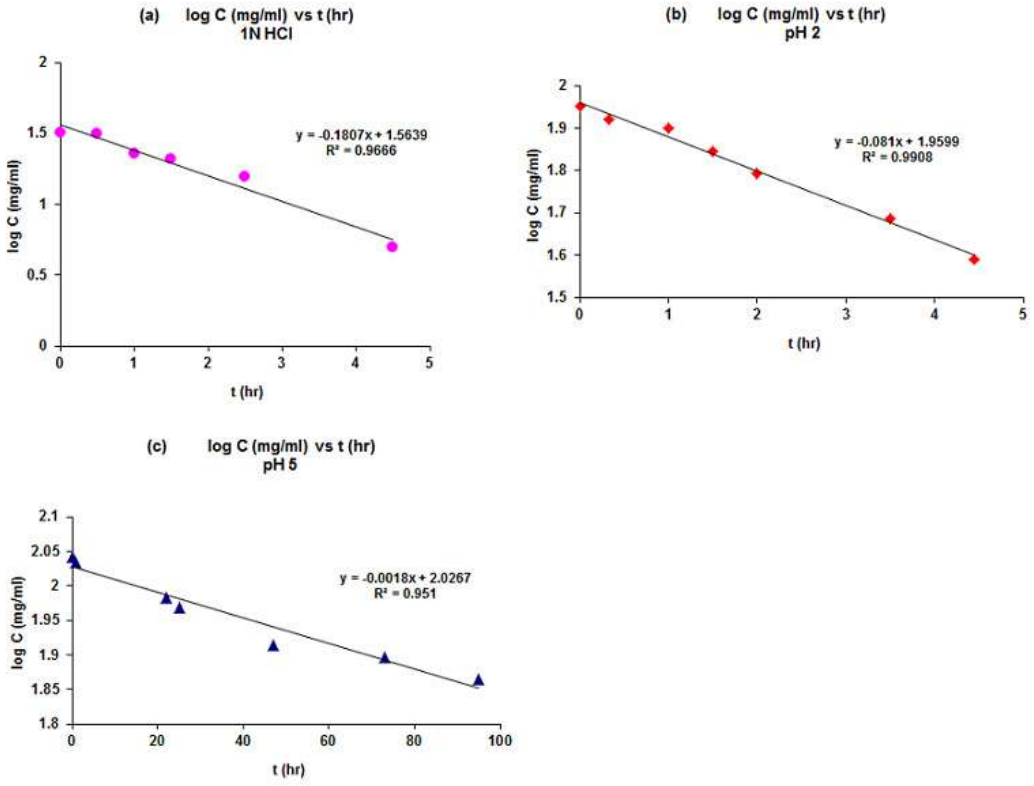


Figure 21: First order hydrolysis plot of atenolol ProD 1 in (a) 1N HCl, (b) buffer pH 2 and (c) buffer pH 5.

ATENOLOL PROD 1 KINETIC STUDY

The kinetics of the acid-catalyzed hydrolysis study for atenolol **ProD 1** was carried out in an aqueous buffer in a similar manner to that done by Kirby on N-alkylmaleamic acids **7**-**15**. Acid-catalyzed hydrolysis kinetics of the synthesized atenolol **ProD 1** was studied in four different aqueous media: 1 N HCl and buffers pH 2, pH 5 and pH 7.4. Under the experimental conditions, the target prodrug (atenolol **ProD 1**) was hydrolyzed to release the parent drug, atenolol, (Fig. **21**) as was evident by HPLC measurements. The reaction displayed strict first order kinetics as the k_{obs} was fairly constant and a straight line was obtained from a plot of log concentration of residual prodrug verses time. The rate constant (k_{obs}) and the corresponding half-lives ($t_{1/2}$) for atenolol **ProD 1** in the different media were

calculated from the linear regression equation correlating the log concentration of the residual prodrug *vs*. time. The kinetic data, k_{obs} and $t_{1/2}$ values, are listed in Table **6**. Acid-catalyzed hydrolysis of the atenolol **ProD 1** was found to be higher in 1N HCl than at pH 2 and 5 (Fig. **21**). At 1N HCl the atenolol **ProD 1** was hydrolyzed to release the parent drug in 2.53 hours. On the other hand, at pH 7.4, the prodrug was entirely stable and no release of the parent drug was observed. Since the pK_a of atenolol **ProD 1** carboxylic acid is in the range of 3-4, it is expected at pH 5 the anionic form of the prodrug will be dominant and the percentage of the free acidic form that undergoes the acid-catalyzed hydrolysis will be relatively low. At 1N HCl and pH 2 most of the prodrug will exist as the free acid form, whereas at pH 7.4 most of the prodrug will be in the anionic form. Thus, the difference in rates at the different pH buffers.

Table 6: First order hydrolysis plot of atenolol ProD 1 in (a) 1N HCl, (b) buffer pH 2 and (c) buffer pH 5

Medium	k_{obs} (hours^{-1})	$t_{1/2}$ (hours)
1 N HCl	4.95 x 10 $^{-4}$	2.53
Buffer pH 2	2.22 x 10 $^{-4}$	3.82
Buffer pH 5	2.75 x 10^{-6}	133
Buffer pH 7.4	----------	---------

The QM calculations at different levels demonstrated that the efficiency of atenolol **ProD 1**- **ProD 2** is largely sensitive to the pattern of substitution on the carbon-carbon double bond and nature of the alkyl group on amide nitrogen.

Using the correlation equation obtained from the plot of the calculated and experimental EM values the $t_{1/2}$ values of two different atenolol prodrugs (**ProD 1**- **ProD 2**) were estimated.

Comparison between the calculated $t_{1/2}$ values 63.2 hours) for atenolol **ProD 1** to the experimental value (3.82 hours) indicates that while the value obtained by B3LYP/6-31G (d,p) is overestimated (about 17 times larger than the experimental) the values obtained by mpwpw91/6-31+G(d,p) were much more closer 6.3 hours. This discrepancy between the calculated and experimental values might be attributed to (1) B3LYP/6-31G(d,p) is a DFT method without dispersion

corrections and (2) PCM solvation model (calculations in presence of water) is not capable of handling calculations in acidic aqueous solvent (medium) since the dielectric constant for pH 2 aqueous solutions is not known. In the study, calculations the value of 78.39 (dielectric constant for pure water) was used instead.

The $t_{1/2}$ experimental value at pH 5 was 133 hours and at pH 7.4 no interconversion was observed. The lack of the reaction at the latter pH might be due to the fact that at this pH atenolol **ProD 1** exists solely in the ionized form (pK_a about 3-4). As mentioned before the free acid form is a mandatory requirement for the reaction to proceed.

Future strategy to achieve more efficient atenolol prodrugs capable of increasing the liquid formulation stability, eliminating atenolol bitterness and releasing the parent drug in a programmable manner is: (a) synthesis of atenolol prodrugs having *pKa* around 6 (intestine pH) *in vitro* kinetics for this prodrug will be performed in at pH 6.5 (intestine) and pH 7.4 (blood circulation system) (iii) *in vivo* pharmacokinetic studies will be done in order to determine the bioavailability and the duration of action of the tested prodrug.

BITTERLESS AMOXICILLIN AND CEPHALEXIN PRODRUGS

Most of the antibacterial agents that are commonly used suffer unpleasant taste and a respected number of them are characterized with bitter taste. For example, amoxicillin, cephalexin and cefuroxime axetil have an extremely unpleasant and bitter taste which is difficult to mask. This is a particular problem in geriatric patients who cannot swallow whole tablets or when small doses are required. Even the antibacterial suspension is difficult for pediatrics to administer due to its better and unpleasant taste. It is widely assumed that the extremely bitter and unpleasant taste of these antibacterial drugs is due to the formation of intermolecular force/s between the drug and the active site of the bitter taste receptor/s. The intermolecular bond/s is/are most likely due to formation either *via* hydrogen bond or ionic bond of the amide (in cefuroxime) or amine (in amoxicillin and cephalexin) group to the active site of the bitter taste receptors. Antimicrobial agents are classified according to their specific mode of action

against bacterial cell. By which these agents may interfere with cell wall synthesis, inhibit protein synthesis, interfere with nucleic acid synthesis or inhibit a metabolic pathway. They have a broad spectrum of activity against both gram-positive and gram-negative bacteria. Among these agents: β-lactams- penicillins, cephalosporins, carbapenems and monobactams. They are preferred because of their efficacy, safety, and because their activity can be extended or restored by chemical manipulation. Inevitably, however, their usage has been restricted because of their bacterial resistance [169-170].

AMOXICILLIN

Amoxicillin is an oral semi-synthetic penicillin, moderate-spectrum, bacteriolytic, lactam antibiotic used to treat bacterial infections caused by susceptible microorganisms by which it is susceptible to the action by the β-lactamases. Amoxicillin has a bactericidal action and acts against both Gram positive and Gram-negative microorganisms by inhibiting the biosynthesis and repair of the bacterial mucopeptide wall. It is usually the drug of choice within its class because it is well absorbed following oral administration. Amoxicillin presents some outstanding advantages in comparison with other aminopenicillins, such as: a better absorption from the intestinal tract, better capacity for reaching effective concentrations at the sites of action and a more rapid capacity for penetrating the cellular wall of Gram-negative microorganisms. Aminopenicillins are frequently prescribed agents for the oral treatment of lower respiratory tract infections and are generally highly effective against *S. pneumonia* and non- β-lactamase-producing *H. influenza.* Amoxicillin is mostly common antibiotics prescribed for children. It has high absorption after oral administration which is not altered and affected by the presence of food. Amoxicillin dose reaches C_{max} about 2 hours after administration and is quickly distributed, and eliminated by excretion in urine (about 60%- 75%). The antibacterial effect of amoxicillin is extended by the presence of a benzyl ring in the side chain. Because amoxicillin is susceptible to degradation by β-lactamase-producing bacteria, which are resistant to a broad spectrum of β-lactam antibiotics, such as penicillin; for this reason, it is often combined with clavulanic acid, a β-lactamase inhibitor. This increases effectiveness by reducing its susceptibility to β-lactamase resistance. Amoxicillin has two ionizable groups in the physiological range (the amino group in α-

position to the amide carbonyl group and the carboxyl group). Amoxicillin has a good pharmacokinetics profile with bioavailability of 95% if taken orally, its half-life is 61.3 minutes and it is excreted by the renal and less than 30 % bio-transformed in the liver [169-172].

CEPHALEXIN

Cephalexin is a first-generation cephalosporin antibiotic, which was chosen as the model drug candidate to obtain dosage with improved stability, palatability and attractive pediatric elegance, cost effective with ease of administration. Cephalosporins are widely used for the treatment of skin infections because of their safety profile, and their wide range of activity against both gram positive and gram negative microorganism. Cephalexin is also used for the treatment of articular infections as a rational first-line treatment for cellulitis. It is a useful alternative to penicillins hypersensitivity, and is thought to be safe in a patient with penicillin allergy; but caution should always be taken, that's because cephalexin and other first-generation cephalosporins are known to have a modest cross-allergy in patients with penicillin hypersensitivity. In addition, cephalexin is also effective and used in the treatment of β-hemolytic streptococcal throat infections. Cephalexin works by interfering with the bacteria's cell wall formation, causing it to rupture, and thus killing the bacteria. The compound is zwitterion; contains both a basic and an acidic group, the isoelectric point of cephalexin in water is approximately 4.5 to 5. Cephalexin has a good pharmacokinetic profile by which it is well absorbed, 80% excreted unchanged in urine within 6 hours of administration. Cephalexin's half-life is 0.5-1.2 hours and it is excreted *via* the renal. It is used for the treatment of infections including otitis media, streptococcal pharyngitis, bone and joint infections, pneumonia, cellulitis and UTI [173-180].

IN VITRO INTRACONVERSION OF AMOXICILLIN AND CEPHALEXIN PRODRUGS TO THEIR PARENT DRUGS

Based on our previously reported DFT calculations and on the experimental data for the acid-catalyzed hydrolysis of amide acids **7**-**15** (Fig. **10**) [58, 93], two amoxicillin and cephalexin prodrugs were proposed (Figs. **22** and **23**, respectively). As shown in Figs. **22** and **23**, the antibacterial prodrugs, amoxicillin

ProD 1 and cephalexin **ProD 1** molecules are composed of an amide acid promoiety, containing a carboxylic acid group (hydrophilic moiety) and the rest of the antibacterial prodrug molecule (a lipophilic moiety).

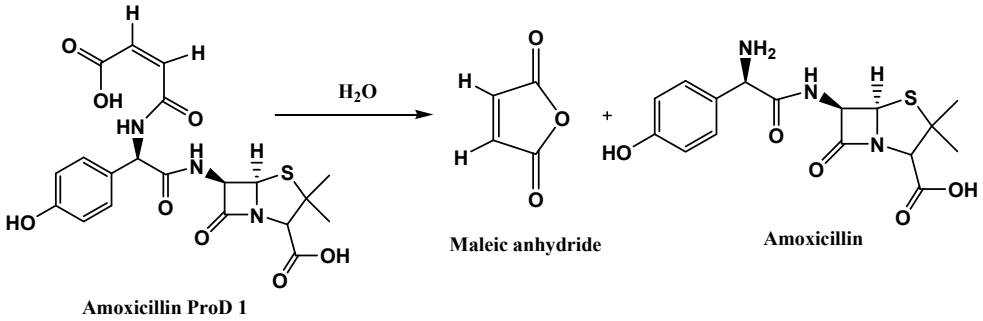


Figure 22: Acid-catalyzed hydrolysis of amoxicillin ProD 1.

The combination of both, the hydrophilic and lipophilic groups provides a prodrug entity with a potential to be with a high permeability (a moderate HLB). It should be emphasized, that the HLB value of the prodrug entity will be determined upon the pH of the target physiological environment. In the stomach where the pH is in the range 1-2, it is expected that prodrugs, amoxicillin **ProD 1** and cephalexin **ProD 1** will be in a free carboxylic acid form (a relatively high hydrophobicity) whereas in the blood stream circulation where the is pH 7.4 a carboxylate anion (a relatively low hydrophobicity) is expected to be predominant form. Our strategy was to prepare amoxicillin **ProD 1** and cephalexin **ProD 1** as sodium or potassium carboxylates due to their high stability in neutral aqueous medium. It should be indicated that compounds **7**-**15** undergo a relatively fast hydrolysis in acidic aqueous medium whereas they are quite stable at neutral pH.

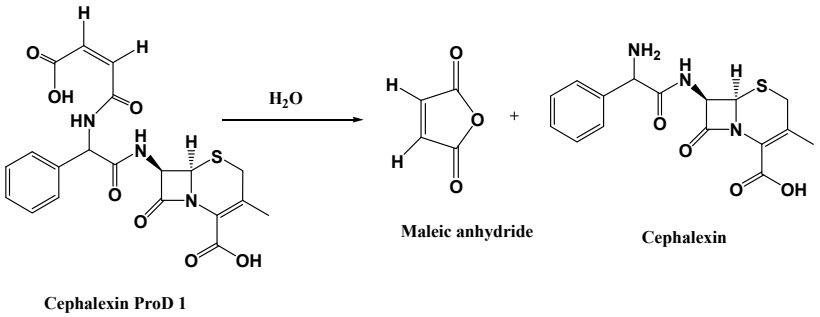


Figure 23: Acid-catalyzed hydrolysis of cephalexin ProD 1.

The hydrolysis kinetic studies for amoxicillin **ProD 1** and cephalexin **ProD 1** were carried out in aqueous buffers in the same manner to that executed by Kirby *et al.* on maleamic acids **7-15**. This is to investigate whether the antibacterial prodrugs undergo hydrolysis in aqueous medium and to what extent or not, suggesting the fate of the prodrugs in the system. The kinetics for the acid-catalyzed hydrolysis of the synthesized amoxicillin **ProD 1** and cephalexin **ProD 1** were carried out in four different aqueous media: 1 N HCl and pH 2.5, pH 5 and pH 7.4 buffers. Under the experimental conditions the two antibacterial prodrugs intraconverted to release the parent drugs (Figs. **24** and **25**) as was determined by HPLC analysis. For both amoxicillin and cephalexin prodrugs, at constant temperature and pH the hydrolysis reaction displayed strict first order kinetics as the k_{obs} was quite constant and a straight line was obtained on plotting log concentration of residual prodrug verves time. The rate constant (k_{obs}) and the corresponding half-lives ($t_{1/2}$) for amoxicillin **ProD 1**and cephalexin **ProD 1** in the different media were calculated from the linear regression equation obtained from the correlation of log concentration of the residual prodrug verses time. The kinetic data for amoxicillin **ProD 1**and cephalexin **ProD 1** are listed in Tables **7** and **8**, respectively. Acid-catalyzed hydrolysis of both, amoxicillin **ProD 1** and cephalexin **ProD 1** was found to be much higher in 1N HCl than at pH 2.5 and 5 (Figs. **24** and **25**). At 1N HCl the $t_{½}$ values for the intraconversion of amoxicillin **ProD 1** and cephalexin **ProD 1** were about 2.5 hours. On the other hand, at pH 7.4, both prodrugs were quite stable and no release of the parent drugs was observed. At pH 5 the hydrolysis of both prodrugs was too slow**.** This is because the pK_a of amoxicillin **ProD 1** and cephalexin **ProD 1** is in the range of 3-4, it is expected that at pH 5 the anionic form of the prodrug will be dominant and the percentage of the free acidic form that undergoes an acid-catalyzed hydrolysis will be relatively low. At 1N HCl and pH 2.5 most of the prodrug will exist as the free acid form and at pH 7.4 most of the prodrug will be in the anionic form. Thus, the discrepancy in rates at the different pH buffers.

Table 7: The observed *k* value and $t_{1/2}$ of amoxicillin ProD 1 in 1N HCl and at pH 2, 5 and 7.4

Medium	k_{obs} (h^{-1})	$t_{1/2}$ (h)
1 N HCl	2.33×10^{-4}	2.5
Buffer pH 2.5	9.60×10^{-5}	7
Buffer pH 5	7.55×10^{-6}	81
Buffer pH 7.4	No reaction	----

Table 8: The observed *k* value and $t_{1/2}$ of cephalexin ProD 1 in 1N HCl and at pH 2, 5 and 7.4

Medium	k_{obs} (h^{-1})	$t_{1/2}$ (h)
1 N HCl	2.41 x 10^{-4}	2.4
Buffer pH 2.5	4.17 x 10^{-5}	14
Buffer pH 5	No reaction	---
Buffer pH 7.4	No reaction	---

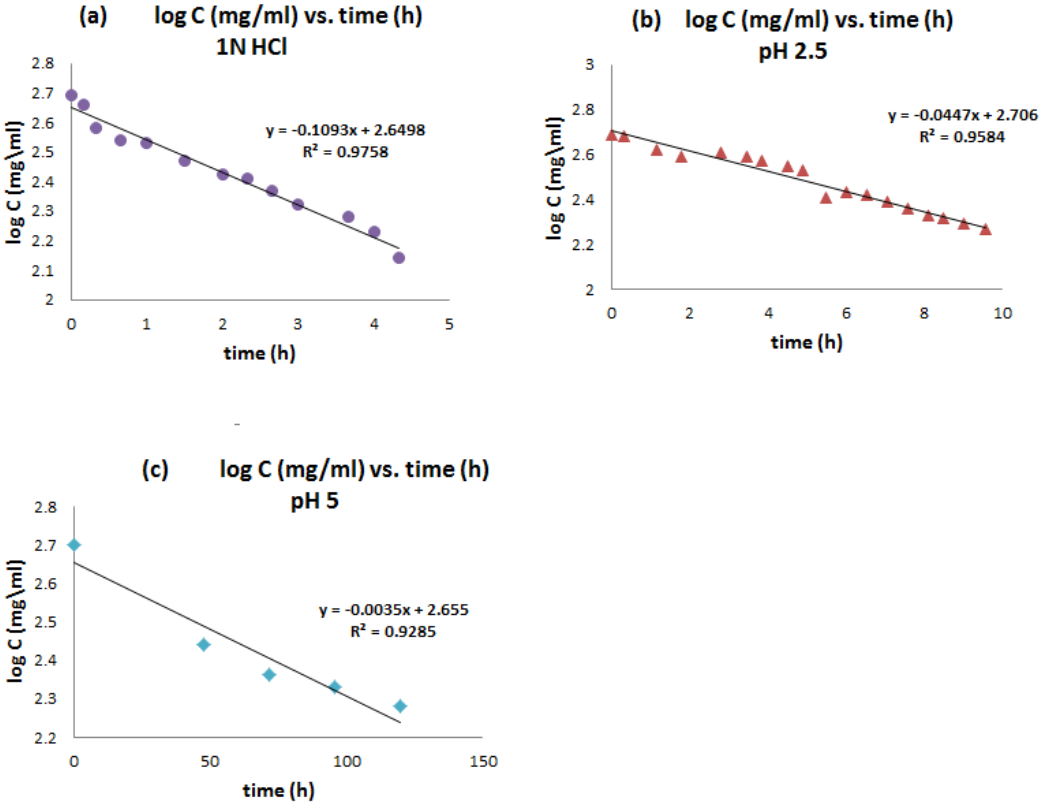


Figure 24: First order hydrolysis plot of amoxicillin ProD 1 in (a) 1N HCl, (b) buffer pH 2.5 and (c) buffer pH 5.

CONCLUDING REMARKS

Our recent studies on intramolecularity have demonstrated that exploring the reaction mechanisms for all enzyme models mentioned in this chapter has helped

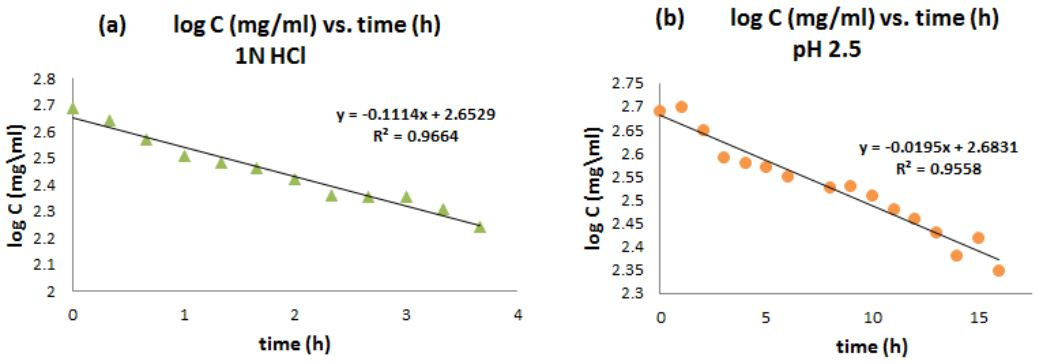


Figure 25: First order hydrolysis plot of cephalexin ProD 1 in (a) 1N HCl, (b) buffer pH 2.5 and (c) buffer pH 5.

to determine the factors affecting the intramolecular reaction rate. Unraveling the mechanisms has allowed for better design of efficient chemical devices that have been utilized as prodrugs linkers that can be covalently attached to active drugs which can chemically, but not enzymatically, be converted to release the active drugs in a controlled manner. For example, studying the mechanism for a proton transfer in Kirby's N-alkylmaleamic acids (enzyme model) was explored for the design of a number of prodrugs such as tranexamic acid for bleeding conditions, acyclovir as antiviral drug for the treatment for herpes simplex [141], atenolol for treating hypertension with enhanced stability and bioavailability without bitter sensation [142] and statins for lowering cholesterol levels in the blood [181]. In addition, prodrugs for masking the bitter taste of paracetamol and antibacterial drugs such as cefuroxime, amoxicillin and cephalexin were also designed and synthesized [143]. The role of the linkers in the antibacterial prodrugs is to block the free amine, which is responsible for the drug bitterness, and to enable the release of the drug in a controlled manner. Menger's Kemp acid enzyme model was utilized for the design of dopamine prodrugs for the treatment for Parkinson's disease [182]. Prodrugs for dimethyl fumarate for the treatment psoriasis was also designed, synthesized and studied [183]. Furthermore, unraveling the mechanism of Kirby's acetals has led to the design and synthesis of novel prodrugs of aza-nucleosides for the treatment for myelodysplastic syndromes [184], atovaquone prodrugs for the treatment for malaria [112], less bitter paracetamol prodrugs to be administered to children and elderly as antipyretic and pain killer [139], and

prodrugs of phenylephrine as decongestant [185]. In these examples, the prodrug moiety was linked to the hydroxyl group of the active drug such that the drug-linker moiety (prodrug) has the potential to interconvert when exposed into physiological environments such as stomach, intestine, and/or blood circulation, with rates that are solely dependent on the structural features of the pharmacologically inactive promoiety (Kirby's enzyme model).

ACKNOWLEDGEMENTS

The author would like to acknowledge funding by the German Research Foundation (DFG, ME 1024/8-1).

CONFLICT OF INTEREST

The author states no conflict of interest and has received no payment in preparation of this chapter.

REFERENCES

[1] DiMasi, J. A.; Hansen, R. W.; Grabowski H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ*. **2003**, *22*,151-185.

[2] DiMasi, J. A. The value of improving the productivity of the drug development process: faster times and better decisions. *Pharmacoeconomics* **2002**, *20*, 1-10.

[3] Huttunen, K. M.; Raunio, H.; Rautio, J. Prodrugs from serendipity to rational design. *Pharmacol Rev*. **2011**, *63*, 750-771.

[4] Stella, V. J.; Nti-Addae, K. W. Prodrug strategies to overcome poor water solubility. *Adv. Drug Deliv. Rev*. **2007**, *59*, 677-694.

[5] Dahan, A.; Khamis, M.; Agbaria, R.; Karaman, R. Targeted prodrugs in oral delivery: the modern molecular biopharmaceutical approach. *Expert Opinion on Drug Delivery* **2012**, *9*, 1001-1013.

[6] Karaman, R.; Fattash, B.; Qtait, A. The future of prodrugs - design by quantum mechanics methods. *Expert Opinion on Drug Delivery* **2013**, *10*, 713-729.

[7] Karaman, R. Prodrugs design based on inter- and intramolecular processes. *Chem. Biol. Drug Des.* **2013**, *82*, 643-668.

[8] Gonzalez, F.J.; Tukey, R. H. Drug metabolism. In: Brunton L.L., Lazo J.S., Parker K.L., editors. Goodman and Gilman's The Pharmacological Basis of Therapeutics. New York: The McGraw-Hill Companies, Inc., 2006, p. 71-91.

[9] Testa, B.; Kramer, S.D. The biochemistry of drug metabolism-an introduction: part 2. Redox reactions and their enzymes. *Chem. Biodivers* **2007**, *4*, 257-405.

[10] Gangwar, S.; Pauletti, G.M.; Wang, B.; Siahaan, T.J.; Stella, V.J.; Borchardt, R.T. Prodrug strategies to enhance the intestinal absorption of peptides. *DDT* **1997**, *2*, 148-155.

[11] Wang, W.; Jiang, J.; Ballard, C.E.; Wang, B. Prodrug approaches to the improved delivery of peptide drugs. *Curr. Pharm. Des*. **1999**, *5*, 265-287.

[12] Chan, O.H.; Stewart, B. H. Physicochemical and drug-delivery considerations for oral drug bioavailability. *Drug Discov. Today* **1996**, *1*, 461-473.
[13] Ohlstein, E. H.; Ruffolo, R. R. Jr; Elliott, J. D. Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol*. **2000**, *40*, 177-91.
[14] Harper, N,J. Drug latentiation. *J Med Pharm Chem*. **1959**, *1*, 467-500.
[15] Huttunen, K.M.; Raunio, H.; Rautio, J. Prodrugs-from serendipity to rational design. *Pharmacol Rev*. **2011**, *63*(3), 750-71.
[16] Albert, A. Selective Toxicity: The PhysicoChemical Basis of Therapy. 7 ed. New York: Chapman and Hall 1985.
[17] Bertolini, A.; Ferrari, A.; Ottani, A.; Guerzoni, S.; Tacchi, R.; Leone, S. Paracetamol: new vistas of an old drug. *CNS Drug Rev*. **2006**, *12*(3-4), 250-75.
[18] Burke, A.S., Emer; F.; Garret, A. Goodman and Gilman's the pharmacological basis of therapeutics. 11 ed. New York: McGraw-Hill; 2006.
[19] Stella, V.J.; Borchardt, R.T.; Hageman, M.J.; Oliyai, R.; Maag, H.; Tilley, Jw. editors. *Prodrugs: Challenges and rewards*. 2007 Vol. 5. New York, NY:: Springer.
[20] Liu, K.S.; Tzeng, J.I.; Chen, Y.W.; Huang, K.L,.; Kuei, C.H.; Wang, J.J. Novel depots of buprenorphine prodrugs have a long-acting antinociceptive effect. *Anesthesia and analgesia*. **2006**, *102*(5), 1445-51.
[21] Singh, Y.; Palombo, M.; Sinko, P.J. Recent trends in targeted anticancer prodrug and conjugate design. *Current medicinal chemistry*. **2008**, *15*(18), 1802-26.
[22] Yu, L.X.; Straughn, A.B.; Faustion, P.J.; Yang, Y.; Parekh, A.; Ciavarella, A.B.; Asafu-Adjaye, E.; Mehta, M.U.; Conner, D.P.; Lesko, L.J.; Hussain, A.S. The effect of food on the relative bioavailability of rapidly dissolving immediate-release solid oral products containing highly soluble drugs. *Mol. Pharm.* **2004**, *1*, 357-362.
[23] Shan, D.; Nicolaou, M.G.; Borchardt, R.T.; Wang, B. Prodrug strategies based on intramolecular cyclization reactions. *J. Pharm. Sci.* **1997**, *86*, 765-767.
[24] Testa, B.; Mayer, J.M. Design of intramolecularly activated prodrugs. *Drug Metab. Rev.***1998**, *30*, 787-807.
[25] Wang, W.; Jiang, J.; Ballard, C.E.; Wang, B. Prodrug approaches to the improved delivery of peptide drugs. *Curr. Pharm. Des.* **1999**, *5*, 265-287.
[26] Vinšová, J.; Imramovský, A. Intramolekulární cycklizace využívané k uvolňování účinných látek z proléčiv. *Chem. Listy* **2005**, *99*, 21-29.
[27] Papot, S.; Tranoy, I.; Tillequin, F.; Florent, J.-C-; Gesson, J.-P. Design of selectively activated anticancer prodrugs: elimination and cyclization strategies. *Curr. Med. Chem. - Anti-Cancer Agents* **2002**, *2*, 155-185.
[28] Stella, Borchardt, Hageman, Oliyai, Maag and Tilley (eds.). Prodrugs: challenges and rewards.Part 1 and 2*,* Springer-AAPS Press: New York, 2007.
[29] Amsberry, K.L.; Borchardt, R.T. Amine prodrugs which utilize hydroxy amide lactonization. I. A potential redox-sensitive amide prodrug. *Pharm. Res.* **1991**, *8*, 323-330.
[30] Wolfe, J.L.; Vander Velde, D.G.; Borchardt, R.T. Facilitated intramolecular conjugate addition of N-(p- methoxypheny1)-3-(3',6'-dioxo-2',4'-dimethylcyclohexa-1',4'-dienyl)-3,3-dimethylpropionamide. 1. Product characterization. *J. Org. Chem.* **1992**, *57*, 6138-6142.
[31] Nicolaou, M.G.; Wolfe, J.L.; Schowen, R.L.; Borchardt, R.T. Facilitated intramolecular conjugate addition of amides of 3-(3',6'-dioxo-2',4'-dimethyl-1',4'-cyclohexadienyl)-3,3-dimethylpropionic acid. 2. Kinetics of degradation. *J. Org. Chem.* **1996**, *61*, 6633-6638.

[32] Amsberry, K.L.; Gerstenberger, E.; Borchardt, R.T. Amine prodrugs which utilize hydroxy amide lactonization. II. A potential esterase-sensitive amide prodrug. *Pharm. Res.* **1991**, *8*, 455-461.

[33] Wang, B.; Gangwar, S.; Pauletti, G.M.; Siahaan, T.J.; Borchardt, R.T. Synthesis of a novel esterase-sensitive cyclic prodrug system for peptides that utilizes a "trimethyl lock"-facilitated lactonization reaction. *J. Org. Chem.* **1997**, *62*, 1363-1367.

[34] Pauletti, G.M.; Gangwar, S.; Wang, B.; Borchardt, R.T. Esterase-sensitive cyclic prodrugs of peptides: evaluation of a phenylpropionic acid promoiety in a model hexapeptide. *Pharm. Res.* **1997**, *14*, 11-17.

[35] Nicolaou, M.G.; Yuan, C.-S.; Borchardt, R.T. Phosphate prodrugs for amines utilizing a fast intramolecular hydroxy amide lactonization. *J. Org. Chem.* **1996**, *61*, 8636-8641.

[36] Hershfield, R.; Schmir, G.L. Lactonization of coumarinic acids. Kinetic evidence for three species of the tetrahedral intermediate. *J. Am. Chem. Soc.* **1973**, *95*, 8032-8040.

[37] Bundgaard, H.; Falch, E.; Larsen, C.; Mesher, G.L.; Mikkelson, T. Pilocarpic acid esters as novel sequentially labile pilocarpine prodrugs for improved ocular delivery. *J. Med. Chem.* **1985**, *28*, 979-981.

[38] Konschim, H.; Ekholm, M. Molecular modeling of pilocarpine prodrugs: A theoretical investigation of pilocarpic acid esters. *International Journal of Quantum Chemistry* **1991**, *40* (Supp.18), 247-267.

[39] Hanson, K.R.; Havir, E. A. The enzymatic elimination of ammonia. In: Boyer P.D., editor. The Enzymes, 3rd edn. vol. 7. New York: Academic Press, 1972, p 75- 166.

[40] Czarnik, A. W. Intramolecularity: proximity and strain. In: Liebman J.F., Greenberg A., editors. Mechanistic Principles of Enzyme Activity. New York, NY: VCH Publishers, 1988, p. 75-117.

[41] Sweigers, G. F. Mechanical Catalysis. Hoboken, NJ: John Wiley & Sons, 2008.

[42] Fersht, A. Structure and Mechanism in Protein Science: A guide to Enzyme Catalysis and Protein Folding. New York: Freeman, W. H. and Company, 1979.

[43] Nelson, D.L.; Cox, M.M. Lehninger Principles of Biochemistry. New York: Worth Publishers, 2003.

[44] Dafforn, A.; Koshland, D.E. Jr (1973) Proximity, entropy and orbital steering. *Biochem. Biophys. Res. Commun*. **1973**, *52*, 779-785.

[45] Lightstone, F. C.; Bruice, T. C. Separation of ground state and transition state effects in intramolecular and enzymatic reactions. 2. A theoretical study of the formation of transition states in cyclic anhydride formation. *J. Am. Chem. Soc*.**1997**, *119*, 9103-9113.

[46] Bruice, T. C.; Pandit, U. K. The effect of geminal substitution ring size and rotamer distribution on the intra molecular nucleophilic catalysis of the hydrolysis of monophenyl esters of dibasic acids and the solvolysis of the intermediate anhydrides. *J. Am. Chem. Soc*. **1960**, *82*, 5858-5865.

[47] Bruice, T. C.; Pandit, U. K. Intramolecular models depicting the kinetic importance of ''Fit'' in enzymatic catalysis. *Proc. Natl. Acad. Sci. USA* **1960**, *46*, 402-404.

[48] Milstein, S.; Cohen, L.A. Concurrent general-acid and general-base catalysis of esterification. *J. Am. Chem. Soc*.**1970**, *92*, 4377-4382.

[49] Milstein, S.; Cohen, L. A. Rate acceleration by stereopopulation control: models for enzyme action. *Proc. Natl. Acad. Sci. USA* **1970**, *67*, 1143-1147.

[50] Milstein, S.; Cohen, L. A. Stereopopulation control I. Rate enhancement in the lactonizations of o-hydroxyhydrocinnamic acids. *J. Am. Chem. Soc*.**1972**, *94*, 9158-9165.

[51] Menger, F. M.; Ladika, M. Fast hydrolysis of an aliphatic amide at neutral pH and ambient temperature. A peptidase model. *J. Am. Chem. Soc*.**1988**, *110*, 6794-6796.

[52] Menger, F. M. On the source of intramolecular and enzymatic reactivity. *Acc. Chem. Res*. **1985**, *18*, 128-134.

[53] Menger, F. M.; Chow, J. F.; Kaiserman H.; Vasquez P. C. Directionality of proton transfer in solution. Three systems of known angularity. *J. Am. Chem. Soc*.**1983**, *105*, 4996-5002.

[54] Menger, F. M.; Galloway, A. L.; Musaev D. G. Relationship between rate and distance. *Chem. Commun*. **2003**, 2370-2371.

[55] Menger, F. M. An alternative view of enzyme catalysis. *Pure Appl. Chem*. **2005**, *77*, 1873-1876.

[56] Kirby, A. J.; Hollfelder, F. From Enzyme Models to Model Enzymes, RSC Publishing, Cambridge UK, 2009, pp. 1-273,

[57] Barber, S. E.; Dean, K. E. S.; Kirby, A. J. A mechanism for efficient proton-transfer catalysis. Intramolecular general acid catalysis of the hydrolysis of 1-arylethyl ethers of salicylic acid. *Can. J. Chem.* **1999**, *77*, 792-801.

[58] Kirby, A. J.; Lancaster, P. W. Structure and efficiency in intramolecular and enzymatic catalysis. Catalysis of amide hydrolysis by the carboxy-group of substituted maleamic acids. *J. Chem. Soc., Perkin Trans. 2*, **1972**, 1206-1214.

[59] Kirby, A. J.; de Silva, M. F.; Lima, D.; Roussev, C. D.; Nome, F. Efficient intramolecular general acid catalysis of nucleophilic attack on a phosphodiester. *J. Am. Chem. Soc*. **2006**, *128*, 16944-16952.

[60] Kirby, A. J.; Williams, N. H. Efficient intramolecular general acid catalysis of enol ether hydrolysis. Hydrogen-bonding stabilization of the transition state for proton transfer to carbon. *J. Chem. Soc., Perkin Trans. 2*, **1994**, 643-648.

[61] Kirby, A. J.; Williams N. H. Efficient intramolecular general acid catalysis of vinyl ether hydrolysis by the neighbouring carboxylic acid group. *J. Chem. Soc. Chem. Commun* **1991**, 1643-1644.

[62] Kirby, A. J. Enzyme Mechanisms, Models, and Mimics. *Angewandte Chemie International Edition in English* **1996**, *35*, 706-724.

[63] Fife, T. H.; Przystas, T. J. Intramolecular general acid catalysis in the hydrolysis of acetals with aliphatic alcohol leaving groups. *J. Am. Chem. Soc.* **1979**, *101*, 1202-1210.

[64] Kirby, A. J. Efficiency of proton transfer catalysis in models and enzymes. *Acc. Chem. Res.* **1997**, *30*, 290-296.

[65] Reddy, M. R.; Erion, M. D. Free Energy Calculations in Rational Drug Design, Kluwer Academic/Plenum Publishers, **2001**.

[66] Parr, R. G.; Craig, D. P.; Ross I. G. Molecular Orbital Calculations of the Lower Excited Electronic Levels of Benzene, Configuration Interaction included. *Journal of Chemical Physics* **1950**, *18* (12), 1561-1563.

[67] Parr, R. G. On the genesis of a theory. *Int. J. Quantum Chem* **1990**, *37* (4), 327-347.

[68] Dewar, M. J. S.; Thiel W. Ground states of molecules. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc*. **1977**, *99*, 4899-4907.

[69] Bingham, R. C.; Dewar, M. J. S.; Lo, D. H. Ground states of molecules. XXV. MINDO/3. An improved version of the MNDO semiempirical SCF-MO method. *J. Am. Chem. Soc*. **1975**, *97*, 1285-1293.

[70] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc*. **1985**, *107*, 3902-907.

[71] Dewar, M. J. S.; Jie, C.; Yu, J. The first of new series of general purpose quantum mechanical molecular models. *Tetrahedron* **1993**, *49*, 5003-5038.

[72] Parr, R. G.; Yang W. Density Functional Theory of Atoms and Molecules. Oxford University Press, Oxford, 1989.

[73] Mourik, T. van; Robert J.G. A critical note on density functional theory studies on rare-gas dimers. *Journal of Chemical Physics* **2002**, *116* (22), 9620-9623.

[74] Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* **2004**, *25* (12), 1463-1473.

[75] Burker, U.; Allinger, N. L. Molecular Mechanics, American Chemical Society, Washington, DC, USA, **1982**.

[76] Warshel, A.; Levitt, M. Theoretical studies of enzymatic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology* **1976**, *103* (2), 227-249.

[77] Field, M. J. Simulating enzyme reactions: challenges and perspectives. *Journal of Computational Chemistry* **2002**, *23*(1), 48-58.

[78] Mulholland, A. J. Modeling enzyme reaction mechanisms, specificity and catalysis. *Drug Discovery Today* **2005**, *10*(20), 1393-1402.

[79] Gaussian 09, Revision **A.1**, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2009.

[80] Casewit, C. J.; Colwell, K. S.; Rappe', A. K. Application of a universal force field to main group compounds. *J. Am. Chem. Soc*.**1992**, *114*, 10046-53.

[81] Murrell, J. N.; Laidler, K. J. Symmetries of activated complexes. *Trans. Faraday Soc*.**1968**, *64*, 371-377.

[82] Muller, K. Reaction paths on multidimensional energy hypersurfaces. *Angew Chem. Int. Ed. Engl.***1980**, *19*, 1-13.

[83] Cancès, M. T.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: theoretical background and applications to isotropic and anisotropic dielectrics. *J. Chem. Phys*.**1997**, *107*, 3032-3041.

[84] Mennucci, B.; Tomasi, J. Coninuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. *J. Chem. Phys*.**1997**, *106*, 5151.

[85] Mennucci, B.; Cancès, M. T.; Tomasi, J. Evaluation of solvent effects in isotropic and anisotropic dielectrics and in ionic solutions with a unified integral equation method: Theoretical bases, computational implementation, and numerical applications. *J. Phys.Chem. B* **1997**, *101*, 10506-10517.

[86] Tomasi, J.; Mennucci B.; Cancès, M. T. The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. *J. Mol.Struct. (Theochem)* **1997**, *464*, 211-226.

[87] Karaman, R. Analysis of Menger's spatiotemporal hypothesis. *Tetrahedron Lett*. **2008**, *49*, 5998-6002.

[88] Karaman, R. Cleavage of Menger's aliphatic amide: a model for peptidase enzyme solely explained by proximity orientation in intramolecular proton transfer. *J. Mol. Struct. (THEOCHEM)* **2009**, *910*, 27-33.

[89] Karaman, R. The efficiency of proton transfer in Kirby's enzyme model, a computational approach. *Tetrahedron Lett*. **2010**, *51*, 2130-2135.

[90] Karaman, R.; Pascal, R. A. Computational Analysis of Intramolecularity in Proton Transfer Reactions. *Org. & Bimol. Chem*. **2010**, *8*, 5174-5178.

[91] Karaman, R. A General Equation Correlating Intramolecular Rates with Attack" Parameters: Distance and Angle. *Tetrahedron Lett*. **2010**, *51,* 5185-5190.

[92] Karaman, R. Analyzing the efficiency of proton transfer to carbon in Kirby's enzyme model- a computational approach. *Tetrahedron Lett*. **2011**, *52*, 699-704.

[93] Karaman, R. Analyzing the efficiency in intramolecular amide hydrolysis of Kirby's N-alkylmaleamic acids - A computational approach. *Comput Theor Chem* **2011**, *974*, 133-142.

[94] Karaman, R. A New Mathematical Equation Relating Activation Energy to Bond Angle and Distance: A Key for Understanding the Role of Acceleration in Lactonization of the Trimethyl Lock System. *Bioorganic Chemistry* **2009**, *37*, 11-25.

[95] Karaman, R. Revaluation of Bruice's Proximity Orientation. *Tetrahedron Lett*. **2009**, *50*, 452-458.

[96] Karaman, R. Accelerations in the Lactonization of Trimethyl Lock Systems are Due to Proximity Orientation and not to Strain Effects. *Research Letters in Org. Chem.* **2009***,* doi: 10.1155/2009/240253.

[97] Karaman, R. The *gem*-disubstituent effect- a computational study that exposes the relevance of existing theoretical models. Tetrahedron Lett. **2009**, *50*, 6083-6087.

[98] Karaman, R. Analyzing Kirby's amine olefin - a model for amino-acid ammonia lyases. *Tetrahedron Lett*. **2009**, *50*, 7304-7309.

[99] Karaman, R. The Effective Molarity (EM) Puzzle in Proton Transfer Reactions. *Bioorganic Chemistry* **2009**, *37*,106-110.

[100] Karaman, R. Effects of substitution on the effective molarity (EM) for five membered ring-closure reactions- a computational approach. *Journal of Molecular Structure (THEOCHEM)* **2010***, 939*, 69-74.

[101] Karaman, R. The Effective Molarity (EM) Puzzle in Intramolecular Ring-Closing Reactions. *Journal of Molecular Structure (THEOCHEM)* **2010**, *940*, 70-75.

[102] Menger, F. M.; Karaman, R. A Singularity Model for ChemicalReactivity. *Eur. J. Chem.* **2010**, *16*, 1420-1427.

[103] Karaman, R. The Effective Molarity (EM) - a Computational Approach. *Bioorganic Chemistry* **2010**, *38*,165-172.

[104] Karaman, R. Proximity *vs*. Strain in Ring-Closing Reactions of Bifunctional Chain Molecules- a Computational Approach. *J. Mol. Phys*. **2010**, *108*, 1723-1730.

[105] Karaman, R. The role of proximity orientation in intramolecular proton transfer reactions. *Journal of Computational and Theoretical Chemistry* **2011**, *966*, 311-321.

[106] Houk, K.N.; Tucker, J.A.; Dorigo, A.E. Quantitative modeling of proximity effects on organic reactivity. *Acc Chem Res* **1990**, *23*, 107-113.
[107] Hitani, A.; Nakamura, T.; Ohtomo, H.; Nawa, Y.; Kimura, M. Efficacy and safety of atovaquone-proguanil compared with mefloquine in the treatment of nonimmune patients with uncomplicated P. falciparum malaria in Japan. Journal of infection and chemotherapy: official journal of the Japan Society of Chemotherapy **2006**, *12* (5), 277-82.
[108] www.cdc.gov/malaria/about/disease.htm
[109] Nayyar, G.M.L.; Breman, J.G.; Newton, P. N.; Herrington, J. (2012). Poor-quality antimalarial drugs in Southeast Asia and sub-Saharan Africa. *Lancet Infectious Diseases* **2012**, *12* (6), 488-496.
[110] Chung, M. C.; Ferreira, E. I.; Santos, J. L.; Giarolla, J.; Rando, D. G.; Almeida, A. E.; Bosquesi, P. L.; Menegon, R. F. & Blau, L. Prodrugs for the treatment of neglected diseases. *Molecules* **2008**, *13* (3), 616-77.
[111] Peterson, A. T. Shifting suitability for malaria vectors across Africa with warming climates. BMC infectious diseases **2009**, *9*, 59.
[112] Karaman, R. & Hallak, H. Computer-assisted design of pro-drugs for antimalarial atovaquone. *Chemical biology & drug design* **2010**, *76* (4), 350-360.
[113] Reinaldo T. Antimalarial drug discovery: old and new approaches. *J Exp Biol* **2003**, *206*, 3735-3744.
[114] Comley, J.; Yeates, C.; Frend, T. Antipneumocystis activity of 17C91, a prodrug of atovaquone. *Antimicrob Agents Chemother*. **1995**, *39*, 2217-2219.
[115] Karaman R. Antimalarial Atovaquone Prodrugs Based on Enzyme Models - Molecular Orbital Calculations Approach.in Antimalarial Drug Research and Development, Nova Cooperation Publisher, NY, USA, 2013 pp 1-67.
[116] Kirby, A. J. effective molarities for intramolecular reactions. *J. Phys. Org. Chem*. **2005**, *18*, 101-278.
[117] Sohi, H.; Sultana, Y.; Khar R.K. Taste Masking Technologies in oral pharmaceuticals, recent development and approaches. *Drug Develop Ind. Pharm*. **2004**, *30*, 429-448.
[118] Reilly, W.J. Pharmaceutical Necessities in Remington: The Science and Practice of Pharmacy. Balti- more, MD: Mack Publishing Company, 2002, p. 1018-1020.
[119] Drewnowski, A., Gomez-Carneros C. Bitter taste, phytonutrients, and the consumer: a review. *Am. J. Clin. Nutr*. **2000**, *72*, 1424-1435.
[120] Hofmann, T. Identification of the key bitter compounds in our daily diet is a prerequisite for the understanding of the hTAS2R gene polymorphisms affecting food choice. *Ann. N. Y. Acad. Sci*. **2009**, *1170*, 116-125.
[121] Rodgers, S.; Busch, J.; Peters, H.; Christ-Hazelhof E. Building a tree of knowledge: analysis of bitter molecules. *Chem. Senses*. **2005**, *30*, 547-557.
[122] Rodgers, S.; Glen, R.C.; Bender, A. Characterizing bitterness: identification of key structural features and development of a classification model. *J. Chem.Inf. Model*. **2006**, *46*, 569-576.
[123] Maehashi, K.; Huang, L. Bitter peptides and bitter taste receptors. *Cell Mol Life Sci*. **2009**, *66*, 1661-1671.
[124] Behrens, M.; Meyerhof, W. Bitter taste receptors and human bitter taste perception. *Cell. Mol. Life Sci.* **2006**, *63*, 1501-1509.
[125] Meyerhof, W.; Born, S.; Brockhoff, A.; Behrens, M. (2011) Molecular biology of mammalian bitter taste receptors. *A review. Flavour Frag. J*. **2011**, *26*, 260-268.

[126] Behrens, M.; Meyerhof, W. Mammalian bitter taste perception. *Results Probl. Cell. Differ.* **2009**, *47*, 203-220.
[127] Ayenew, Z.; Puri, V.; Kumar, L.; Bansal, A.K. Trends in pharmaceutical taste masking technologies: a patent review. *Recent Pat Drug Deliv Formul*. **2009**, *3*, 26-39.
[128] Fawzy, A. A. Pleasant Tasting Aqueous Liquid Composition of a Bitter-Tasting Drug*, PCT Int. Appl*. **1998**, WO9805312, 2.
[129] Gowan, W. G. Aliphatic Esters as Solventless Coating pharmaceuticals", *Can. Pat. Appl*. **1993**, CA2082137, 11.
[130] Gowthamarajan, K.; Kulkarni, G. T.; Kumar, M. N. Pop the pills without bitterness Taste-Masking technologies for bitter drugs. *resonance*. **2004**, 25.
[131] Bakan, J. A. Microencapsulation, Theory and practice of Industrial Pharmacy, Third Edition, 1986, pp 412-429.
[132] Iyer, V. S. Srinivas SC (2007) WO2007060682.
[133] Bush, L. Bitter taste bypass need for sugar spoon. *Pharm Technol*. **2004**.http://pharmtech.findpharma.com/pharmtech/data/articlest and ard//pharmtech/072004/84521/article.pdf.
[134] www.pharmainfo.net. Ion exchange resin complex: an approach to mask the taste of bitter drugs.
[135] Bress, W.S.; Kulkarni, N.; Ambike, S.; Ramsay, M. P. **2006**, EP1674078.
[136] Mendes, W. R. Theory and practice of Industrial pharmacy, Third Edition, 1976.
[137] Redondo, A. M. J.; Abanades, L. B. **2003**, WO047550.
[138] Scotti, L.; Scotti, M. T.; Ishiki, H. M.; Ferreira, M. G. P. Emerenciano VP, Menezes CMS, Ferreira EI, Quantitative elucidation of the structure-bittereness relationship of cynaropicrin and grosheimin derivatives. *Food Chem*. **2007**, *105*, 77-83.
[139] Hejaz, H.; Karaman R.; Khamis M. Computer- Assisted Design for paracetamol Masking Bitter Taste Prodrugs. *J. Mol. Model*. **2012**, *18*, 103-114.
[140] Karaman, R.; Ghareeb, H.; Dajani, K.K.; Scrano, L.; Hallak, H.; Abu-Lafi, S.; Bufo, S. A. Design, synthesis and in-vitro kinetic study of tranexamic acid prodrugs for the treatment of bleeding conditions. *J. Molecular Aided Computer Design*. **2013**, *27*(7), 615-635.
[141] Karaman, R.; Dajani, K.K.; Qtait, A.; Khamis, M. Prodrugs of Acyclovir - A Computational Approach. *Chem. Biol. Drug Des*. **2012**, *79*, 819-834.
[142] Karaman, R.; Dajani, K. K., Hallak, H. Computer-assisted design for atenolol prodrugs for the use in aqueous formulations. *J. Mol. Model*. **2012**, *18*, 1523-1540.
[143] Karaman, R. Prodrugs for Masking Bitter Taste of Antibacterial Drugs- A Computational Approach. *J. Mol. Model*. **2013**, *19*, 2399-2412.
[144] Karaman, R. Computationally Designed Enzyme Models to Replace Natural Enzymes In Prodrug Approaches. *Journal of Drug Designing*. **2012**, 1:e111doi:10.4172/2169-0138.1000e111.
[145] Karaman, R. Prodrug Design *vs*. Drug Design. *Journal of Drug Designing*. **2013**, 2: e114. doi:10.4172/2169-0138.1000e114
[146] Kluger R.; Chin J. Carboxylic acid participation in amide hydrolysis. Evidence that separation of a nonbonded complex can be rate determining. *J. Am. Chem. Soc*. **1982**, *104*, 2891-2897.
[147] Katagi, T. AM1 study of acid-catalyzed hydrolysis of maleamic (4-amino-4-oxo-2-butenoic) acids. *J. Comp. Chem*. **1990**, *11*(9), 1094-1100.
[148] CRASH-2 Trial Collaborators Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant hemorrhage (CRASH-2): a randomized, placebo controlled trial. *Lancet* **2010**, *6736*, 60835-5.

[149] Gohel, M.; Patel, P.; Gupta, A.; Desai, P. Efficacy of tranexamic acid in decreasing blood loss during and after cesarean section: a randomized case controlled prospective study. *J. Obstet. Gynecol. India.* **2007**, *57*, 227-230.

[150] Giancarlo, L.; Francesco, B.; Angela, L.; Pierluigi, P.; Gina, R. Recommendations for the transfusion management of patients in the peri-operative period. II. *The intra-operative period. Blood Transfus*. **2011**, *9*, 189-217.

[151] Lukes, A.S.; Kouides, P.A.; Moore, K.A. Tranexamic acid: a novel oral formulation for the treatment of heavy menstrual bleeding. *Women's Health*. **2011**, *7*, 151-158.

[152] Lukes A.S., Moore K.A., Muse K.N., Gersten J.K., Hecht B.R., Edlund M Richter H.E., Eder S.E., Attia G.R., Patrick D.L., Rubin A., Shangold G.A. Tranexamic acid treatment for heavy menstrual bleeding: a randomized controlled trial. *Obstet Gynecol*. **2010**, *116*, 865-875.

[153] Pilbrant, A.; Schannong, M.; Vessman, J. Pharmacokinetics and bioavailability of tranexamic acid. Eur. *J. Clin. Pharmacol*. **1981**, *20*, 65-72.

[154] McEvoy, G. K. *AHFS Drug Information-93*," American Society of Health-System Pharmacists 1993, pp. 916-920.

[155] http://homepage.ntlworld.com/bhandari/Imperial/Atenolol/Synthesis.htm

[156] Melander, A.; Stenberg, P.; Liedholm H. *et al*. Food- induced reduction in bioavailability of atenolol. *Eur J Clin Pharmacol,* **1979**, *16*, 327-30.

[157] J. Mcainsh, WT. Simpson, BF. Holmes *et al*. "Bioavailability of atenolol formulations. *Biopharm Drug Dispos* **1980**, *1*, 223-232.

[158] Standing, F. J.; Tuleu, C. Pediatric formulations-getting to the heart of the problem. *Int. J. Pharm.* **2005**, *300*, 56-66.

[159] Garner, S.; West, D.; Reynolds, E. Stability of atenolol in an extemporaneously compounded oral liquid. *American Journal of Hospital Pharmacy* **1994***, 51*, 508-511.

[160] Foppa, T.; Murakami, F. S.; Silva, M. A. Development, validation and stability study of pediatric atenolol syrup. *Pharmazie* **2007***, 62*, 519-521.

[161] Patel, D.; Doshi, D. H.; Desia, A. Short term stability of atenolol in oral liquid formulations. *Int. J. Pharmaceut. Compd*. **1997**, 437-439.

[162] Chan, K.; Swenden, J. Pilot study of the short-term physico-chemical stability of atenolol tablets stored in a multi-compartment compliance aid. *Eur. J. Hosp. Pharm. Sci*. **2007**, *13*, 60-66.

[163] Anroop, B.; Ghosh, B.; Parcha, V.; Khanam, J. Transdermal delivery of atenolol: effect of prodrugs and iontophoresis. *Curr Drug Delivery* **2009***, 6*, 280-290.

[164] Sohi, H.; Sultana, Y.; Khar, R. K. Taste Masking Technologies in oral pharmaceuticals, recent development and approaches. Drug Develop. *Ind. Pharm*. **2004**, *30*(5), 429-448.

[165] Garner, S.; West, D.; Reynolds, E. Stability of atenolol in an extemporaneously compounded oral liquid. *American Journal of Hospital Pharmacy* **1994**, *51*, 508-511.

[166] Shin,SC.;Choi, JS. Enhanced bioavailability of atenolol by transdermal administration of the ethylene-vinyl acetate matrix in rabbits. *Eur J Pharm Biopharm* **2003**, *56*, 439-443.

[167] Montes-Gil, Ana C., *et al*. "Pharmacokinetic profile of atenolol aspirinate. *Archiv der Pharmazie* **2007**, *340*, 445-455.

[168] Stella, V. J. Prodrugs as therapeutics. *Expert Opin Ther Patents* **2004**, *14*, 277-280.

[169] Neu HC., Winshell EB. Pharmacological studies of 6 (D(-) -amino-p-hydroxyphenylacetamido) penicillanic acid in humans. *Antimicrob Agents Chemother* **1970**, *10*, 423-426.

[170] Parr Jr, TR.; Bryan, L. E. Non-enzymatic resistance to B-lactam antibiotics and resistance to other cell wall synthesis inhibitions. Antimicrobial drug resistance. Academic Press, Inc., New York; 1984, pp 81-111.

[171] Waxman, DJ.; Strominger, J. L. Penicillin-binding proteins and the mechanism of action of beta-lactam antibiotics. *Annu Rev Biochem* **1983**, *52*, 825-69.

[172] Brogeden, R. N.; Speight, T. M.; Avery, G. S. Amoxycillin: A review of its antibacterial and pharmacokinietic properties and therapeutic use. *Drugs* **1975**, *9*(2), 88-140.

[173] Davis, J. L.; Salmon, J. H.; Papich, M.G. Pharmacokinetics and tissue fluid distribution of cephalexin in the horse after oral and i.v. administration. *J Vet Pharmacol Ther* **2005**, *28*(5), 425-431.

[174] Tack, K. J.; Keyserling, C. H.; McCarty, J.; Hedrick, J. A. Study of use of cefdinir *versus* cephalexin for treatment of skin infections in pediatric patients. The Cefdinir Pediatric Skin Infection Study Group. *Antimicrobial agents and chemotherapy* **1997**, *41*(4), 739-742.

[175] Jacobs, M. R.; Jones, R. N.; Giordano, P. A. Oral β-lactams applied to uncomplicated infections of skin and skin structures. *Diagnostic microbiology and infectious disease* **2007**, *57*(3), S55-S65.

[176] Disney, F. A.; Dillon, H.; Blumer, JL.; Dudding, B. A.; McLinn, S. E.; Nelson, D. B.; Selbst, S. M. Cephalexin and Penicillin in the Treatment of Group A {beta}-Hemolytic Streptococcal Throat Infections. *Archives of Pediatrics & Adolescent Medicine* **1992**, *146*(11), 1324.

[177] Perry, CM.; Brogden, R. N. Cefuroxime axetil. A review of its antibacterial activity, pharmacokinetic properties and therapeutic efficacy. *Drugs*. **1996**, *52*(1), 125-58.

[178] Kees, F. K.; Lukassek, U.; Naber, K. G.; Grobecker, H. Comparative investigations on the bioavailability of cefuroxime axetil. *Arzneimittel-Forschung* **1991**, *41*(8), 843-846.

[179] Ehrnebo, M.; Nilsson, S. O.; Boréus, L. O. Pharmacokinetics of ampicillin and its prodrugs bacampicillin and pivampicillin in man. *Journal of pharmacokinetics and Biopharmaceutics* **1979**, *7*(5), 429-451.

[180] Jones, K. H. Bioavailability of talampicillin. *Br. Med. J*. **1977**, *2*, 232-233.

[181] Karaman, R.; Amly, W.; Scrano, L.; Mecca, G.; Bufo, S. A. Computationally designed prodrugs of statins based on Kirby's enzyme model. *Journal of molecular modeling* **2013**, *19*(9), 3969-3982.

[182] Karaman, R. Computational aided design for dopamine prodrugs based on novel chemical approach. *Chem. Biol. Drug Design* **2011**, *78*, 853-863.

[183] Karaman, R.; Dokmak, G.; Bader, M.; Hallak, H.; Khamis, M.; Scrano, L.; Bufo, S. A. Prodrugs of fumarate esters for the treatment of psoriasis and multiple sclerosis(MS)- A computational approach. *J. Molecular Modeling* **2013**, *19*, 439-452.

[184] Karaman, R. Prodrugs of Aza Nucleosides Based on Proton Transfer Reactions. *J. Comput. Mol. Des*. **2010**, *24*, 961-970.

[185] Karaman, R.; Karaman, D.; Ziadeh, I. Computationally Designed Phenylephrine Prodrugs-A Model for Enhancing Bioavailability. *J. Molecular Physics* **2013**, ***111*(21), 3249-3264.**

CHAPTER 6

Structural and Vibrational Investigation on a Benzoxazin Derivative with Potential Antibacterial Activity

María V. Castillo, Elida Romano, Ana B. Raschi and Silvia A. Brandán*

General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán, Tucumán, Argentina

Abstract: In this chapter, the structural and vibrational properties of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one were studied by using the available experimental infrared spectrum and the hybrid B3LYP/6-31G*and B3LYP/6-311++G** methods. The bonds order, charge-transfers and stabilization energies for the compound were calculated employing the Natural Bond Orbital (NBO) analysis while the topological properties at the same levels of theory were calculated using the Atoms in Molecules theory (AIM). Furthermore, the frontier molecular HOMO and LUMO orbitals for the compound were also computed and later the values were compared with those reported for 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one and 2-phenyl-4*H*-3,1-benzoxazin-4-one. On the other hand, the harmonic vibrational frequencies at the same levels of theory were calculated using the optimized geometries of the compound. Then, the Pulay´s scaled quantum mechanical force field (SQMFF) methodology was used together with the corresponding normal internal coordinates in order to perform the complete assignment of the vibrational spectra. In addition, the scaled force constants were also presented together with the force fields by using both levels of approximation. In this chapter, the Raman spectrum for the compound at the B3LYP/6-31G* level of theory was predicted.

Keywords: 1-benzoxazin-4-one, 2-(4-methylphenyl)-4H-3, DFT calculations, force field, molecular structure, vibrational spectra.

INTRODUCTION

Recently, a series of twenty derivatives of 2-aryl-4*H*-3,1-benzoxazin-4-one have been synthesized and their potential therapeutically significance tested against two

***Corresponding author Silvia A. Brandán:** General Chemistry, Inorganic Chemistry Institute, Faculty of Biochemistry, Chemistry and Pharmacy, National of Tucuman University, Ayacucho 471, (4000)., San Miguel de Tucumán, Tucumán, Argentina; Tel: +54-381-4247752; Fax: +54-381-4248169;
E-mail: sbrandan@fbqf.unt.edu.ar

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

strains of Gram positive bacteria and four strains of Gram negative bacteria have been reported by Khan *et al*. [1]. The study has revealed that 2-aryl-4*H*-3,1-benzoxazin-4-ones possess good bactericidal activity against a panel of bacteria causing common bacterial diseases and therefore opens the possibility of finding latest clinically useful antibacterial compounds [1]. Many previous studies have also demonstrated that the benzoxazinones derivatives are promising as antifungal and antibacterial agents [2-5] and, moreover, they are broadly used directly or indirectly in numerous applications [6-13], such as, intermediates in organic synthesis to obtain 4(3*H*)quinazolinones [6-10], compounds with interesting biological activities (antifungal, antibacterial, antiviral, antitubercular and anticancer) [11-13]. Industrially, the syntheses of poly-benzoxazines are of importance in the electronics and aerospace industries because they constitute thermosetting resins with remarkable properties [13]. In this context, the study of the structural and spectroscopic properties of these (4*H*)-3,1-benzoxazinones derivatives are of great importance to know the relation existent between structure and reactivity and to identify these species by means of vibrational spectroscopy. For the 2-phenyl-4H-3,1-benzoxazin-4-one derivative, vibrational spectroscopic studies and theoretical calculations were reported by Ambujakshan *et al*. [14] while the structures and vibrational properties of the 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one derivative were recently reported by Castillo et. al. [15] by using theoretical calculations based on the density functional theory (DFT). In this chapter, as part of our studies on compounds of pharmacological and industrial interest [16-32] and, with the objective of calculating the structural and vibrational properties of other 4*H*-3,1-benzoxazinones derivative, we considered here the study of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one. The preparation of this compound was previously reported by Rai [9] who has characterized it compound by means of thin layer chromatography (TLC) and IR, ^{1}H-NMR and Mass spectra. The crystal and molecular structure of this compound is unknown and, so far, there is no theoretical studies related with the geometry and assignments of the infrared and Raman spectra. The goals of this work are: (i) determine the best calculation level that reproduce the structure and the available experimental infrared spectrum of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one, (ii) determine the more stable structure of the compound and to calculate their structural properties and, (iii) perform the complete assignments of all the bands

observed in the vibrational spectra. For these purposes, the optimizations of the structures and the calculations of the corresponding frequencies were performed by using the B3LYP/6-31G* and B3LYP/6-311++G** levels of approximations. In this work, two basis sets were used in order to observe the influence of the size of basis sets on the structural and vibrational properties. For this derivative, the molecular force field was calculated at the B3LYP/6-31G* level using the normal internal coordinates and the generalized valence force field (GVFF) with the SQM methodology [33]. The study of structural parameters such as the bonds order, charge-transfers and stabilization energies are important in this compound due to the presence in their structure of the methyl phenyl and benzoxazin rings. Hence, these properties were obtained by using the Natural Bond Orbital (NBO) calculations [34, 35] and both basis sets while the topological properties were calculated by means of the Atoms in Molecules (AIM) [36, 37] theory. On the other hand, the prediction of the reactivity in this compound is of interest specially if it is used as a pharmacologic drug and, for this reason, the energy gap values were also calculated at the same approximation levels by means of the frontier highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) orbitals. Finally, a comparison on the calculated structural and vibrational properties for this derivative with those corresponding to 2-(4-chlorophenyl) -4H-3,1-benzoxazin-4-one [15] was presented.

COMPUTATIONAL DETAILS

Here, all the studied properties were calculated by using the optimized geometries for the stable structure of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one (MPB) employing the hybrid B3LYP/6-31G*and B3LYP/6-311++G** methods. The most stable MPB structure with C_1 symmetry by using both levels of approximations can be seen in Fig. **1** together with the atoms labelling. The topological properties for MPB were calculated employing the AIM2000 program [37] while the natural charges, bond orders and stabilization energies were computed with the NBO 3.1 program [35], as specified by the Gaussian 03 program [38]. The structural formulas corresponding to MPB, 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one (CPB) and 2-phenyl-4H-3,1-benzoxazin-4-one (PB) are $C_{15}H_{11}NO_2$, $C_{14}H_8NO_2Cl$ and $C_{14}H_9NO_2$, respectively.

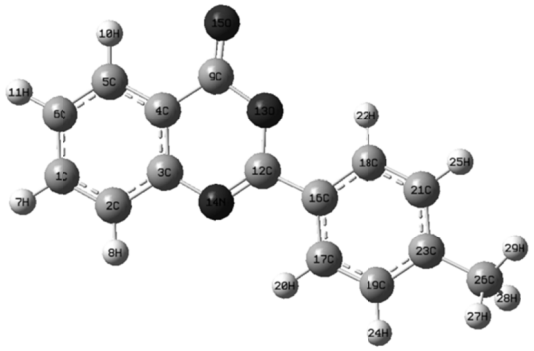


Figure 1: Theoretical structure and atoms numbering of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one.

The calculated force fields in Cartesian coordinates were transformed to natural internal coordinates using the Molvib program [39]. These natural internal coordinates for MPB are given in Table **1** and they were defined in agree with similar molecules [15-20]**.**

Table 1: Definition of Natural internal coordinates for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one

S_1 = r (9-15)	ν C9-O15
S_2 = r(9-13)	ν C9-O13
S_3= r (12-13)	ν C12-O13
S_4 = r (12-14)	ν C12-N14
S_5 = r (3-14)	ν C3-N14
S_6 = r (23-26)	ν C23-C26
S_7 = r (4-9)	ν C4-C9
S_8 = r (3-4)	ν C3-C4
S_9 = r (2-3)	ν C2-C3
S_{10} = r (1-2)	ν C1-C2
S_{11} = r (6-1)	ν C6-C1
S_{12} = r (5-6)	ν C5-C6
S_{13} = r (4-5)	ν C4-C5
S_{14} = r (12-16)	ν C12-C16
S_{15} = r (16-17)	ν C16-C17
S_{16} = r (16-18)	ν C16-C18
S_{17} = r (17-19)	ν C17-C19
S_{18} = r (19-23)	ν C19-C23

Table 1: contd…

S_{19} = r (18-21)	ν C18-C21
S_{20} = r (21-23)	ν C21-C23
S_{21} = r (1-7)	ν C-H
S_{22} = r (2-8)	ν C-H
S_{23} = r (5-10)	ν C-H
S_{24} = r (6-11)	ν C-H
S_{25} = r (17-20)	ν C-H
S_{26} = r (18-22)	ν C-H
S_{27} = r (19-24)	ν C-H
S_{28} = r (21-25)	ν C-H
S_{29} = 2 α (C26-H28) - α (C26-H27) - α (C26-H29)	νa CH_3
S_{30} = α (C26-H27) - α (C26-H29)	νa CH_3
S_{31} = α (C26-H28) + α (C26-H27) + α (C26-H29)	νs CH_3
S_{32} = α (H7-C1-C2) - α (H7-C1-C6)	β (C-H)
S_{33} = α (H8-C2-C3) - α (H8-C2-C1)	β (C-H)
S_{34} = α (H10-C5-C6) - α (H10-C5-C4)	β (C-H)
S_{35}= α (H11-C6-C1) - α (H11-C6-C5)	β (C-H)
S_{36} = α (H20-C17-C19) - α (H20-C17-C16)	β (C-H)
S_{37} = α (H22-C18-C16) - α (H22-C18-C21)	β (C-H)
S_{38} = α (H25-C21-C18) - α (H25-C21-C23)	β (C-H)
S_{39} = α (H24-C19-C17) - α (H24-C19-C23)	β (C-H)
S_{40} = α (O15-C9-C4) - α (O15-C9-O13)	β (C9-O15)
S_{41} = γ (H7-C1-C2-C6)	γ C1-H7
S_{42} = γ (H8-C2-C3-C1)	γ C2-H8
S_{43} =γ (H11-C6-C1-C5)	γ C6-H11
S_{44} = γ (H10-C5-C6-C4)	γ C5-H10
S_{45}= γ (H20-C17-C19-C16)	γ C17-H20
S_{46} = γ (H24-C19-C23-C17)	γ C19-H24
S_{47}= γ (H25-C21-C18-C23)	γ C21-H25
S_{48} = γ (H22-C18-C16-C21)	γ C18-H22
S_{49} = γ (O15-C9-C4-O13)	γ C9-O15
$S_{50} = 6^{-1/2}$ [α (C6-C5-C4) + α (C4-C3-C2) + α (C2-C1-C6) - α (C5-C4-C3) - α (C3-C2-C1) - α (C1-C6-C5)]	βR_1 (A1)
$S_{51} = 12^{-1/2}$ [2α (C6-C5-C4) - α (C5-C4-C3) - α (C4-C3-C2) + 2α (C3-C2-C1) - α (C2-C1-C6) - α (C1-C6-C5)]	βR_2 (A1)
S_{52} = ½ [α (C5-C4-C3)- α (C4-C3-C2)+ α (C2-C1-C6)- α (C1-C6-C5)]	βR_3 (A1)

Table 1: contd…

$S_{53} = 6^{-1/2}$ [τ (C5-C4-C3-C2) - τ (C4-C3-C2-C1) + τ (C3-C2-C1-C6) - τ (C2-C1-C6-C5) + τ (C1-C6-C5-C4) - τ (C6-C5-C4-C3)]	τR_1 (A1)
S_{54} = ½ [-τ (C3-C2-C1-C6)+ τ (C5-C4-C3-C2)- τ (C6-C5-C4-C3) + τ (C2-C1-C6-C5)]	τR_2 (A1)
$S_{55} = 12^{-1/2}$ [-τ (C5-C4-C3-C2) + 2τ (C4-C3-C2-C1) - τ (C3-C2-C1-C6) - τ (C2-C1-C6-C5) + 2τ (C1-C6-C5-C4) - τ (C6-C5-C4-C3)]	τR_3 (A1)
$S_{56} = 6^{-1/2}$ [α (C4-C9-O13) + α (O13-C12-N14) + α (N14-C3-C4) - α (C9-O13-C12) - α (C12-N14-C3) - α (C3-C4-C9)]	βR_1 (A2)
$S_{57} = 12^{-1/2}$ [2α (C4-C9-O13) - α (C9-O13-C12) - α (O13-C12-N14) + 2α (C12-N14-C3)- α (N14-C3-C4) - α (C3-C4-C9)]	βR_2 (A2)
S_{58} = ½ [α (C9-O13-C12)- α (O13-C12-N14)+ α (N14-C3-C4)- α (C3-C4-C9)]	βR_3 (A2)
$S_{59} = 6^{-1/2}$ [τ (C9-O13-C12-N14) - τ (O13-C12-N14-C3) + τ (C12-N14-C3-C4) - τ (N14-C3-C4-C9) + τ (C3-C4-C9-O13) - τ (C4-C9-O13-C12)]	τR_1 (A2)
S_{60} = ½ [-τ (C12-N14-C3-C4)+ τ (C9-O13-C12-N14)- τ (N14-C3-C4-C9) + τ (C4-C9-O13-C12)]	τR_2 (A2)
$S_{61} = 12^{-1/2}$ [-τ (C9-O13-C12-N14) + 2τ (O13-C12-N14-C3) - τ (C12-N14-C3-C4) - τ (N14-C3-C4-C9) + 2τ (C3-C4-C9-O13) - τ (C4-C9-O13-C12)]	τR_3 (A2)
$S_{62} = 6^{-1/2}$ [α (C21-C23-C19) + α (C19-C17-C16) + α (C16-C18-C21) - α (C23-C19-C17) - α (C17-C16-C18) - α (C19-C17-C16)]	βR_1 (A3)
$S_{63} = 12^{-1/2}$ [2α (C21-C23-C19) - α (C23-C19-C17) - α (C19-C17-C16) + 2α (C17-C16-C18)- α (C16-C18-C21) - α (C18-C21-C23)]	βR_2 (A3)
S_{64} = ½ [α (C23-C19-C17)- α (C19-C17-C16)+ α (C16-C18-C21)- α (C18-C21-C23)]	βR_3 (A3)
$S_{65} = 6^{-1/2}$ [τ (C23-C19-C17-C16) - τ (C19-C17-C16-C18) + τ (C17-C16-C18-C21) - τ (C16-C18-C21-C23) + τ (C18-C21-C23-C19) - τ (C21-C23-C19-C17)]	τR_1 (A3)
S_{66} = ½ [-τ (C17-C16-C18-C21)+ τ (C23-C19-C17-C16)- τ (C21-C23-C19-C17) + τ (C16-C18-C21-C23)]	τR_2 (A3)
$S_{67} = 12^{-1/2}$ [-τ (C23-C19-C17-C16) + 2τ (C19-C17-C16-C18) - τ (C17-C16-C18-C21) - τ (C16-C18-C21-C23) + 2τ (C18-C21-C23-C19) - τ (C21-C23-C19-C17)]	τR_3 (A3)
S_{68} = τ(N14-C3-C4-C5) + τ(N14-C3-C4-C9) - τ(C2-C3-C4-C9) - τ(C2-C3-C4-C5)	Butt
S_{69} = γ (C12-C16-C17-C18)	γ C12-C16
S_{70} = γ (C16-C12-O13-N14)	γ C16-C12
S_{71} = τ(O13-C12-C16-C17) + τ(O13-C12-C16-C18) - τ(N14-C12-C16-C17) + τ(N14-C12-C16-C18)	τwA2A3
S_{72}= α (C18-C16-C12) - α (C17-C16-C12)	β (C12-C16)
S_{73} = α (O13-C12-C16) - α (N14-C12-C16)	β (C12-O13)
S_{74}= 2α (H27-C26-H29) - α (H29-C26-H28) - α (H27-C26-H28)	δa CH_3
S_{75}= α (H29-C26-H28) - α (H27-C26-H28)	δa CH_3
S_{76}= α (H29-C26-H27) + α (H29-C26-H28) + α (H27-C26-H28) - β (H29-C26-C23) - β (H28-C26-C23) - β (H27-C26-C23)	δs CH_3

Table 1: contd…

S_{77} = 2β (H29-C26-C23) -β (H28-C26-C23) - β (H27-C26-C23)	ρ CH_3
S_{78} = β (H28-C26-C23) - β (H27-C26-C23)	ρ' CH_3
S_{79} = τ (C21-C23-C26-H29) + τ (C21-C23-C26-H27) + τ (C21-C23-C26-H28)	τw(CH_3)
S_{80}= γ (C26-C23-C21-C19)	γ C26-C23
S_{81} = α (C26-C23-C19) - α (C26-C23-C21)	β (C23-C26)

Abbreviations:ν, stretching; β, deformation in the plane; γ, deformation out of plane; wag, wagging; τ , torsion; β_R, deformation ring τ_R, torsion ring; ρ, rocking; tw, twisting; α, angular deformation; δ, deformation; Butt, butterfly; a, antisymmetric; s, symmetric; A1, benzyl ring; A2, pyrrol ring; a= cos 144º, b=cos 72º

The SQMFF methodology [33] was employed to evaluate the harmonic force field of the compound by using the B3LYP/6-31G* method and, then, the complete assignments of the vibrational spectra was performed taking into account the potential energy distribution (PED) components ≥ 10%. The *GaussView* program [40] was used to analyze the vibration normal modes.

RESULTS AND DISCUSSION

Geometry Optimization

Table **2** show a comparison of the total energy and dipole moment values for MPB by using both levels of theory with those reported for 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one (CPB) [15]. The energy values for MPB with both combinations are enough higher than the other ones, as it is expected, due to that CPB has in their structure a Cl atom linked to the phenyl ring. The similar dipolar moment values for both molecules show that the size and the electronegativity of that atom have practically no influence on those values. This way, only the energy values are strongly dependent on the method used, as can be seen in Table **2**.

Table 2: Calculated total energy (*E*) and dipolar moments for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one

B3LYP		
2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]		
Property	6-31G*	6-311++G**
E (Hartrees)	-783.4568	-783.6576
μ (D)	3.01	3.26
2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]		
E (Hartrees)	-1203.7336	-1203.9517
μ (D)	3.10	3.30

[a]This work
[b]From Ref [15]

The calculated geometrical parameters for MPB by using the two methods were compared with the experimental values determined for 2-phenyl-4*H*-3,1-benzoxazin-4-one (PB) by X-ray diffraction [41], as can be seen in Table **3**. Experimentally, the PB molecule is nearly planar with a dihedral angle of 3.72º (4) between the planar phenyl ring and the 3,1-benzoxazin-4-one moiety. On the contrary, the theoretical calculations by using both basis sets predict the MPB structure as essentially planar with the dihedral angles values between both rings of 0.2 and 0.3º. Comparing the theoretical results with the experimental ones by means of the root means of square deviations (rmsd) the values shows that both basis sets reproduce quite well the bond lengths (0.004 Å) and the bond angles (0.15º). Moreover, the calculated C12-N14 distances at the same levels calculations show that these bonds have a certain double character, as experimentally observed in CPB [15]. Probably, the molecular MPB structure is stabilized by intermolecular C-H---O hydrogen bonds as in CPB, whose crystal packing is stabilized by weak intermolecular C-H---O and π-π stacking interactions [41], as can be seen in Fig. **2**.

Table 3: Calculated geometrical parameters for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one

2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]			**2-Phenyl-4*H*-3,1-benzoxazin-4-one[b]**
Parameter	**6-31G***	**6-311++G****	**Experimental**
Bond Length (Å)			
C1-C2	1.388	1.385	1.369 (2)
C2-C3	1.406	1.405	1.394 (2)
C3-C4	1.413	1.411	1.393 (2)
C4-C5	1.402	1.401	1.387 (2)
C5-C6	1.388	1.385	1.371 (2)
C4-C9	1.461	1.461	1.448 (2)
C1-C6	1.405	1.404	1.381 (3)
C9-O15	1.205	1.200	1.1926 (19)
C9-O13	1.401	1.400	1.3791 (19)
C12-O13	1.370	1.369	1.3702 (18)
C12-N14	1.288	1.284	1.275 (2)
C12-C16	1.471	1.470	1.462(2)
C3-N14	1.387	1.385	1.394 (2)
C16-C17	1.406	1.403	1.384(2)
C16-C18	1.403	1.400	1.388(2)

Table 3: contd…

C17-C19	1.388	1.386	1.371(2)
C18-C21	1.393	1.391	1.376 (3)
C19-C23	1.405	1.403	1.374(3)
C21-C23	1.401	1.398	1.367 (3)
C23-C26	1.510	1.508	
RMSD[#]	0.004	0.004	
Bond angle (degrees)			
C1-C2-C3	119.98	120.00	119.49 (16)
C2-C3-C4	118.86	118.85	119.05 (15)
C3-C4-C5	120.74	120.70	120.63 (15)
C4-C5-C6	119.71	119.73	119.55 (16)
C5-C6-C1	119.84	119.84	120.00 (17)
C2-C1-C6	120.88	120.88	121.27 (16)
C3-C4-C9	118.86	118.77	118.68 (15)
C5-C4-C9	120.40	120.53	120.68 (15)
C12-C16-C17	119.47	119.57	119.16 (15)
C17-C16-C18	119.02	118.95	119.27 (17)
O13-C9-C4	114.36	114.27	115.34 (14)
O15-C9-C4	127.81	127.94	127.66 (16)
C16-C17-C19	120.25	120.29	120.20 (18)
C12-C16-C18	121.51	121.49	121.56 (16)
C17-C19-C23	121.21	121.23	119.9 (2)
C19-C23-C21	118.09	118.05	120.64 (19)
C21-C23-C26	121.18	121.24	119.7
C19-C23-C26	120.72	120.70	119.7
C18-C21-C23	121.29	121.28	119.9 (2)
C12-O13-C9	122.21	122.38	121.64 (12)
N14-C12-O13	124.64	124.43	124.73 (15)
N14-C12-C16	122.48	122.71	122.90 (15)
C3-N14-C12	118.08	118.27	117.80 (14)
C2-C3-N14	119.29	119.28	119.22 (15)
C4-C3-N14	121.84	121.87	121.73 (14)
O13-C12-C16	112.88	112.85	112.37 (14)
O15-C9-O13	117.83	117.79	117.00 (15)
C16-C18-C21	120.13	120.19	120.05 (19)
RMSD[#]	0.15	0.15	
Dihedral angles (degrees)			
C19-C17-C16-C18	- 0.050	-0.077	0.9 (2)
C17-C16-C18-C21	0.042	0.060	−0.4 (3)

Table 3: contd…

C16-C18-C21-C23	0.118	0.102	0.0 (3)
C18-C21-C23-C19	-0.262	-0.243	−0.1 (3)
C21-C23-C19-C17	0.254	0.227	0.7 (3)
C23-C19-C17-C16	-0.101	-0.070	−1.0 (3)
C19-C17-C16-C12	-179.89	-179.92	−179.33 (15)
C2-C3-C4-C9	179.99	179.99	178.16 (14)
C2-C3-C4-C5	-0.003	-0.006	−0.8 (2)
N14-C3-C4-C5	-180.00	179.99	178.74 (14)
C6-C1-C2-C3	0.00	0.00	−0.3 (3)
N14-C3-C4-C9	-0.00	-0.010	−2.3 (2)
C5-C6-C1-C2	-0.00	-0.004	−0.6 (3)
C4-C5-C6-C1	0.00	-0.000	0.8 (3)
N14-C12-O13-C9	0.00	-0.015	0.1 (2)
C2-C3-N14-C12	-179.99	-179.99	179.81 (15)
C4-C3-N14-C12	0.005	0.013	0.2 (2)
C3-N14-C12-C16	-179.97	-179.98	−179.59 (12)
C3-N14-C12-O13	-0.005	-0.000	0.9 (2)
C1-C2-C3-N14	180.00	-180.00	−178.56 (14)
C1-C2-C3-C4	0.00	0.00	1.0 (2)
N14-C12-C16-C17	0.00	0.091	−3.3 (2)
C5-C4-C9-O13	179.99	179.99	−177.93 (14)
C5-C4-C9-O15	-0.003	0.005	3.1 (3)
O13-C12-C16-C18	0.198	0.268	−3.9 (2)
O13-C12-C16-C17	-179.97	-179.89	176.31 (13)
C14-C12-C16-C18	-179.83	-179.75	176.50 (15)
C12-C16-C18-C21	179.88	179.90	179.88 (15)
C9-O13-C12-C16	179.97	179.96	−179.43 (13)
C3-C4-C9-O15	180.00	180.00	−175.84 (17)
O15-C9-O13-C12	-180.00	-179.99	176.92 (16)
C3-C4-C9-O13	-0.000	-0.005	3.1 (2)
C4-C9-O13-C12	0.000	0.017	−2.1 (2)
C3-C4-C5-C6	0.00	0.005	−0.1 (2)
C9-C4-C5-C6	-179.99	-179.99	−179.02 (15)
C17-C19-C23-C26	-178.88	-178.92	
C18-C21-C23-C26	178.87	178.90	
[c]RMSD[#]	**30.66**	**27.02**	

[a]This work
[b]From Ref [41]
[c]RMSD
[#]See text

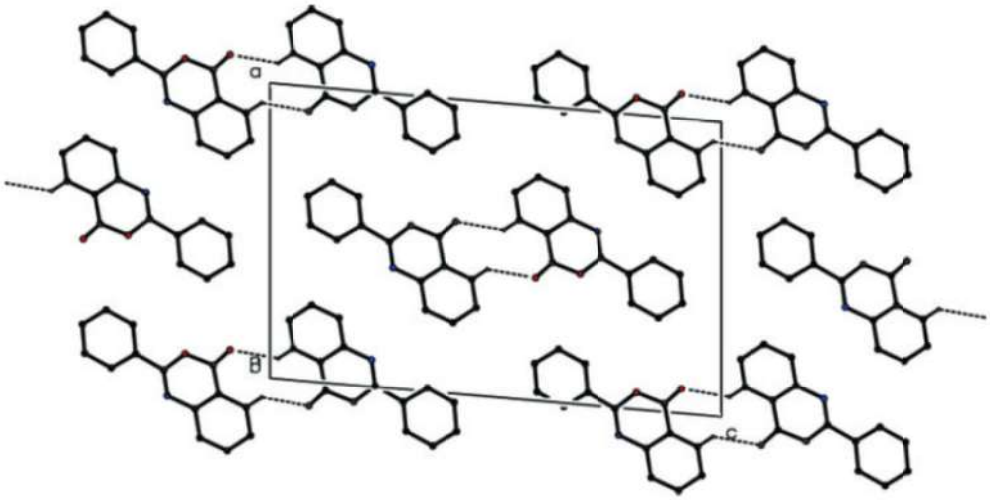


Figure 2: The crystal packing for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one viewed down the *b* axis. C-H---O hydrogen bonds are shown as dashed lines (According to Ref. [42]).

NBO Analysis

The stability of MPB was studied by means of NBO and AIM calculations. In this study, the atomic charges, as natural population analysis (NPA), the bond orders, normally expressed by Wiberg's index, and the second order perturbation energies were calculated by using NBO calculations [35]. The atomic charges by using both calculation levels compared with those obtained for the CPB derivative [15] are given in Table **4**.

Table 4: Atomic charges (NPA) for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one at different theory levels

B3LYP METHOD					
2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]			**2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]**		
Atoms	**6-31G**	**6-311++G****	**Atoms**	**6-31G**	**6-311++G****
1. C	-0.204	-0.175	1. C	-0.203	-0.174
2. C	-0.226	-0.175	2. C	-0.224	-0.174
3. C	0.175	0.167	3. C	0.173	0.165
4. C	-0.206	-0.201	4. C	-0.205	-0.200
5. C	-0.175	-0.141	5. C	-0.174	-0.140
6. C	-0.240	-0.207	6. C	-0.238	-0.204

Table 4: contd…

7. H	0.243	0.209	7. H	0.245	0.210
8. H	0.252	0.219	8. H	0.253	0.220
9. C	0.808	0.792	9. C	0.808	0.792
10. H	0.260	0.230	10. H	0.261	0.230
11. H	0.244	0.211	11. H	0.245	0.212
12. C	0.612	0.599	12. C	0.609	0.596
13. O	-0.530	-0.537	13. O	-0.531	-0.537
14. N	-0.511	-0.512	14. N	-0.507	-0.508
15. O	-0.558	-0.554	15. O	-0.554	-0.549
16. C	-0.139	-0.108	16. C	-0.132	-0.101
17. C	-0.177	-0.153	17. C	-0.172	-0.146
18. C	-0.186	-0.162	18. C	-0.183	-0.156
19. C	-0.233	-0.204	19. C	-0.249	-0.224
20. H	0.258	0.228	20. H	0.265	0.233
21. C	-0.232	-0.204	21. C	-0.247	-0.223
22. H	0.258	0.228	22. H	0.264	0.233
23. C	-0.009	-0.003	23. C	-0.030	-0.020
24. H	0.237	0.204	24. H	0.257	0.224
25. H	0.238	0.205	25. H	0.258	0.225
26. C	-0.692	-0.593	26. Cl	0.008	0.015
27. H	0.244	0.212			
28. H	0.249	0.217			
29. H	0.240	0.208			

[a]This work
[b]From Ref [15]

The results with both basis sets show clearly that the charges on the atoms of phenyl ring fused with the oxazin moiety are practically constants; indicating that the stability of MPB is mainly associated with the variations on the charges of the atoms belonging to the methylphenyl or chlorophenyl rings, as consequence of exchange Cl by CH_3. Hence, the effect of this exchange is the decreasing in the charge values on the atoms belonging to the methylphenyl ring. For MPB, the calculated bond orders are given in Table **5** compared with the corresponding values for the CPB derivative [15].

Table 5: Wiberg bond Index for the 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one at different theory levels

B3LYP METHOD					
2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]			**2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]**		
Atoms	**6-31G**	**6-311++G****	**Atoms**	**6-31G***	**6-311++G****
1. C	3.945	3.963	1. C	3.944	3.963
2. C	3.944	3.959	2. C	3.943	3.958
3. C	3.985	3.986	3. C	3.984	3.986
4. C	3.984	3.987	4. C	3.984	3.987
5. C	3.935	3.953	5. C	3.934	3.953
6. C	3.945	3.963	6. C	3.945	3.962
7. H	0.942	0.959	7. H	0.941	0.958
8. H	0.939	0.956	8. H	0.938	0.955
9. C	3.829	3.838	9. C	3.829	3.837
10. H	0.935	0.950	10. H	0.934	0.950
11. H	0.942	0.958	11. H	0.941	0.957
12. C	3.887	3.889	12. C	3.889	3.892
13. O	2.210	2.216	13. O	2.209	2.215
14. N	3.087	3.092	14. N	3.089	3.093
15. O	2.078	2.079	15. O	2.083	2.084
16. C	3.997	3.995	16. C	3.998	3.997
17. C	3.934	3.954	17. C	3.931	3.952
18. C	3.935	3.955	18. C	3.933	3.953
19. C	3.946	3.963	19. C	3.940	3.957
20. H	0.935	0.951	20. H	0.932	0.948
21. C	3.946	3.962	21. C	3.940	3.957
22. H	0.935	0.950	22. H	0.932	0.948
23. C	3.999	4.004	23. C	4.016	4.019
24. H	0.945	0.961	24. H	0.935	0.952
25. H	0.945	0.960	25. H	0.935	0.952
26. C	3.836	3.884	26. Cl	1.217	1.232
27. H	0.942	0.957			
28. H	0.940	0.955			
29. H	0.943	0.959			

[a]This work
[b]From Ref [15]

For MPB, we observed that the bond order values corresponding to the atoms of phenyl ring fused with the oxazin moiety remain practically constant, as observed

previously. The second order perturbation energies $E^{(2)}$ (donor → acceptor) obtained by NBO analysis for the most important delocalizations of MPB are presented in Table **6.** The values by using both methods were compared with those obtained for the CPB derivative [15]. The results with both calculation levels, show that the contributions of the stabilization energies to the $\Delta ET_{\sigma^* \rightarrow \sigma^*}$ charge transfers due mainly to the chlorophenyl ring are drastically higher for CPB than those corresponding to the methylphenyl ring for MPB.

Table 6: Main delocalization energy (in kJ/mol) for 2-(4-methylphenyl)-4H- 3,1-benzoxazin-4-one at different theory levels

B3LYP Method					
2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]			**2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]**		
Delocalization	**6-31G***	**6-311++G****	**Delocalization**	**6-31G***	**6-311++G****
σ C1-C2 → σ C3-C4*	90.00	90.79	*σ C1-C2 → σ* C3-C4*	90.71	91.46
σ C1-C2 → σ C5-C6*	70.98	70.89	*σ C1-C2 → σ* C5-C6*	71.10	71.02
σ C3-C4→ σ C1-C2*	62.49	61.45	*σ C3-C4→ σ* C1-C2*	62.53	61.49
σ C3-C4 → σ C5-C6*	83.60	83.77	*σ C3-C4 → σ* C5-C6*	82.76	82.93
σ C3-C4→ σ C9=O15*	110.81	105.71	*σ C3-C4→ σ* C9=O15*	110.18	104.96
σ C5-C6→ σ C1-C2*	85.10	85.15	*σ C5-C6→ σ* C1-C2*	84.85	84.94
σ C5-C6→ σ C3-C4*	73.53	73.65	*σ C5-C6→ σ* C3-C4*	74.45	74.57
σ C12-N14→ σ C3-C4*	74.40	75.37	*σ C12-N14→ σ* C3-C4*	72.56	73.57
σC16-C18 → σ C12-N14*	95.26	93.42	*σC16-C18 → σ* C12-N14*	90.75	89.03
*σC16-C18 → σ*C17-C19*	84.39	84.85	*σC16-C18 → σ*C17-C19*	85.48	86.02
*σC16-C18 → σ*C21-C23*	73.78	73.11	*σC16-C18 → σ*C21-C23*	86.15	85.27
σ C17-C19→ σ C16-C18*	74.53	74.70	*σ C17-C19→ σ* C16-C18*	75.45	75.74
σ C17-C19→ σ C21-C23*	91.17	92.34	*σ C17-C19→ σ* C21-C23*	90.83	92.25
σ C21-C23→ σ C16-C18*	99.23	100.03	*σ C21-C23→ σ* C16-C18*	84.27	84.52
σ C21-C23→ σ C17-C19*	71.98	71.56	*σ C21-C23→ σ* C17-C19*	72.73	72.48
ΔET σ→σ*	**1241.25**	**1236.79**	***ΔET σ→σ****	**1234.8**	**1230.25**
σ C9-O13→ σ* C12-O13*	101.53	89.70	*σ* C9-O13→ σ* C12-O13*	88.11	78.96
σ C9-O15→ σ* C3-C4*	891.51		*σ* C9-O15→ σ* C3-C4*	915.59	-
σ C12-N14→ σ* C3-C4*	757.75		*σ* C12-N14→ σ* C3-C4*	667.17	807.91
σ C12-N14→ σ* C16-C18*	359.31	363.20	*σ* C12-N14→ σ* C16-C18*	458.59	464.44
			σ C3-C4→ σ* C5-C6*	1188.54	1157.19

Table 6: contd…

			σ C21-C23→ σ* C16-C18*	968.05	1033.76
			σ C21-C23→ σ* C17-C19*	589.25	650.62
ΔET $_{\sigma^* \rightarrow \sigma^*}$	**2110.1**	**452.9**	***ΔET*** $_{\sigma^* \rightarrow \sigma^*}$	**4875.3**	**4192.88**
LP(2)O13 → σ C9-O15*	149.23	139.44	*LP(2)O13 → σ* C9-O15*	146.51	136.73
LP(2)O13 → σ C12-N14*	164.32	158.97	*LP(2)O13 → σ* C12-N14*	165.70	160.22
LP(1)N14 → σ C12-O13*	85.69	82.30	*LP(1)N14 → σ* C12-O13*	85.73	82.35
LP(2)O15 → σ C4-C9*	72.11	68.18	*LP(2)O15 → σ* C4-C9*	72.15	68.34
LP(2)O15 → σ C9-O13*	169.54	162.14	*LP(2)O15 → σ* C9-O13*	171.55	164.40
			LP(3)Cl26 → σ C21-C23*	52.67	54.34
ΔET $_{LP \rightarrow \sigma^*}$	**640.89**	**611.03**	***ΔET*** $_{LP \rightarrow \sigma^*}$	**694.31**	**666.38**
ΔETotal	**3992.23**	**2258.91**	***ΔETotal***	**6804.41**	**6089.51**

[a]This work
[b]From Ref [15]

On the contrary, the $\Delta ET_{\sigma \rightarrow \sigma^*}$ charge transfers are higher for MPB than CPB, as observed in Table **6**. Thus, the ΔE_{Total} larger energies for CPB than MPB show a higher stability for the first compound attributed principally to the hyperconjugation of the chlorophenyl ring.

AIM Studies

The presence of a methyl group in the phenyl ring linked to the benzoxazin ring in MPB slightly modify the stability of this compound in relation to CPB, as observed previously by NBO analysis by using the 6-31G* and 6-311++G** basis sets. For these reasons, the topological properties in the ring critical points (RCPs) of the benzoxazin and methylphenyl rings of MPB were also investigated by means of the AIM theory [36] and, then, their values were compared with those reported for CPB and with those calculated, in this chapter, at the same levels of approximation for the PB derivative. The values of the topological properties, such as the calculated charge electron density, (ρ) and the *Laplacian* values, $\nabla^2\rho(r)$ in the ring critical points for the three benzoxazin derivatives are shown in Table **7**. It is necessary to clarify that for MPB, CPB and PB, RCP_1 corresponds to the phenyl rings fused with the oxazin moiety; RCP_2 corresponds to the oxazin ring while RCP_3 corresponds to the methylphenyl, chlorophenyl or phenyl rings. The analysis shows clearly for the

MPB and CPB molecules, by using both methods, that the three expected RCPs have slightly different topological properties and, that when increase the size of the basis set the values in all the cases increase too, as indicated in Table **7**. It is important to note that for the three molecules, by using both basis sets, the topological properties of RCP_1 do not change while between RCP_2 and RCP_3 the properties are different. Thus, in the RCPs analyzed for the three molecules the same relations in the (ρ) and $\nabla^2\rho(r)$ values were observed. These relations following the trend: $(\rho)RCP_2 > (\rho)RCP_3 > (\rho)RCP_1$, and $\nabla^2(\rho)RCP_3 > \nabla^2(\rho)RCP_1 > \nabla^2(\rho)RCP_2$, as can be seen in Table **7**. The AIM results are in complete agreement with the previous studies performed by means of NBO calculations.

Table 7: An analysis of the Ring Critical Points (RCP) for 2-(4-methylphenyl)-4H-3,1-benzoxazin -4-one at different theory levels

2-(4-Methylphenyl)-4H-3,1-benzoxazin-4-one[a]						
Parameters	**6-31G***			**6-311G****		
	RCP_1	**RCP_2**	**RCP_3**	**RCP_1**	**RCP_2**	**RCP_3**
$\rho(r)$	0.01990	0.02150	0.02015	0.02141	0.02292	0.02173
$\nabla^2\rho(r)$	0.15817	0.15735	0.16029	0.15652	0.15652	0.15864
λ_1	-0.01495	-0.01834	-0.01507	-0.01668	-0.02008	-0.01684
λ_2	0.08338	0.07840	0.08274	0.08365	0.08058	0.08324
λ_3	0.08974	0.09729	0.09261	0.08956	0.09603	0.09223
$\lvert\lambda_1\rvert/\lambda_3$	0.16659	0.18851	0.16272	0.18624	0.20910	0.18259
2-(4-Chlorophenyl)-4H-3,1-benzoxazin-4-one[b]						
Parameters	RCP_1	RCP_2	RCP_3	RCP_1	RCP_2	RCP_3
$\rho(r)$	0.01991	0.02148	0.02016	0.02142	0.02290	0.02175
$\nabla^2\rho(r)$	0.15824	0.15723	0.15996	0.15660	0.15636	0.15852
λ_1	-0.01497	-0.01833	-0.01517	-0.01670	-0.02005	-0.01692
λ_2	0.08355	0.07819	0.08529	0.08381	0.08032	0.08519
λ_3	0.08964	0.09737	0.08982	0.08948	0.09611	0.09024
$\lvert\lambda_1\rvert/\lambda_3$	0.16700	0.18825	0.16889	0.18663	0.20861	0.18750
2-Phenyl-4H-3,1-benzoxazin-4-one[a]						
Parameters	RCP_1	RCP_2	RCP_3	RCP_1	RCP_2	RCP_3
$\rho(r)$	0.01990	0.02149	0.02109	0.02142	0.02292	0.02282
$\nabla^2\rho(r)$	0.15823	0.15732	0.16474	0.15660	0.15648	0.16269
λ_1	-0.01496	-0.01833	-0.01652	-0.01670	-0.02007	-0.01855
λ_2	0.08346	0.07834	0.08354	0.08375	0.08049	0.08453

Table 7: contd…

λ_3	0.08973	0.09730	0.09772	0.08954	0.09606	0.09671
$/\lambda_1//\lambda_3$	0.16675	0.18839	0.16905	0.18651	0.20893	0.19181

[a]This work
[b]From Ref [15]

VIBRATIONAL ANALYSIS

Fig. **3** show the infrared spectrum of MPB in solid phase recorded and reported by Rai [9] whiles the corresponding theoretical predicted by us at B3LYP/6-31G* level can be seen in Fig. **4.** A comparison between the predicted infrared spectrum for MPB by using the B3LYP/6-31G* method with the corresponding experimental one is presented in Fig. **5**.

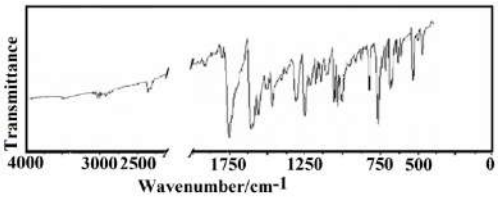


Figure 3: Experimental infrared spectrum of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one (From Ref [9]).

The calculations with both methods show that the stable structure of MPB has C_1 symmetry and a total of 29 atoms in their structure, thus, 81 normal vibration modes (3N-6= 3x29-6 = 81) are expected for this molecule. Here, the entire vibration normal modes are IR and Raman active. For MPB, the experimental and calculated wavenumbers, the 6-31G*/SQM and the corresponding assignments are summarized in Table **8** together with those corresponding to CPB [15] and PB [42]. On the other hand, the theoretical and observed wavenumbers, potential energy distribution and assignment for MPB are given in Table **9**. In the higher wavenumbers region, those between 3000-2500 and 2000-500 cm^{-1} are observed a set of characteristic broad IR bands that could probably be assigned to the hydrogen bonds due to the crystal packing of the molecules of MPB in the solid phase, as was experimentally reported for CPB (see Fig. **2**) [41]. For MPB, the complete vibrational assignment was performed by comparison with assignments reported for similar molecules [15-20, 24-29, 42, 43] and taking into account the results of the calculations performed by using the B3LYP/6–31G* method. The

scale factors used to calculate the force field of MPB are defined only for the B3LYP/6-31G* method and they were reported by Pulay *et al*. [33].

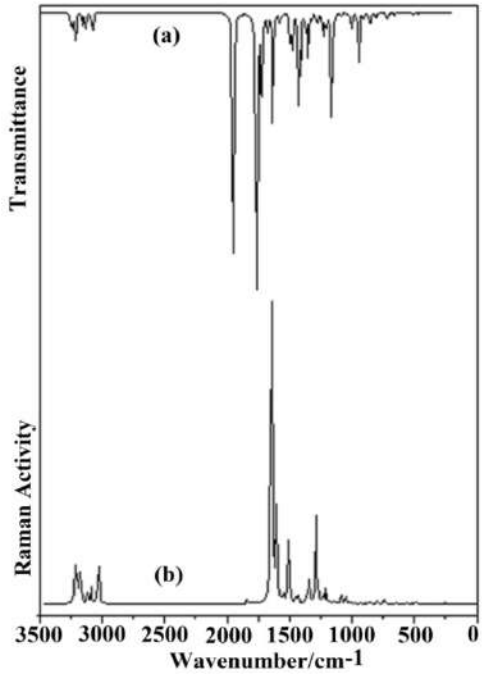


Figure 4: Theoretical spectra of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one at the B3LYP/6-31G*: (a), infrared spectrum and (b) Raman spectrum.

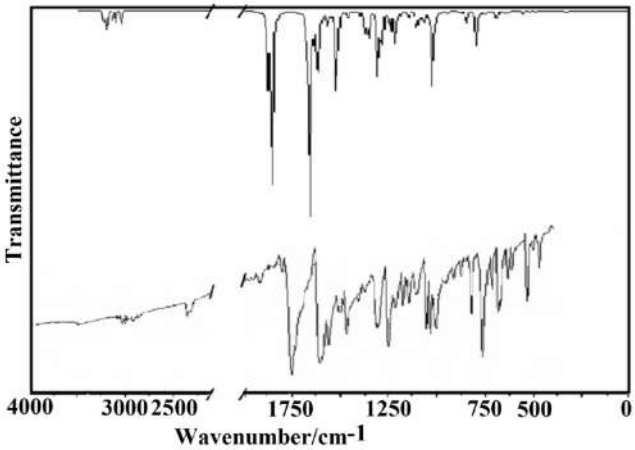


Figure 5: Comparison between the calculated infrared spectrum of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one by using B3LYP/6–31G* level (Upper) with the corresponding experimental (From Ref [9]).

Table 8: Observed and calculated wavenumbers (cm^{-1}) and assignment for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one

2-phenyl-4H-3,1-benzoxazin-4-one[b]			2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[c]			2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]		
IR Solid	**Raman Solid**	**Assignment**	**IR Solid**	**SQM[d]**	**Assignment**	**IR Solid**	**SQM[d]**	**Assignment**
		νCH I 20a		3107	νC18-H22			
		νCH I 20b		3103	νC17-H20		3101	ν C18-H22
		νCH II 20a		3094	νC5-H10		3097	ν C17-H20
		νCH II 20b		3091	νC21-H25		3094	ν C5-H10
		νCH I 2		3090	νC2-H8		3089	ν C2-H8
		νCH II 7b		3089	νC19-H24		3076	ν C6-H11
3060 w	3070 s	νCH I 13	3076 w	3077	νC6-H11	3063 s	3061	ν C1-H7
3050 w	3022 w	νCH II 2					3056	ν C21-H25
3040 m							3052	ν C19-H24
			3025 vw	3063	νC1-H7			
						3009 s	3000	va CH_3
						2920 s	2971	va CH_3
						2854 s	2917	vs CH_3
1763 vs	1757 m	νC=O	1768 vs	1788	νC9-O15	1759 vs	1785	ν C9-O15
1692 s		νC=N	1657 s	1616	νC1-C2			
			1624 s	1603	νC12-N14			
1613 s	1623 vs	νPh II 8b	1600 sh	1595	νC18-C21	1608 s	1618 1615 1598	ν C18-C21 ν C1-C2 ν C12-N14
1585 sh	1599 vs	νPh I 8a	1566 m	1568	νC3-C4	1568 s	1568	ν C16-C18
1578 s	1574 s	νPh I 8b	1566 m	1559	νC16-C17		1560	ν C16-C17
		νPh II 8a	1490 sh	1492	β(C18-H22)	1510 s	1513	β(C19-H24)
1490 m		νPh I 19a	1475 s	1470	β(C5-H10)			
1474 s	1476 s	νPh II 19b	1460 sh	1465	β(C17-H20), β(C6-H11)	1471 s	1470 1466 1458 1450	νC5-C6 β(C6-H11) δaCH_3 δaCH_3
1450 m	1451 w	νPh II 19a	1425 m	1403	νC17-C19			
						1412 m	1411	νC17-C19
						1381 m	1378	δsCH_3
		νPh I 19b	1320 sh	1321	νC2-C3		1322	νC4-C5
							1316	β(C2-H8)
	1322 s	νPh I 14, νCN	1315 s	1315	β(C2-H8)		1314	β(C18-H22)
1315 s		νPh II 3	1300 sh	1309	νC16-C18	1311 s	1311	ν C21-C23
	1300 w	δCH II 14	1290 sh	1295	νC19-C23			
		δCH I 3	1271 m	1254	νC3-N14	1259 s	1254	ν C3-N14
1258 s	1259 vs	?C(X)X II 13 ?CO	1240 sh	1243	νC4-C5	1221 s	1244	β(C5-H10)
1236 m		δ CH I 9a	1222 m	1207	νC4-C9		1208	νC4-C9
1183 m	1183 w	δ CH II 15	1175 sh	1176	β(C21-H25)	1180 s	1199 1183	νC19-C23 β(C21-H25)

Table 8: contd…

		δ CH I 9b	1150 w	1156	β(C1-H7)	1155 s	1155	β(C1-H7)
1153 w	1157 m	δ CH II 9b δCH II	1110 w 1100 sh	1109 1108	νC5-C6 β(C17-H20), β(C19-H24)	1113 m	1119 1109	β(C17-H20) βR_1(A1)
1112 m		νC(X)X II 1	1089 m	1074	νC21-C23			
		νC(X)X I	1050 m	1056	νC12-O13	1061 s	1057	ν C12-O13
							1049	ρ'CH_3
1058 m	1064 w	δCH I 15	1029 sh	1025	νC6-C1	1038 s	1026	νC6-C1
1027 m	1028 w	νCOC	1010 sh	1015	βR_1(A3)	1009 s	1019 993 992	βR_1(A3) γC6-H11 ρCH_3
1018 m		δCH I 18a	1001 s	995	γC1-H7			
1012 s	1002 s	δCH II 18b	990 sh	986	νC9-O13		989	νC9-O13
984 m		Breathing I 1	980 sh	980	γC17-H20		982	γC17-H20
		δCH I	970 sh	971	γC18-H22		970	γC18-H22
		γCH I 5	947 w	967	γC5-H10	962 m	966	γC5-H10
		γCH II 5	921 w	915	βR_1(A2)	922 m	916	βR_1(A2)
929 w		γCH II 17b	887 w	888	γC2-H8	881 w	887	γC2-H8
		γCH I 17a	863 w	856	βR_1(A1)	859 w	857	νC2-C3
882 w		γCH I 17b	850 vw	848	γC19-H24		852	γC19-H24
858 w	858 w	γCH II 17b	820 sh	836	γC21-H25	825 s	840	γC21-H25
		γCH II 17a	780 sh	785	γC6-H11		787	νC23-C26
		γCH I 10a	762 s	772	Butt	769 s	785 772	γC1-H7 Butt
	784 w	δPh(X) II	740 sh	746	βR_2(A3)			
765 s		γCH I, II 11	730 sh	730	γC16-C12			
	749 w	γCH I, 11	725 m	714	βR_2(A1)	725 m	732 729	νC3-C4 γC16-C12
		δ Ph (X) II 4	686 s	683	γC9-O15		686	γC9-O15
	680 w	δPh(X) I 12	657 m	669	τR_1(A3)	677 s	675	τR_1(A3)
684 s		δ Ph I 4	640 sh	643	βR_3(A3)		651	βR_3(A3)
		γ Ph II 4	625 sh	629	β(C9-O15)	640 w	638	βR_2(A1)
		γ Ph (X) II 6a	619 w	597	βR_2(A1), βR_3(A1)	615 w	622	βR_3(A1)
628 m	618 w	δ C=O	576 vw	542	βR_3(A2)	581 vw	563	βR_2(A2)
600 w		δ Ph I 6b	532 vw	536	τR_1(A1)	540 s 511 vw	537 512	τR_1(A1) βR_2(A3)
537 m	568 w	δ Ph (X) II 6b γ C=O	476 vvw	487 480	βR_2(A2) τR_3(A3)	471 w	487 480	βR_3(A2) τR_3(A3)
	520 w	γ Ph(X) II 16a	448 vw	448	νC23-Cl26			
483 w	487 w	γ Ph(X) II 16b	430 sh	425	τR_2(A1)		425	τR_2(A1)
		γ Ph II 18a	415 sh	405	τR_2(A3)		403	τR_2(A3)
425 w		γ Ph(X) I 16b		384	β(C12-C16)		393	β(C23-C26)
		γ Ph I 16[a]		326	τR_3(A1)		327	γC23-C26
	358 w	δ Ph (X) I 6a		304	βR_3(A1)		316	β(C9-O15)
							276	β(C12-C16)

Table 8: contd…

	305 w	δ CX (X) II		251	β(C23-Cl26)		251	τR_3 (A1)
		γ Ph(X) II 7b		246	γC23-Cl26		243	νC12-C16
	260 w	γ Ph(X) II 9a		217	νC12-C16			
	178 w	δ CX (X) I 18b		151	τR_3 (A2)		151	τR_2 (A2)
	136 w	γ CX (X) I 10b		136	τR_1 (A2)		142	τR_1 (A2)
		t Ph I		96	τR_2 (A2)		103	τR_3 (A2)
	97 m	t Ph II		82	β(C12-O13)		93	β(C12-O13)
							47	γC12-C16
		t Ph II		45	τwRing(A3)		40	τwA2A3
		t Ph II		39	γC12-C16		31	$\tau w(CH_3)$

Abbreviations:ν, stretching; β, deformation in the plane; γ, deformation out of plane; wag, wagging; τ , torsion; β_R, deformation ring τ_R, torsion ring; ρ, rocking; τw, twisting; δ, deformation; Butt, butterfly; a, antisymmetric; s, symmetric; A1, phenyl Ring 1; A2, oxazin Ring 2; A3, methylphenyl Ring3,
[a] This work, [b]From Ref [42],[c]From Ref [15], [d]DFT B3LYP/6-31G*

Fig. **6** show a comparison between the calculated infrared spectrum of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one by using B3LYP/6–31G* level with the corresponding experimental obtained from Ref [44] for 2-Phenyl-4H-3,1-benzoxazin-4-one. Here, the presence of bands associated to the CH_3 vibration modes in the theoretical spectrum of MPB justifies the difference between both spectra. For MPB, the resultant B3LYP/6–31G*/SQM force field can be obtained at request of the authors. The assignment of the most important groups is discussed below.

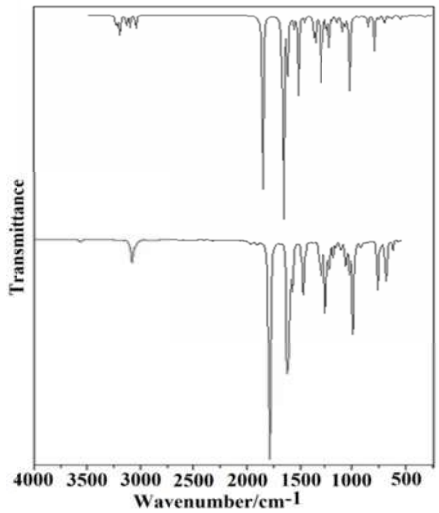


Figure 6: Comparison between the calculated infrared spectrum of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one by using B3LYP/6–31G* level (upper) and the experimental of 2-Phenyl-4H-3,1-benzoxazin-4-one (From Ref [44]).

BANDS ASSIGNMENTS

CH modes. In compounds containing this group [15-20,24-29,42,43], the C-H stretching modes are assigned in the 3140-2849 cm^{-1} region. Hence, the strong IR band at 3063 cm^{-1} can be assigned to those vibration modes, as indicated in Table **8**.

For MPB, the bands located between 1510 and 1113 cm^{-1} are assigned to the eight expected in-plane deformation modes while the corresponding out-plane deformation modes are associated with the IR bands between 1009 and 769 cm^{-1}, as observed in Table **8**.

CH_3 modes. Here, the PED contribution show that the antisymmetric and symmetric stretching modes of the methyl group are calculated as totally pure modes; hence, they are easily assigned respectively to the strong IR bands at 3009, 2920 and 2854 cm^{-1}.

Table 9: Observed and calculated wavenumbers (cm^{-1}), potential energy distribution and assignment for the 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one

Modes	IR[a] Sol	calculated[b]	IR Int[c]	Raman Int.	SQM	PED[a] ≥ 10%
1		3235	1.8	63.5	3101	S_{26}(96)
2	3063 s	3230	2.7	40.4	3097	S_{25}(96)
3		3227	10.8	237	3094	S_{23}(71), S_{24}(16)
4		3223	7.3	63.4	3089	S_{22}(78) S_{23}(13)
5		3209	13.7	180.5	3076	S_{24}(60),S_{21}(18),S_{23}(13)
6		3193	6.2	88.1	3061	S_{21}(69) S_{24}(23)
7		3188	21.7	128.8	3056	S_{28}(95)
8		3184	18.8	110.7	3052	S_{27}(95)
9	3009 s	3130	15.3	77.9	3000	S_{30}(93)
10	2920 s	3099	19.3	128.3	2971	S_{29}(87)
11	2854 s	3043	31.2	406	2917	S_{31}(92)
12	1759 vs	1854	430	40.2	1785	S_{1}(80)
13		1673	12.6	79	1618	S_{19}(20),S_{17}(19),S_{15}(13),S_{63}(10)
14	1608 s	1671	59.9	565	1615	S_{10}(24),S_{4}(20)
15		1652	404.5	2750	1598	S_{4}(27)
16	1568 s	1621	7.9	31.7	1568	S_{11}(16),S_{12}(15),S_{15}(12)
17		1614	167.2	1288.5	1560	S_{4}(17),S_{15}(17)

Table 9: contd…

18	1510 s	1561	20.7	63	1513	S_{20}(16),S_{39}(14),S_{37}(14),S_{36}(13),S_{38}(12)
19	1471 s	1521	22.4	118.8	1470	S_{12}(22),S_{33}(20),S_{34}(15),S_{5}(12),S_{32}(10)
20		1518	122.8	392.3	1466	S_{35}(27),S_{13}(22),S_{32}(15)
21		1516	8.7	50.6	1458	S_{75}(61),S_{74}(14)
22		1510	0.2	16.3	1450	S_{74}(75),S_{75}(15)
23	1412 m	1457	13.5	52.4	1411	S_{17}(20),S_{19}(13),S_{75}(12),S_{38}(12)
24	1381 m	1442	1.0	72.6	1378	S_{76}(89)
25		1372	35.5	84.8	1322	S_{8}(25) S_{13}(19)
26		1361	6.0	16.1	1316	S_{14}(19) S_{33}(11),S_{56}(10)
27		1353	35.2	212.4	1314	S_{37}(13) S_{8}(13) S_{36}(12),S_{39}(12) S_{13}(11) S_{38}(10)
28	1311 s	1350	15.8	26.5	1311	S_{20}(28) S_{18}(23) S_{16}(12)
29	1259 s	1298	113.5	705.4	1254	S_{5}(31) S_{3}(13)
30	1221 s	1282	52.2	132	1244	S_{34}(17),S_{14}(16),S_{33}(11)
31		1244	7.7	72.3	1208	S_{7}(25),S_{50}(14) S_{9}(10)
32		1241	15.0	10.2	1199	S_{6}(38),S_{18}(14) S_{62}(11)
33	1180 s	1217	62.9	132.1	1183	S_{38}(22),S_{37}(20) S_{36}(15) S_{39}(15),S_{19}(10)
34	1155 s	1189	6.2	11.0	1155	S_{32}(30),S_{35}(21) S_{33}(15) S_{11}(10)
35		1153	6.4	17.4	1119	S_{36}(18),S_{39}(17) S_{17}(14) S_{38}(12),S_{37}(10)
36	1113 m	1139	10.4	2.7	1109	S_{50}(18),S_{35}(12) S_{34}(10) S_{32}(10)
37	1061 s	1093	35.7	87.3	1057	S_{3}(36)
38		1074	14.7	0.24	1049	S_{78}(55),S_{80}(11)
39	1038 s	1061	14.6	48.4	1026	S_{11}(28),S_{2}(14)
40	1009 s	1039	7.2	0.94	1019	S_{62}(51),S_{20}(15)
41		1021	45.0	2.9	993	S_{41}(37),S_{43}(25),S_{42}(14),S_{44}(11),S_{53}(11)
42		1017	109.6	8.0	992	S_{77}(39),S_{2}(14)
43		1002	0.05	0.7	989	S_{2}(26),S_{77}(16),S_{40}(13)
44		991	1.0	3.3	982	S_{45}(46),S_{46}(24),S_{65}(12)
45		980	0.2	0.9	970	S_{48}(45),S_{47}(32)
46	962 m	976	1.3	0.4	966	S_{44}(46),S_{42}(19),S_{41}(14),S_{43}(12)
47	922 m	936	3.9	6.4	916	S_{56}(31),S_{2}(17),S_{3}(14)
48	881 w	899	1.6	4.4	887	S_{42}(38),S_{44}(18),S_{43}(14)
49	859 w	876	3.2	8.7	857	S_{50}(29),S_{52}(10),S_{9}(9)
50		862	0.3	5.2	852	S_{46}(37),S_{45}(26),S_{47}(19),S_{48}(17)
51	825 s	851	19.3	0.2	840	S_{47}(26),S_{48}(17),S_{46}(15),S_{69}(10)
52		808	1.3	21.7	787	S_{63}(24),S_{6}(23),S_{15}(12),
53		797	1.9	2.3	785	S_{41}(23),S_{49}(22),S_{43}(17),S_{42}(12)
54	769 s	789	57.7	0.1	772	S_{68}(26),S_{53}(21),S_{59}(19),S_{43}(12)

Table 9: contd…

55		753	4.1	32.9	732	S_{51}(16),S_{2}(10),
56	725 m	744	4.7	5.4	729	S_{65}(34),S70(21),S69(13)
57		701	5.1	2.4	686	S_{65}(28),S49(18),S53(14),
58	677 s	694	17.2	0.2	675	S_{53}(28),S65(27),S70(10)
59		656	2.0	7.9	651	S_{64}(74)
60	640 w	647	4.2	5.5	638	S_{51}(38),S_{40}(10)
61	615 w	627	6.0	2.8	622	S_{52}(29),S_{40}(20),S_{73}(18)
62	581 vw	571	0.5	6.8	563	S_{57}(28),S_{51}(24),S_{63}(10)
63	540 s	553	8.9	0.2	537	S_{68}(23),S_{55}(15),S_{53}(15)
64	511 vw	518	2.2	5.6	512	S_{63}(21),S_{57}(18),S_{52}(13),S_{58}(13)
65		492	1.4	11.3	487	S_{58}(34),S_{57}(25),S_{7}(10)
66	471 w	492	1.6	1.1	480	S_{67}(24),S_{80}(22),S_{69}(14)
67		438	0.09	0.2	425	S_{54}(52),S_{55}(27)
68		417	0.04	0.06	403	S_{66}(60),S_{67}(19)
69		398	0.9	1.5	393	S_{81}(34),S_{72}(20)
70		335	3.0	0.6	327	S_{55}(33),S_{80}(17),S_{65}(11)
71		319	1.9	1.9	316	S_{81}(25),S_{40}(21),S_{58}(15)
72		280	1.2	1.3	276	S_{81}(25),S_{72}(16),S_{52}(10)
73		257	0.001	6.3	251	S_{55}(25),S_{69}(11),S_{60}(10)
74		248	0.07	0.8	243	S_{14}(25),S_{63}(24)
75		156	0.01	2.4	151	S_{60}(51),S_{54}(20)
76		146	1.3	0.5	142	S_{60}(35),S_{59}(28),S_{67}(12)
77		105	1.2	3.2	103	S_{61}(38),S_{59}(14),S_{67}(11),S_{55}(10)
78		93	0.4	0.3	93	S_{73}(43),S_{72}(34)
79		48	0.2	0.4	47	S_{71}(27),S_{69}(13),S_{60}(11)
80		41	1.0	2.2	40	S_{71}(54),S_{61}(12)
81		35	0.2	0.7	31	S_{79}(66)

[a] This work
[b] DFT/B3LYP/6-31G*
[c] Units are km.mol^{-1}
[d] Raman activities in Å^{4} (amu)$^{-1}$
[e] From scaled quantum mechanics force field

The strong band at 1461 cm^{-1} is assigned to the two antisymmetric CH_3 bending modes while the corresponding symmetrical mode is associated with the band of the medium intensity at 1381 cm^{-1} because these modes are predicted by calculations at 1450, 1458 and 1378 cm^{-1}; for this reason, they were assigned in this region. For MPB, the rocking and twisting modes related with this group are

clearly predicted in the expected regions [16, 18, 24, 26, 29, 43] thus, they were assigned in those regions, as observed in Table **8**.

Skeletal modes: The skeletal modes in the compound are predicted by SQM calculation strongly mixed among them, as observed in Table **9**. Here, in accordance with the values previously reported for molecules with similar rings [15, 16-19, 24-27, 43] and at our theoretical results, the strong IR bands at 1758, 1608 and 1598 cm^{-1} are respectively assigned to the C=O, C=C and C=N stretchings corresponding to the different rings, as indicated in Table **8**. Note that the C=O stretching modes for CPB (1768 cm^{-1}) and PB (1763 cm^{-1}) are assigned at higher wavenumbers than the corresponding to MPB (1759 cm^{-1}) while for the C=N stretching modes are assigned in the following order: PB (1692 cm^{-1}), CPB (1624 cm^{-1}) and MPB (1608 cm^{-1}). The C=C stretching modes for the three compounds are assigned approximately in the same region, as observed in Table **8**. The C-N stretching modes for CPB and MPB are predicted by calculations at 1254 cm^{-1} but they are assigned respectively at 1271 and 1259 cm^{-1}. The two C-O stretching modes corresponding to the oxazin rings in the three compared compounds are assigned approximately in the same region, thus, in PB that mode is assigned at 1027 cm^{-1}, in CPB are assigned in the 1050-990 cm^{-1} region and, in MPB are assigned between 1061 and 1009 cm^{-1}. On the other hand, the benzoxazin-methylphenyl stretching inter-rings are predicted at lower wavenumbers, it is, in MPB at 243 cm^{-1} while in CPB at 217 cm^{-1}, and for this reason, they were not assigned. Obviously, the differences observed in the wavenumbers show clearly the effect of exchanging the atom of chlorine or the methyl group on the phenyl ring linked to the benzoxazin ring. In general, all the stretching modes in PB are observed at higher wavenumbers than in CPB and MPB. Hence, the introduction of a Cl atom or CH_3 group in the phenyl ring linked to the benzoxazin ring generates a decrease in the wavenumbers corresponding to skeletal stretching modes, as can be seen in Table **8**. The remaining C-C stretching modes for MPB were assigned in accordance with assignments reported for similar compounds [15, 16-19, 24-27, 43] and at our theoretical calculations. The deformations and torsions corresponding to the phenyl, oxazin and methylphenyl rings of MPB were assigned taking into consideration their relative

position predicted by calculations and the assignments for molecules with similar rings [15, 16-19, 24-27, 43], as can be seen in Table **8**. The other skeletal modes expected for MPB were assigned as observed in Table **8**.

FORCE FIELD

For MPB, the scaled force fields at the two approximation levels were obtained employing the SQM methodology [33] with the MOLVIB program [39], then, the force constants were calculated at the same levels of theory, as was described previously in section computational details. The calculated force constants were expressed in terms of internal coordinates and their values can be seen in Table **10** together with those corresponding to CPB [15]. In general, the *f*(C=O) and *f*(C-N) force constants by using both basis sets have higher values in CPB than MPB, as it is expected because they are in accordance with the observed wavenumbers. Also, a similar variation is observed in the calculated *f*(C12-C16) force constants related to the stretching inter-rings. Thus, in MPB the value of that force constant is slightly higher (4.964 mdyn $Å^{-1}$) than the corresponding to CPB (4.948 mdyn $Å^{-1}$), in accordance with the observed wavenumbers, as can be seen in Table **10**. The same variations are observed by using both calculation levels.

Table 10: Comparison of scaled internal force constants for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one)

2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]			**2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]**		
Force constant	**B3LYP 6-31G***	**B3LYP 6-311++G****	**Force constant**	**B3LYP 6-31G***	**B3LYP 6-311++G****
f(C=O)	12.629	12.305	*f*(C=O)	12.685	12.385
f(C-N)	7.556	7.434	*f*(C-N)	7.575	7.458
f(C-O)	4.640	4.455	*f*(C-O)	4.639	4.450
f(C12-C16)	4.964	4.873	*f*(C12-C16)	4.948	4.858
f(C23-C26)	4.375	4.306	*f*(C-Cl)	3.304	3.256
$f(C\text{-}C)_{ring}$	6.341	6.209	$f(C\text{-}C)_{ring}$	6.358	6.230
$f(C\text{-}H)_{ring}$	5.200	5.124	*f*(C-H)	5.235	5.155
$f(C\text{-}H)_{methyl}$	4.846	4.779			

Units are mdyn $Å^{-1}$ for stretching and stretching/stretching interaction and mdyn Å rad^{-2} for angle deformations
[a]This work
[b]From Ref [15]

HOMO-LUMO Study

For MPB, the frontier molecular HOMO and LUMO orbitals were calculated by using both calculations levels and the values were compared with those calculated in this work for PB and with those reported for CPB [15]. The results are summarized in Table **11**. The results show that in the three compared molecules the HOMO-LUMO orbitals are mainly localized on the rings, demonstrating this way that the HOMO-LUMOs are mostly π-antibonding type orbitals. The energy gap values show the following relation: MPB > CPB > PB, indicating that the two first compounds are less reactive and have higher chemical hardness than PB while PB has a higher chemical reactivity than the other ones. Clearly, we observed that the introduction of a Cl atom or a CH_3 group in the phenyl ring linked to the benzoxazin ring increase the chemical hardness of the corresponding benzoxazin derivative. On the contrary, the lower HOMO value for the chlorinate derivative indicate a higher stability for this compound, in accordance with the NBO calculations. These results are very important especially when the derivatives are used as a pharmacologic compound.

Table 11: The frontier molecular HOMO and LUMO orbitals for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one and 2-Phenyl-4H-3,1-benzoxazin-4-one

Orbital	2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one[a]		Orbital	2-Phenyl-4H-3,1-benzoxazin-4-one[a]		Orbital	2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one[b]	
	6-31G*	[#]		6-31G*	[#]		6-31G*	[#]
HOMO (62)	-0.22797	-0.24081	HOMO (57)	-0.23369	-0.24772	HOMO (66)	- 0.23626	- 0.23650
LUMO (63)	-0.06613	-0.08047	LUMO (58)	-0.08772	-0.10609	LUMO (67)	- 0.07604	- 0.07521
GAP (a.u.)	-0.16184	-0.16034	GAP (a.u.)	-0.14597	-0.14163	GAP (a.u.)	- 0.16022	- 0.16129
GAP (eV)	-4.40392	-4.36311	GAP (eV)	-3.97208	-3.85398	GAP (eV)	- 4.35984	- 4.38895

[a]This work
[b]From Ref [15]
[#]6-311++G**

CONCLUSION

In this chapter, the theoretical molecular structures of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one were determined by using the hybrid B3LYP/6-31G* and B3LYP/6-311++G** methods. The stability of 2-(4-methylphenyl)-4H-3,1-

benzoxazin-4-one was studied by means of NBO, HOMO-LUMO and AIM studies. The charge values involved on the atoms belonging to the methylphenyl ring of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one decrease in relation to 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one. The reactivity predicted by using 6-31 basis set follow the trend: 2-Phenyl-4H-3,1-benzoxazin-4-one > 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one > 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one while the tendency change when the other basis set is used. On the other hand, the presence of a CH_3 group in the phenyl ring linked to the benzoxazin ring increases the HOMO energy value and, as consequence, the stability of 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one is lower than 2-Phenyl-4H-3,1-benzoxazin-4-one and 2-(4-chlorophenyl)-4H-3,1-benzoxazin-4-one, in accordance with the NBO results. Furthermore, the NBO and HOMO-LUMO studies reveal the strong dependence of the size of the basis set on the delocalization and frontier orbitals energies. A complete assignment of the 81 normal vibration modes for 2-(4-methylphenyl)-4H-3,1-benzoxazin-4-one was performed. In fact, all the bands observed in the vibrational spectra were completely assigned for which this derivative can be easily identified by means of vibrational spectroscopy. The SQM force fields by using the B3LYP/6-31G* and B3LYP/6-311++G** combinations were obtained together as a set of scaled force constants.

ACKNOWLEDGEMENTS

This work was supported with grants from CIUNT (Consejo de Investigaciones, Universidad Nacional de Tucumán). The authors thank Prof. Tom Sundius for his permission to use MOLVIB.

CONFLICT OF INTEREST

The authors confirm that this chapter content has no conflict of interest.

REFERENCES

[1] Khan, Z.A.; Raza Naqvi, S.A., Shahzad, S.A., Mahmood, N.; Yar, M.; Zahoor, A.F. Synthesis and Antimicrobial Activity of 2-Aryl-4H-3,1-benzoxazin-4-ones, *Asian J. Chem.*, **2013**, *25*(1), 152-156.

[2] Bouillant, M.L.; Favre-Bonvin, J.; Ricci, P. Dianthalexine, nouvelle phytoalexine, de type benzoxazinone, isolee de l'oillet dianthus caryophyllus l. (caryophyllacees). *Tetrahedron Lett*., **1983**, *24*(1), 51-52.

[3] Hauteville, M.; Ponchet, M.; Ricci, P.; Favre-Bonvin, J. Novel synthesis of dianthalexin (phytoalexin) analogues preparation. *J. Heterocycl.Chem*., **1988**, *25*(3), 715-718.

[4] El-Din, N.S. Synthesis and potential antibacterial activity of some 1,3-benzoxazin-4-one derivatives, *Acta Pharm.*, **2000**,*50*, 239-248.

[5] Shalaby, A.A.; El-Khamry, A.M.; Shiba, S.A.; Ahmed, A.A.; Hanafi, A.A. Synthesis and antifungal activity of some new quinazoline and benzoxazinone derivatives. *Arch. Pharm*.,**2000**, *333*, 365-372.

[6] Madkour, H.M.F. Reactivity of 4*H*-3,1-benzoxazin-4-ones towards nitrogen and carbon nucleophilic reagents: applications to the synthesis of new heterocycles. *ARKIVOC*, **2004**, *(i)*, 36-54.

[7] Rajeev, K.S.; Joseph, A.; Paia, A.; Srinivasan, K. K.; Kedar, T.; Thomas, A. T.; Jessy, E. M.; Singla, R. K. Synthesis of novel 3-8substituted 1,3,4-thiadiazol-2-yl)-quinazolin-4(3H)-ones as anticancer agents, *Orbital Electron. J. Chem. Campo Grande,* **2010**, *2*(2), 158-167.

[8] Mosad, S.M., Mohammed, K.I.; Ahmed, M.A.; Abdel-Hamide, S.G. New quinazoline related derivatives with antimicrobial activity. *Part 1 Pak. J. Biol. Sci*., **2004**, *7*, 1262-1268.

[9] Rai, L. *Synthesis of quinazoline analogues of biological interest*, PhD Master, Rajiv Gandhi University of Health Sciences, March **2010**.

[10] Selvam, P.; Vanitha, K.; Chandramohan, M.; De Clercq, E. Synthesis And Antimicrobial Activity Of Some Novel 6-Bromo-2-Methyl/phenyl-3-(Sulphonamido) quinazolin-4(3H)-Ones. *Indian J. Pharm. Scien*., **2004**, *66*(1), 82-86.

[11] Jatav, V.; Mishra, P.; Kashaw, S.; Stables, J. P. CNS depressant and anticonvulsant activities of some novel 3-[5-substituted 1,3,4-thiadiazole-2-yl]-2-styryl quinazoline-4(3H)-ones. *Eur. J. Med. Chem*., **2008**, *43*, 1945-1954.

[12] Venkataraman, S.; Meera, R.; Pandiarajan, Devi, P. Synthesis and biological activity of some novel quinazolinone derivatives. *J. Chem. Pharm. Res*., **2010**, *2*(5), 461-475.

[13] Altinkök, Ç. *Synthesis and characterization of benzoxazine-sulfones as precursors for thermally stable thermosets*. PhD Thesis, Istanbul Technical University, Institute of Science and Technology **2011**.

[14] Ambujakshan, K.R.; Tresa Varghese, H.; Mathew, S.; Ganguli, S.; Kumar Nanda, A.; Panicker, C.Y. Vibrational spectroscopic studies and theoretical calculations of 2-phenyl-4H-3,1-benzoxazin-4-one. *Oriental J. Chem*., **2008,** *24*(3), 865-874.

[15] Castillo, M.V.; Romano, E.; Argañaraz, G.R.; Rudyk, R.A.; Brandán, S.A. Theoretical Structural and Vibrational Investigation on the 2-(4-chlorophenyl) -4H-3,1-benzoxazin-4-one compound. In: *Structural Analysis and Modelling: Research and Development*; Liu, Y. Ed. Collection, Nova Science Publishers, Inc, **2012**; pp. 1-26.

[16] Brizuela, A.B.; Raschi, A.B.; Castillo, M.V.; Leyton, P.; Romano, E.; Brandán. S.A. Comparison between the structural and vibrational properties of the artificial sweetener sucralose with those obtained for sucrose. *Comp. Theoret. Chem*., **2013**, *1008*, 52–60.

[17] Romano, E.; Soria, N.A.J.; Rudyk, R.; Brandán, S.A. Structural and Vibrational Properties of a 1,3,4-Oxadiazole derivative with diverse biological activities. ISSN: 09719237. *J. Asia Spectros*. **2013**, *17*, 1-28.

[18] Romano, E.; Castillo, M.V.; Pergomet, J.L.; Zinczuk, J.; Brandán. S.A. Synthesis, structural study and vibrational spectra of (5-chloro-quinolin-8-yloxy) acetic acid. *Open Journal Synthesis, Theory and Applications*, **2013**, *2*, 8-22.

[19] Romano, E.; Brizuela, A.B.; Guzzetti, K.; Brandán, S. A. An experimental and theoretical study on the Hydration in aqueous medium of the antihypertensive agent tolazoline hydrochloride. ISSN: 0022-2860. *J. Molec. Struct*. **2013**, *1037*, 393-401.

[20] Romano, E.; Ladetto, F.; Brandán, S.A. Structural and Vibrational studies of the potential anticancer agent, 5-difluoromethyl-1,3,4-thiadiazole-2-amino by DFT calculations. *Comput. Theoret. Chem*. **2013**, *1011*, 57-64.

[21] Bichara, L.C.; Brandán, S.A. Hydration of species derived from ascorbic acid in aqueous solution. An experimental and theoretical study by using DFT calculation. *J. Molec. Liq*. **2013**,*181*, 34-43.

[22] Roldán, M.L.; Ledesma, A.E.; Raschi, A.B.; Castillo, M.V.; Romano, E.; Brandán, S.A. A new experimental and theoretical investigation on the structures of aminoethyl phosphonic acid in aqueous medium based on the vibrational spectra and DFT calculations. *J. Molec. Struct*. **2013**, *1041*, 73-81.

[23] Guzzetti, K.; Brizuela, A.B.; Romano, E.; Brandán, S.A. Structural and vibrational study on zwitterions of L-threonine in aqueous phase using the FT-Raman and SCRF calculations. *J. Molec. Struct*. **2013**, *1045*, 171-179.

[24] Lizarraga, E.; Romano, E.; Raschi, A.B.; Leyton, P.; Paipa, C.; Catalán, A.C.N.; Brandán, S.A. A structural and vibrational study of dehydrofukinone combining FTIR, FTRaman, UV–visible and NMR spectroscopies with DFT calculations. *J. Molec. Struct*. **2013**, *1048*, 331-338.

[25] Elida Romano, Noelia Anahi J. Soria, Roxana Rudyk, Silvia A. Brandán, Theoretical structural and vibrational study of a potential bioactive agent: the (5-pyridin-4-yl)-1,3,4-Oxadiazole-2-thiol compound. ISSN: 2319-2933. *Asian Materials Science Letters* **2013**, *2*, 1-14.

[26] Raschi, A.B.; Romano, E.; Castillo, M.V.; Leyton, P.; Paipa, C.; Maldonado, L.M.; Brandán, S.A. Vibrational study of a potential anticancerigen agent, Caffeic Acid Phenethyl Ester, by FTIR-FTRaman and NMR spectroscopies and, DFT calculation. *Vibrat. Spectros*. **2014**, *70*, 100-109.

[27] Márquez, M.B.; Brandán, S.A. A structural and vibrational investigation on the antiviral deoxyribonucleoside thymidine agent in gas and aqueous solution phases. *International J. of Quantum Chem*. **2014**, *114*(3), 209-221.

[28] Brizuela, A.B.; Castillo, M.V.; Raschi, A.B.; Davies, L.; Romano, E.; Brandán, S.A. A complete assignment of the vibrational spectra of sucrose in aqueous medium based on the SQM methodology and SCRF calculations. *Carbohydrate Res*. **2014**, *388*, 112-124.

[29] Chain, F.; Romano, E.; Leyton, P.; Paipa, C.; Catalán, C.A.N.; Fortuna, M.A.; Brandán, S.A. An experimental study of the structural and vibrational properties of sesquiterpene lactone cnicin using FT-IR, FT-Raman, UV-Visible and NMR spectroscopies. J. Mol. Struct. **2014**, 1065-1066, 160-169.

[30] Romano, E.; Castillo, M.V.; Márquez, M.B.; Gramajo, M.B.; Brandán, S.A. Physicochemical properties of the antihypertensive agent 2-benzylimidazoline in aqueous and physiological solutions at 25 and 37ºC.*J. Mol. Liq*. **2014**, *196*C, 285-290.

[31] Brizuela, A.B.; Raschi, A.B.; Castillo, M.V.; Davies, L.; Romano, E.; Brandán, S.A. Vibrational investigation on species derived from cyclamic acid in aqueous solution by using HATR and Raman spectroscopies and SCRF calculations, *J. Mol. Struct*. **2014**,*1074*,144-156.

[32] Romano, E.; Davies, L.; Brandán, S.A. Structural and Vibrational Studies and Molecular Force Field of zinc difluoromethanesulfinate. *J. Mol. Struct*. **2013**, *1044*, 144-151.

[33] Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, E. Combination of theoretical ab initio and experimental information to obtain reliable harmonic force constants. Scaled Quantum

Mechanical (SQM) force fields for glyoxal, acrolein, butadiene, formaldehyde, and ethylene. *J. Am. Chem. Soc*., **1979,** *101*(10), 2550.

[34] Reed, A.E.; Curtis, L.A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.* **1988**, *88*(6), 899-926.

[35] Glendening, E.D.; Badenhoop, J.K.; Reed, A.D.; Carpenter, J.E.; Weinhold, F. NBO 3.1; Theoretical Chemistry Institute, University of Wisconsin; Madison, WI, 1996.

[36] Bader, R.F.W. *Atoms in Molecules, A Quantum Theory*, Oxford University Press, Oxford, 1990, ISBN: 0198558651.

[37] Biegler-Köning, F.; Schönbohm, J.; Bayles, D. AIM2000; A Program to Analyze and Visualize Atoms in Molecules, *J. Comput. Chem*., **2001**, *22*, 545.

[38] *Gaussian 03, Revision B.01,*M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J. A. Jr. Montgomery, T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M.Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, Gaussian Inc*.,* Pittsburgh PA, **2003**.

[39] Sundius, T. Scaling of Ab-initio force fields by MOLVIB. *Vib. Spectrosc*., **2002**, *29*, 89-95.

[40] *Gauss View* 3.0, Version 2000–2003 User's Reference, GAUSSIAN Inc., Pittsburgh, PA,. A.B. Nielsen, A. J. Holder, **2003**.

[41] Thilagavathy, R.; Kavitha, H. P.; Arulmozhi, R.; Vennila, J. P.; Manivannan, V. 2-Phenyl-4*H*-3,1-benzoxazin-4-one, *Acta Cryst*., **2009**, *E65*, o127.

[42] Ambujakshan, K.R.; Tresa Varghese, H.; Mathew, S.; Ganguli, S.; Kumar Nanda, A.; Panicker, C.Y. Vibrational spectroscopic studies and theoretical calculations of 2-phenyl-4H-3,1-benzoxazin-4-one. *Oriental J. Chem.*, **2008**, *24*(3), 865-874.

[43] Piro, O.E., Echeverría, G.A.; Lizarraga, E.; Romano, E.; Catalán, C.A.N.; Brandán, S.A. Molecular structure of 4-hidroxy-3-(3-methyl-2-butenyl) acetophenone, a plant antifungal, by X-ray diffraction, DFT calculation, and NMR and FTIR spectroscopy. *Spectrochim. Acta A*, **2013**, *101*, 196-203.

[44] National Institute of Standards and Technology Chemistry WebBook. USA (http://Webbook.nist.gov/chemistry). [Accessed on March 2012].

CHAPTER 7

First Principles Computational Biochemistry with deMon2k

A. Alvarez-Ibarra[1,*], P. Calaminici[1], A. Goursot[2], C. Z. Gómez-Castro[1], R. Grande-Aztatzi[3], T. Mineva[2], D. R. Salahub[4], J. M. Vásquez-Pérez[1], A. Vela[1], B. Zuniga-Gutierrez[1] and A. M. Köster[1]

[1]Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México; [2]Institut Charles Gerhardt Montpellier, UMR5253 CNRS, ENSCM/UM2/UM1, 8 rue de l'Ecole Normale, 34296 Montpellier Cedex 5, France; [3]Departamento de Física Aplicada, Centro de Investigación y de Estudios Avanzados, Unidad Mérida, Km. 6 Antigua Carretera a Progreso, A.P. 73, Cordemex, Mérida 97310, Yucatán, México and [4]Department of Chemistry, Centre for Molecular Simulation and Institute for Quantum Science and Technology, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

Abstract: The growth of computational power, provided by new hardware technologies and the development of better theoretical methods and algorithms, allows more than ever an improvement in the reliability of computational predictions in medical sciences, along with a better understanding of the underlying molecular mechanisms. However, one limitation of computational chemistry approaches in the field of biological systems is the complexity of the molecules and the environment in which such molecules are to be studied. Important issues such as the determination of molecular properties which depend on the electronic structure face a considerable challenge when all-electron methodologies are required in the investigation. The most rigorous and sophisticated electronic structure methodologies, like density functional theory (DFT), are usually overwhelmed by the molecular size of most pharmacological targets. However, important implementations were recently achieved by the developers group of the computational chemistry code deMon2k. Knowing that the computation of electrostatic interaction integrals is an important bottleneck in all-electron calculations three new implementations have been worked out in order to eliminate such bottleneck. These implementations allow deMon2k now to explore biological and pharmacological systems in the framework of all-electron DFT methodologies.

Keywords: Asymptotic expansions, born-oppenheimer molecular dynamics, copper coordination, density functional theory, electron repulsion integrals, nuclear magnetic resonance, prion protein.

***Corresponding author A. Alvarez-Ibarra:** Departamento de Química, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Av. Instituto Politécnico Nacional 2508, A. P. 14-740 México, D.F. 07000, México; Tel: +52-5557473800, Ext: 4411; E-mail: aalvarezi@cinvestav.mx

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

1. INTRODUCTION

In theoretical chemistry, so-called *ab initio* (meaning *from first principles*) methods are based on the many-electron Schrödinger equation. The most elementary approximation for the wavefunction is obtained by the Hartree-Fock approach [1]. In this approximation the correlation between electrons of opposite spins is completely neglected. The systematic incorporation of electron correlation in wavefunction methods result in algorithms with computational costs that scale badly with the number of basis functions. Hence they become unsuitable for molecules with many hundreds of atoms. An interesting alternative to wavefunction methods was born from the ideas of Thomas [2] and Fermi [3]. They suggested the electronic density instead of the wavefunction as basic ingredient for electronic structure calculations. However, the Thomas-Fermi model failed to give quantitative results of the shell structure of atoms or the bonding in molecules [4-6] among other drawbacks. Nevertheless, it established an interesting alternative to wavefunction calculations. Further works by Dirac [7] as well as Wigner and Seitz [8, 9] improved the model by introducing a local expression for the exchange potential.

Several years later, Slater introduced the idea of approximating the Hartree-Fock exchange operator by an average local potential [10] based on the free-electron gas model. The result is an exchange potential expressed solely in terms of the electron density $\rho(\vec{r})$. Using a different derivation, Gaspar [11] obtained the same $\rho^{1/3}(\vec{r})$ form of Slater's expression but with a pre-factor of 2/3. This pre-factor was a topic of debate for many years. Further development by Slater and Johnson resulted in the so-called Xα methodology [12], a self-consistent field (SCF) methodology using the muffin-tin approximation and the multiple-scattering method [13, 14]. Even though this methodology was developed mainly for solid-state physics applications, it started to be applied to molecules, building a bridge between solid-state physics and chemistry [15].

In the sixties of the twentieth century, along with the Xα development, another theory arose from the ideas of Thomas and Fermi. With the formulation of the Hohenberg-Kohn theorems [16], a solid theoretical framework for density functional theory (DFT) was given. Although this turned DFT into an exact

theory, the missing knowledge of the form of the universal Hohenberg-Kohn functional hampered direct applications. In the pioneering paper of Kohn and Sham [17] in 1965 this obstacle was circumvented by the introduction of an artificial non-interacting reference system. Even though the here appearing exchange-correlation functional is also unknown, its approximation turned out to be much easier as for the Hohenberg-Kohn functional due to the reduced kinetic energy dependency (in fact, the kinetic energy contribution is often completely neglected in exchange-correlation functionals).

Computationally, the Kohn-Sham method resembles the well-studied Hartree-Fock method. The Kohn-Sham equations can be cast in a similar matrix form as the Roothan-Hall equations [18, 19]. Therefore, many algorithms from Hartree-Fock wavefunction calculations can be used in Kohn-Sham calculations. In particular, the linear combination of Gaussian type orbitals (LCGTO) approximation can be used in Kohn-Sham density functional theory calculations. As a result, many, but not all, molecular integrals can be either calculated analytically or reduced to the calculation of an incomplete gamma function [20]. In addition, the use of Gaussians as basis functions allows the implementation of recurrence relations for the molecular integrals, which permits the construction of efficient algorithms for their calculation.

Obviously, there are differences between the Hartree-Fock and the Kohn-Sham methods. The most noticeable one is the treatment of the non-classical interactions between electrons. While the Hartree-Fock method includes only exchange (which is treated exactly), Kohn-Sham calculations usually include both exchange and correlation but, in the vast majority of cases, with an approximate exchange-correlation functional. This term usually requires numerical integration techniques that are not needed in LCGTO Hartree-Fock methods. Even though linear scaling algorithms for numerical integration are well established they usually come with large prefactors and, therefore, represent computationally demanding tasks.

Since computational studies of biological systems have become much more frequent in the last decades and the computational power has grown at a considerable pace, the design and implementation of efficient algorithms to boost

the introduction of new studies on the quantum chemistry field and exploit the increased hardware performance is mandatory. This is of utmost importance for the kind of studies required for an accurate determination of molecular structures and properties, in which electronic structure calculations are the most reliable method to achieve such determinations. As part of the deMon2k developers group we present here a new methodology to improve the computational efficiency of the code, paving the way to enlarge system size and simulation time in *ab-initio* molecular simulations. Then, a couple of case studies are shown encouraging the importance of an efficient implementation of electronic structure methodologies: the search for a coordination model of the prion protein that matches experimental data and the importance of considering dynamic effects for the prediction of the chemical shifts of glycerol.

1.2. Auxiliary Density Functional Theory

The fundamental approach for all-electron first-principles quantum chemistry calculations in deMon2k is the so-called auxiliary density functional theory (ADFT), which uses fitted electronic densities to reduce the formal scaling of the standard Kohn-Sham methodology. It is based on the works by Baerends *et al.* [21], Sambe and Felton [22], Dunlap *et al.* [23] and Mintmire and Dunlap [24] who introduced the so-called variational fitting of the Coulomb potential (VFCP) to reduce the computational cost of the calculation of Coulomb interactions between electrons. In deMon2k different Gaussian type functions are used to expand the Kohn-Sham density and the fitted density. For the Kohn-Sham density and orbitals Cartesian Gaussian type functions, also refered as Gaussian type orbitals (GTOs) are used. They are defined as,

$$a(\vec{r}) = \left(x - A_x\right)^{a_x} \left(y - A_y\right)^{a_y} \left(z - A_z\right)^{a_z} \sum_{k}^{K_a} d_k e^{-\zeta_k \left(\vec{r} - \vec{A}\right)^2} , \tag{1}$$

where $\vec{r} = \left(x, y, z\right)$ are the electron coordinates, $\vec{A} = \left(A_x, A_y, A_z\right)$ are the function center coordinates (which coincide with a nuclear center), ζ_k is the exponent which defines the function extension, d_k are contraction coefficients of a contraction degree K_a and (a_x, a_y, a_z) is a set of non-negative integer numbers related to the angular momentum index of the represented orbital *i.e.* the set

(0,0,0) represents a *s* function, (1,0,0) represents a p_x function, and so on. The basis functions, denoted by *a* or *b*, represent orbitals and define the electronic density as follows (for simplicity, all formulas refers to closed-shell systems),

$$\rho(\vec{r}) = \sum_{a,b} P_{ab} a(\vec{r}) b(\vec{r}), \tag{2}$$

where P_{ab} is the density matrix. For the expansion of the fitted density Hermite Gaussian type functions are used in deMon2k [25, 26]. They are defined as,

$$\overline{c}(\vec{r}) = \left(\frac{\partial}{\partial C_x}\right)^{\overline{c}_x} \left(\frac{\partial}{\partial C_y}\right)^{\overline{c}_y} \left(\frac{\partial}{\partial C_z}\right)^{\overline{c}_z} e^{-\zeta_c \left(\vec{r}-\vec{C}\right)^2}, \tag{3}$$

where $\vec{C} = \left(C_x, C_y, C_z\right)$ are the function center coordinates (which coincide with a nuclear center), ζ_c is the exponent which defines the function extension and $\left(\overline{c}_x, \overline{c}_y, \overline{c}_z\right)$ is a set of non-negative integer numbers related to the angular momentum index of the function. These functions, denoted $\overline{c}$ and $\overline{d}$, are used to expand the auxiliary electronic density $\tilde{\rho}(\vec{r})$ through the following linear combination,

$$\tilde{\rho}\left(\vec{r}\right) = \sum_{\overline{c}} x_{\overline{c}} \overline{c}\left(\vec{r}\right), \tag{4}$$

where $x_{\overline{c}}$ are the corresponding density coefficients. Recently, the use of the auxiliary electronic density from the VFCP has been extended to the calculation of the exchange correlation energy [27], giving birth to the ADFT implementation in deMon2k. Even though the numerical integration of exchange-correlation energy and potential are not avoided in ADFT, the use of the auxiliary density reduces the computational effort substantially. In combination with efficient adaptive grid techniques [28] as implemented in the current version of deMon2k [29], the numerical integration of the ADFT exchange-correlation energy and potential is one order of magnitude faster than the corresponding calculation of the analytical three-center electron repulsion integrals. Thus, by its construction, ADFT permits large-scale parallel DFT calculations in reasonable times with the typical accuracy and reliability of *ab initio* methods. As a result, the

computational effort in the calculation of exchange-correlation terms is reduced. The consequence is that the Coulomb interaction calculation becomes once again the computational bottleneck in the ADFT Kohn-Sham methodology.

In a closed-shell system, the ADFT energy is given by [27],

$$E = \sum_{a,b} P_{ab} H_{ab} + \sum_{a,b} \sum_{\bar{c}} P_{ab} \left\langle ab \middle\| \bar{c} \right\rangle x_{\bar{c}} - \frac{1}{2} \sum_{\bar{c},\bar{d}} x_{\bar{c}} x_{\bar{d}} \left\langle \bar{c} \middle\| \bar{d} \right\rangle + E_{xc}\left[\tilde{\rho}\right], \tag{5}$$

where H_{ab} is the core Hamiltonian matrix which contains all the one-electron quantities (such as electron kinetic energy and nuclear-attraction integrals) and $E_{xc}\left[\tilde{\rho}\right]$ is the exchange-correlation energy, which depends on the auxiliary electronic density. The quantities in brackets represent interaction integrals between electrons, and they are discussed in more detail in the following section.

2. THE ELECTRON REPULSION INTEGRALS BOTTLENECK

The quantity of interest in this discussion is the three-center electron repulsion integral (ERI)

$$\left\langle ab \middle\| \bar{c} \right\rangle = \iint \frac{a(\vec{r}_1) b(\vec{r}_1) \bar{c}(\vec{r}_2)}{\left| \vec{r}_1 - \vec{r}_2 \right|} \mathrm{d}\vec{r}_1 \mathrm{d}\vec{r}_2 \,. \tag{6}$$

In the iterative ADFT self-consistent field (SCF) procedure to calculate the molecular energy these ERIs are required twice in every step. First for the determination of the density fitting coefficients in the VFCP using the equation

$$x_{\bar{c}} = \sum_{\bar{d}} \sum_{a,b} G^{-1}_{\bar{c}\bar{d}} \left\langle ab \middle\| \bar{d} \right\rangle P_{ab} \,, \tag{7}$$

where $G^{-1}_{\bar{c}\bar{d}}$ is an element of the inverse of the so-called Coulomb matrix $G_{\bar{c}\bar{d}} = \left\langle \bar{c} \middle\| \bar{d} \right\rangle$ with elements

$$\left\langle \bar{c} \middle\| \bar{d} \right\rangle = \iint \frac{\bar{c}(\vec{r}_1) \bar{d}(\vec{r}_2)}{\left| \vec{r}_1 - \vec{r}_2 \right|} \mathrm{d}\vec{r}_1 \mathrm{d}\vec{r}_2 \,. \tag{8}$$

Second, in the calculation of the ADFT Kohn-Sham matrix,

$$K_{ab} = H_{ab} + \sum_{\bar{c}} \langle ab||\bar{c}\rangle \left(x_{\bar{c}} + z_{\bar{c}}\right), \tag{9}$$

where $z_{\bar{c}}$ represents the exchange-correlation coefficients,

$$z_{\bar{c}} = \sum_{\bar{d}} G^{-1}_{\bar{c}\bar{d}} \left\langle \bar{d} \middle| v_{xc}[\tilde{\rho}] \right\rangle . \tag{10}$$

Here v_{xc} denotes the exchange-correlation potential. In large-scale calculations, the ERI calculation becomes a bottleneck due to the combination of two issues:

1.- The computation of ERIs require the calculation of a non-analytical integral known as the Boys function or incomplete gamma function [30],

$$F_n(T) = \int_0^1 t^{2n} e^{-Tt^2} \mathrm{d}t \, , \tag{11}$$

that in turn requires tabulation and interpolation to obtain the appropriate values to use in the basic ERI calculation. Basic ERIs are the starting point of a sequence of recurrence relations that finally obtain the desired ERI [31, 32].

2.- The number of ERIs grows rapidly with system size. In the ADFT implementation, the scaling of ERIs is formally $N \bullet M$, where N is the number of basis functions and M is the number of auxiliary functions. Although linear scaling can be achieved using integral screening techniques [32] this is only possible when the number of atoms is huge and the system is widely extended in space. Thus, large molecular systems that do not reach the conditions for taking advantage of linear scaling methods will suffer an ERI bottleneck due to a large number of integrals to compute.

Although a molecular system is not large enough to present an ERI bottleneck due to the amount of integrals, it can present it due to the number of times ERIs have to be calculated. Good examples for this situation arise in first-principle molecular dynamics (MD) simulations. Born-Oppenheimer molecular dynamics (BOMD) requires repeated self-consistent field and molecular gradient calculations, which of

course include ERI computations. As an example, we present the calculation of a relatively small sodium cluster, Na_{55}^{+} (Fig. **1**). Performing a small MD of only 200 steps (1 fs each) as a test, we find that ERI calculation takes more than half of the total time, while other time consuming steps are matrix operations (*i.e.* diagonalization and multiplication) and exchange-correlation interaction calculation.

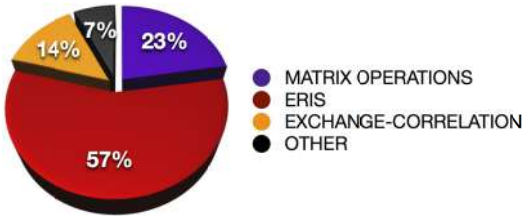


Figure 1: Time distribution in Na_{55}^{+} cluster BOMD calculation.

2.1. Solving the ERI Bottleneck

One way to reduce the computational cost of ERIs in large molecular systems is the use of multipole expansions for the Coulomb operator $1/|\vec{r}_1 - \vec{r}_2|$. Multipole expansions are only reliable if charge density centers are far away from each other (Fig. **2**) [26]. ERIs of this cateogry are called *far-field*, whereas ERIs that do not fulfill this condition are called *near-field*.

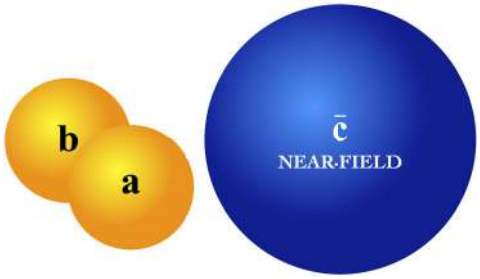


Figure 2: Far-field ERI definition. Since both orbitals (shown as yellow spheres) are outside the defined near-field of the auxiliary function (shown as a blue sphere), the ERI can be calculated using multipole series (see [26] for details).

A breakthrough algorithm for the handling of long-range interactions came with the introduction and development of the Fast Multipole Method (FMM) by Greengard and Rokhlin [33-36]. Several methods have been created in order to adapt the FMM for electronic structure calculations [37-39]. However, the

approach designed and implemented in deMon2k is different from the FMM since systematic space division is avoided. This implementation is named double asymptotic expansion of ERIs [40, 41].

$$\left\langle ab\middle\|\overline{c}\right\rangle^{A} \simeq \left(\frac{\pi}{\zeta_{\overline{c}}}\right)^{3/2} \sum_{m_x,m_y,m_z} \frac{(-1)^{M}}{m_x!m_y!m_z!} T_{AC}(m+c)\left\langle a+m\middle|b\right\rangle, \tag{12}$$

where

$$T_{AC}(m+c) = \left(\frac{\partial}{\partial C_x}\right)^{m_x+\overline{c}_x}\left(\frac{\partial}{\partial C_y}\right)^{m_y+\overline{c}_y}\left(\frac{\partial}{\partial C_z}\right)^{m_z+\overline{c}_z} \frac{1}{\left|\vec{A}-\vec{C}\right|}, \tag{13}$$

m=(m_x, m_y, m_z) are the expansion indices that sum up to M, *i.e.* $M = m_x + m_y + m_z$. In this example, the expansion center for the series is on A, the center of orbital a. This expansion transforms the two-electron ERI $\left\langle ab\middle\|\overline{c}\right\rangle$ into a series of overlap integrals $\left\langle a\middle|b\right\rangle$ with prefactors $T_{AC}(m+c)$ that depend only on atomic coordinates. When a sufficient number of integrals occur in the molecular system, the powerful factorization given by the prefactors and the efficient calculation of the overlap integrals [42] allow a significant reduction in the calculation of far-field ERIs (see Fig. **3**).

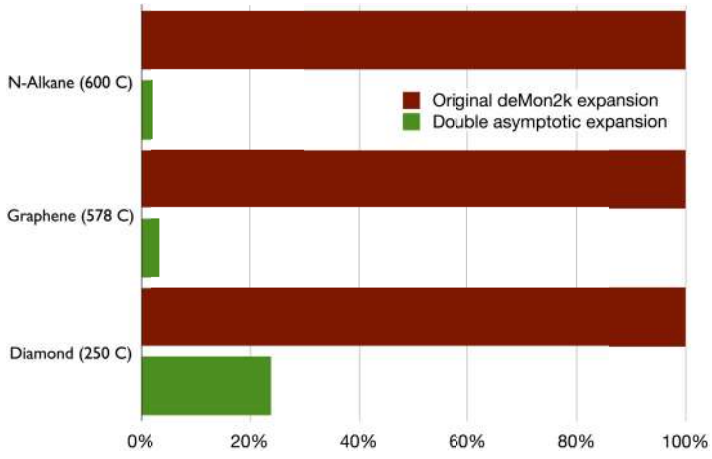


Figure 3: Relative far-field ERIs timing for selected carbon systems: 1D n-alkane, 2D graphene sheet and 3D diamond block. Red bars represent the original deMon2k expansion. Green bars represent the new double asymptotic expansion.

When the double asymptotic expansion is applied, the ERI calculation time is completely dominated by near-field ERIs even when they are fewer in number (see Fig. **4**). Thus, the next implementation step has to solve the near-field ERI bottleneck. Since no approximation shall be done in this kind of ERIs, an interesting alternative is to store them in memory. The so-called mixed SCF scheme [41], which combines the direct [32] and conventional [43] SCF schemes, calculates far-field ERIs in every SCF step using the efficient double asymptotic expansion whereas near-field ERIs are stored in RAM so they have to be read instead of calculated in every SCF step. This implementation has proven to eliminate the ERI bottleneck in the ADFT implementation of deMon2k (see Fig. **4**). The mixed SCF is specially useful in BOMD calculations of molecules and clusters with dozens of atoms since they may reduce the total calculation time to half of the original one (see Fig. **5**). The use of this powerful implementation allows faster calculations with the same accuracy as the original DFT code and, thus, facilitates deMon2k applications in the field of biochemistry.

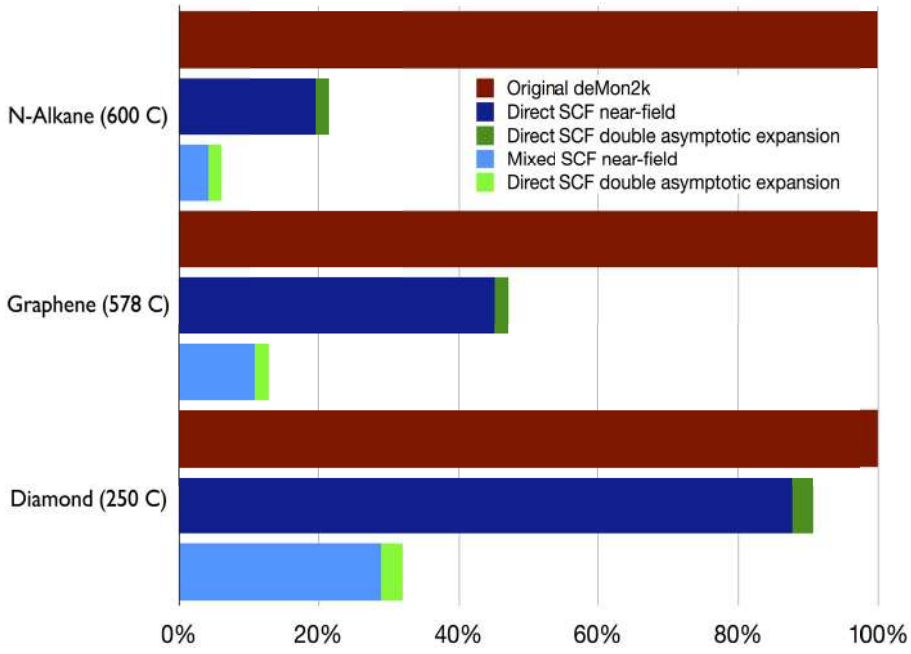


Figure 4: Relative total ERI timing for selected carbon systems: 1D n-alkane, 2D graphene sheet and 3D diamond block. Red bars represent the original deMon2k algorithm. Dark blue and dark green bars represent the direc SCF scheme using the double asymptotic expansion. Light blue and light green bars represent the new mixed SCF scheme.

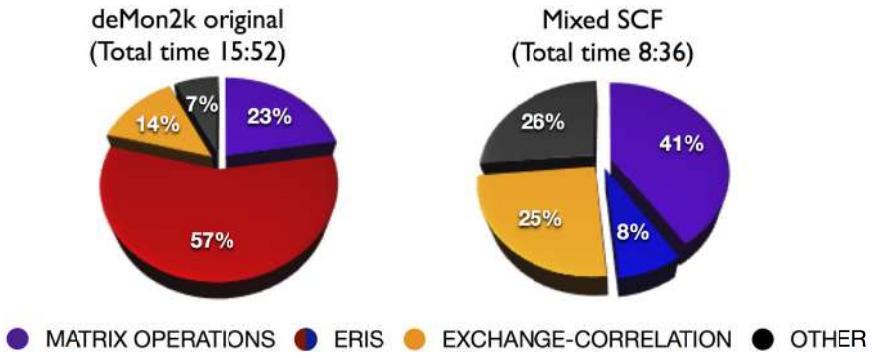


Figure 5: Time distribution in Na_{55}^{+} cluster MD calculation using the original deMon2k implementation (left) and the new mixed SCF implementation (right).

3. COPPER COORDINATION MODES IN THE HUMAN PRION PROTEIN: THE SECOND AND THIRD SITES

In the last three decades, the cellular prion protein (PrP^{C}) has acquired a major interest, due to its ability to coordinate copper ions, a process which may affect its physiological functions and the pathogenesis of the scrapie isoform (PrP^{Sc}) [44]. This isoform is responsible for neurodegenerative diseases known as transmissible spongiform encephalopaties (TSEs). Some examples of TSEs are mad cow disease and Creutzfeldt-Jakob disease in humans. These TSEs result from structural changes of the PrP^{C} keeping the encoded amino acid sequence to produce the pathogenic agent PrP^{Sc} [45-47]. This agent may act as a templete to refold the cellular prion protein into the scrapie isoform [48, 49]. On the other hand, the function of the celluar isoform has not been completely unraveled, but due to bound copper ions in the N-terminal region [50, 51], several studies have proposed that the PrP^{C} is involved in copper homeostasis or uptake into the cell [52, 53], copper-dependent antioxidant processes such as those with superoxide dismutase, signal trasduction, cell-cell adhesion, modulation of NMDA receptor activity and as an antiapoptosis agent [54-59].

The C-terminal domain of the PrP^{C} is formed predominantly by α-helices, while the N-terminal domain has no structure [60]. The N-terminal domain can bind six copper ions [61, 62], four of them in the octarepeat region, which contains the

highly conserved octapeptide PHGGGWGQ (residues 60-91 in the human sequence), while the other two copper ions bind in the region spanning residues 92-96 and 106-113 to the His96 and His111 [61-65]. Copper binds to each of the three coordination sites with different coordination modes, depending on pH and the relative concentration of copper in the protein [62, 64, 66, 67]. The octarepeat region has been the most studied binding site and, in fact, only a small copper complex of this region has been crystallized [68]. Also, theoretical studies have focused on the coordination of the copper ion in the octarepeat region using this structure as a model [44] while the electronic structure of the His111 and His96 copper binding sites have been relatively unexplored in theoretical studies [69-72]. Therefore, the lack of a crystal structure for the PrP(92-96) and PrP(106-113) binding sites makes the construction of appropriate structural models challenging and validation of different coordination models is crucial in order to compare experimental data with computed spectroscopic parameters [73, 74].

3.1. Methodology

The study of the coordination modes to the PrP(92-96) and PrP(106-113) binding sites was done using the GGGTH amino acid sequence for the former and KTNMKHMA sequence for the latter peptide fragment, both with the N-terminal acetylated and the C-terminal amidated. Employing the deMon2k code [29], all models were optimized without any geometry constraints using the nonempirical generalized gradient approximation (GGA) exchange-correlation functional PBE [75], the double-ζ with polarization (DZVP) [76] orbital basis set and the GEN-A2 auxiliary basis set [77, 78]. The complexes without explicit water molecules (*vide infra*) were characterized as stationary points by a frequency analysis at the same level of theory. Table **1** presents the technical details of the smallest copper-peptide complexes of both systems. The optimized structures were used to compute the electron paramagnetic resonance (EPR) parameters, g and A tensors, and in some cases the UV-Vis absorption and circular dichroism (CD) spectra were simulated using time-dependent density functional theory; all spectroscopic data were calculated with the ORCA program [79] and compared with experimental data [73, 74].

Table 1: Technical details of the smallest Cu(II)-PrP(92-96) and Cu(II)-PrP(106-113) complexes

System	Cu(II)-PrP(92-96)	Cu(II)-PrP(106-113)
Number of atoms	59	140
Number of electrons	275	563
Number of orbital	813	1409
Number of auxiliary functions	1608	2674

3.2. Results

For the copper coordination to the PrP(92-96) fragment, four different coordination modes have been proposed experimentally in a pH-dependent process, which vary the number of deprotonated amides of the backbone [80, 81], as shown in Fig. **6**. In the case of the coordination with four nitrogen atoms (4N coordination mode) no problems were found since the calculated data shows an excellent agreement with experiment in solution [73]. However, the coordination with three nitrogen atoms and one oxygen (3NO coordination mode) displays poor agreement with the experimental data. The way to achieve a better agreement was through adding water molecules to the solvation sphere of the copper ion and thus testing several structures, as shown in Fig. **7**. The explicit solvation in the coordination site plays an important role, structural as well as electronic. In summary, after adding water molecules we can group the structures in two blocks, one with no explicit water molecules binding to the metal ion and another with explicit water molecules binding to the metal ion. In both cases the copper ion has a 3NO equatorial coordination mode. These structural changes are reflected in the CD and UV-Vis spectra, as shown Fig. **8**, since the structures with a 3NO coordination mode with the oxygen atom from a carbonyl group, yield the same pattern of the experimental CD spectrum (B in Fig. **8**); while the participation of water molecules in the copper coordination gives a different absorption pattern (D in Fig. **8**). In this way it was possible to establish a structural model that reflects the experimental solution at pH values of 7 [73]. Finally, the characterization of the other two coordination modes, at low pH values (Fig. **6**), was done with a few structures yielding also good matches with experiment [73].

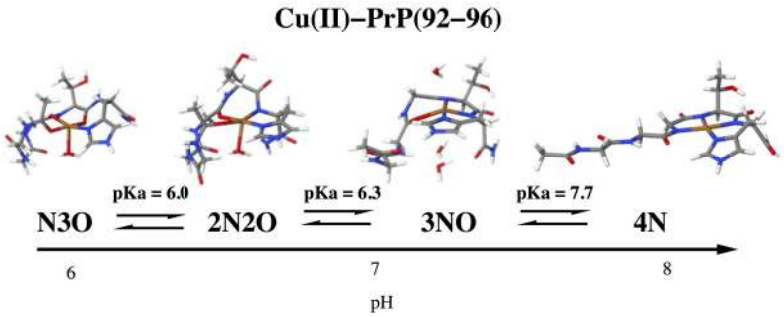


Figure 6: Coordination modes of the Cu(II)-PrP(92-96) complex depending on pH.

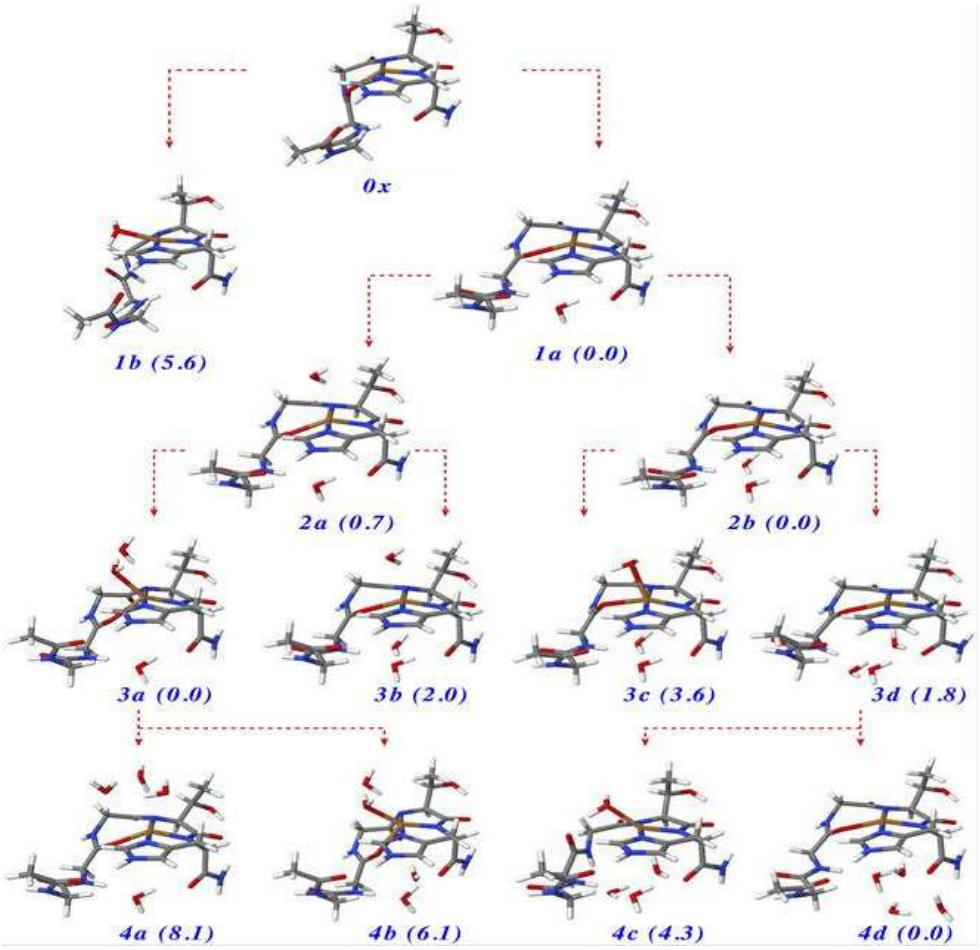


Figure 7: Optimized structures using the PBE functional of the Cu(II)−PrP(92−96) complex with 3NO coordination and with different numbers (0−4) of explicit water molecules. The numbers in parentheses are the relative energies in kcal/mol within each series.

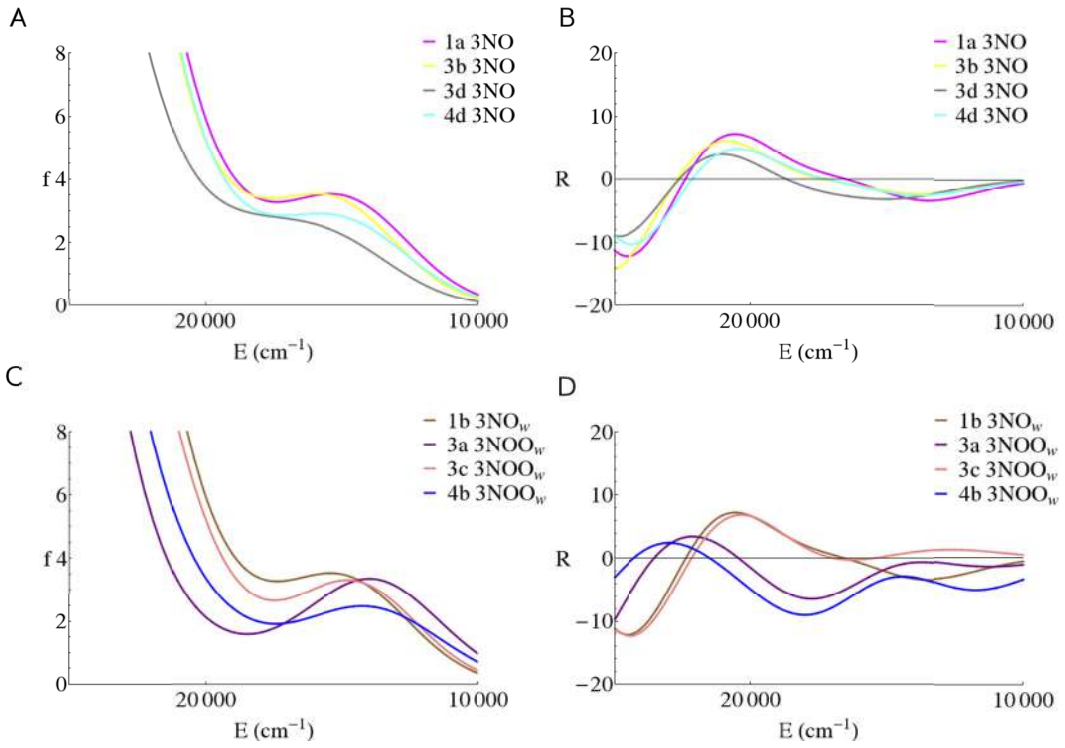


Figure 8: Computed UV-Vis absorption (A, C) and CD (B, D) spectra of the different 3NO coordination modes for the Cu(II)-PrP(92-96) complex with explicit water molecules.

For the study of the copper coordination to the PrP(106-113) fragment, besides the coordination modes with nitrogen and oxygen atoms, the presence of methionine residues in the sequence gives the possibility of sulfur atoms coordinated to the metal ion as previously proposed [71, 82-85]. In this case, the system size is larger than in the previous examples and the study of all relevant combinations was necessary to elucidate the coordination mode present in solution at physiological pH values. For this system, the 4N, 4NS, 3NO, 3NS, 3NOS, 2N2O, 2NOS and 2N2OS coordination modes were tested as initial models. Depending on the number of binding nitrogen atoms we can divide the models into three groups. For the group with four nitrogen atoms, the computed results match with the experimental ones collected in solution at pH 8.0 when the copper ion acquires a 4N equatorial coordination mode without the participation of axial ligands or with a very weak axial interaction with a sulfur atom located more than 4 Å away from the metallic center after structure optimization [74]. A second group collects the combinations

with three binding nitrogen atoms shown in Fig. **9**, where selected structures with different coordination modes of the copper atom are compared. We observed that the energy difference between the 3NO and 3NS coordination modes is small when the microsolvation is not taken into account, and therefore we cannot discard some of them from participating in solution. However, when explicit water molecules are

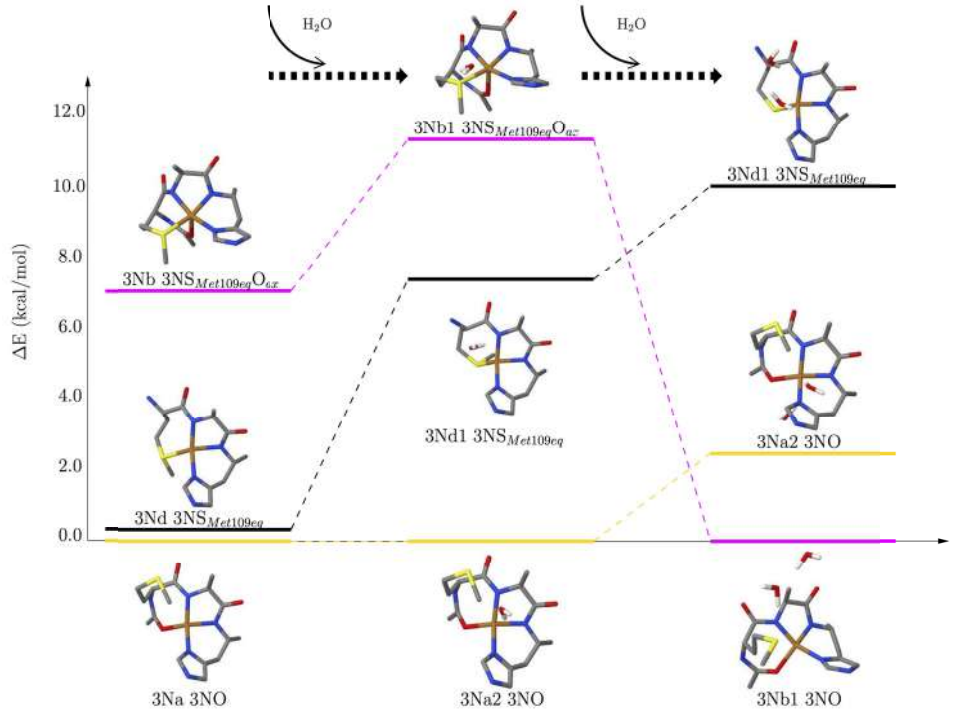


Figure 9: Energy diagram for selected structures of the 3N models of the Cu(II)-PrP(106-113) complex without water (left) and after the successive addition of two explicit water molecules. Most side chains are not shown for clarity; however, the geometry optimizations were done with the complete Cu(II)-PrP(106-113) complex.

incorporated the gap between the coordinaton modes increases *i.e.* the structures with sulfur atoms coordinated equatorially to the copper ion are destabilized by the presence of water molecules. These structure models match better with the experimental data collected at pH 7.0. The agreement reached with these models leads to the conclusion that the participation of sulfur atoms as equatorial ligands in the copper coordination is avoided by the solvent effects in the coordination sphere and that at physiological pH the 3NO coordination mode is preferred, which could

accommodate a sulfur atom as weak axial ligand [74]. The third group includes structures with two binding nitrogen atoms and its results are similar to those described previously *i.e.* without explicit water molecules there is a competition between the 2NOS and 2N2O coordination modes. But when explicit solvent molecules are taken into account, such competition disappears and the copper atom will be preferably coordinated with nitrogen and oxygen atoms on the backbone, as well as with the surrounding water molecules [74].

CONCLUSION

Experimentally, it has been observed that copper coordination to the prion protein is a pH-dependent process. In this study, different coordination modes of copper binding to PrP(92-96) and PrP(106-113) were evaluated and, in both cases, the preferred equatorial coordination mode at pH 8.5 occurs with four nitrogen atoms, one from imidazole ring and three from deprotonated amides from backbone. At pH 7.0 one oxygen atom is coordinated instead of one deprotonated amide. The presence of explicit water molecules in the coordination sphere is an important factor in the determination of coordination modes since explicit solvation helps to improve the agreement with experimental values of UV-Vis absorption, CD and EPR. In the case of the PrP(106-113) fragment, the equatorial coordination of sulfur atoms from methionines is discarded since those structures are destabilized by the presence of water molecules and the poor agreement with the experimental data. Finally, structural models for the copper coordination to the prion protein at lower pH values were studied, showing that the metal ion is mainly bound to oxygen based ligands.

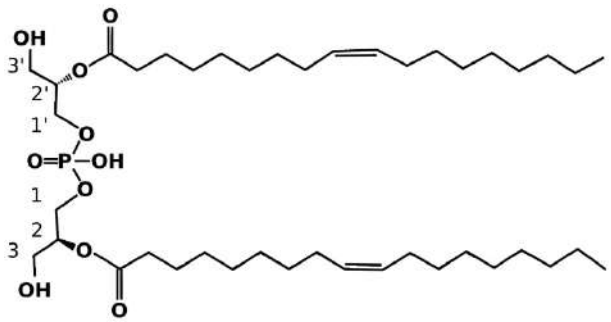


Figure 10: Schematic representation of LBPA with stereospecific numbering (sn) of the glycerol carbons.

4. DFT-BOMD STUDY OF HIGH ENERGY CONFORMATIONS OF GLYCEROL AND THEIR CONTRIBUTION TO NMR CHEMICAL SHIFTS

This section is adapted from [Ref. 86] by permission of the PCCP Owner Societies: http://pubs.rsc.org/en/content/articlelanding/2013/cp/c2cp43514d.

Glycerol is the starting point for the synthesis of triglycerides and phospholipids in living organisms. It has been studied using a wide variety of techniques in solid, liquid and gas phases [87-97]. It has been shown that glycerol conformers are the precursors of the glycerophosphatidyl choline lipid framework conformers [98-100]. In addition, it is accepted that quantum mechanical (QM) calculations are necessary for a quantitative evaluation of the molecular energy of lipid structures, the determination of possible conformers [99] and for the calculation of spectroscopic properties such as NMR, as shown below.

Interestingly, the most stable conformations of the isolated glycerol molecule and of the glycerol group in glycerophosphatidyl choline lipids are similar. It is also the case for the most stable structures of the bis(monoacylglycero)-phosphate lipid (formerly known as lysobisphosphatidic acid or LBPA) formed by two glycerol molecules bound to one phosphate group. This lipid is found in a small amount in all animal tissues (≤1%) (see Fig. **10**). Unlike other phospholipids LBPA is only found in late endosomes where it constitutes approximately 17% mol of the membrane [101, 102] but its function is still a topic of research [101, 103]. The 2,2'-dioleoyl isoform of LBPA spontaneously arranges in multivesicular liposomes if the liposome lumen has the same pH as the *in vivo* endosome forming structures similar to it [104]. Thus, it has been suggested that the phosphodiester functional group is attached to positions sn-1 and sn-1' instead of sn-3 which is a very different stereochemistry than the one observed in other animal glycerophospholipids (see Fig. **10**). The *in vivo* synthesis of most glycerophospholipids proceeds, indeed, through the phosphorylation of glycerol in position 3, following the conventional stereospecific numbering (sn) [105].

This intriguing problem was tackled by calculating and analysing the structures and energies of the constituting blocks of the LBPA and its isomers, namely the

glycerol and glycerophosphate molecules [98]. This systematic study determined a preferred bis(glycerol)phosphate isomer determined by the preponderant glycerol conformer. In the case of the full LBPA molecule (including the oleoyl chains), the positions of the acyl chains determine different structures for the 2,2' and 3,3' LBPA isomers, but for both types the most stable conformations of LBPA glycerols are the same as in the isolated glycerol molecule (glycerol B ($\alpha\gamma$) for the 2,2’ and glycerol F ($\gamma\gamma$) for the 3,3’). These various studies show how glycerolipid structures are strongly correlated with the structure and energetic properties of the glycerol precursor.

NMR has been widely used by chemists and biochemists and, after the development of magic angle spinning (MAS), by solid state scientists to obtain spectroscopic fingerprints of atoms in different environments [106-109]. Consequently, NMR spectroscopy provides a sensitive probe for the electronic environment of nuclei in molecules. Correlations with bond/torsion angles, the presence of intra- or inter-molecular hydrogen bonds and polarization by the medium, to name only a few effects, have been established in a large variety of systems [101, 110-115]. In the case of complex systems, in particular biomolecules, NMR has been used for a long time to gather structural information at the atomic level.

However, it is only recently that structure changes or fluctuations with time have been studied based on NMR experiments. Indeed, most biosystems exist as conformers in dynamic equilibria, and their functional properties cannot be described using only a static structure as determined by X-ray crystallography or even by a set of structures obtained from experimental nuclear magnetic resonance (NMR) constraints (taking into account the lower and upper limits of interatomic distances evaluated through experiment) [116]. For example, the existence and interconversion of three rotational isomers of phosphatidylcholine (PC) lipids has been proposed from NMR studies [117] where the rigidity of the glycerol backbone is utilized to interpret the NMR dipolar coupling data [118]. Also conformational fluctuations of proteins involving multiple conformers of different energies are now studied using NMR spectroscopy at various pressures [119, 120]. This new methodology provides different time-averaged NMR signals

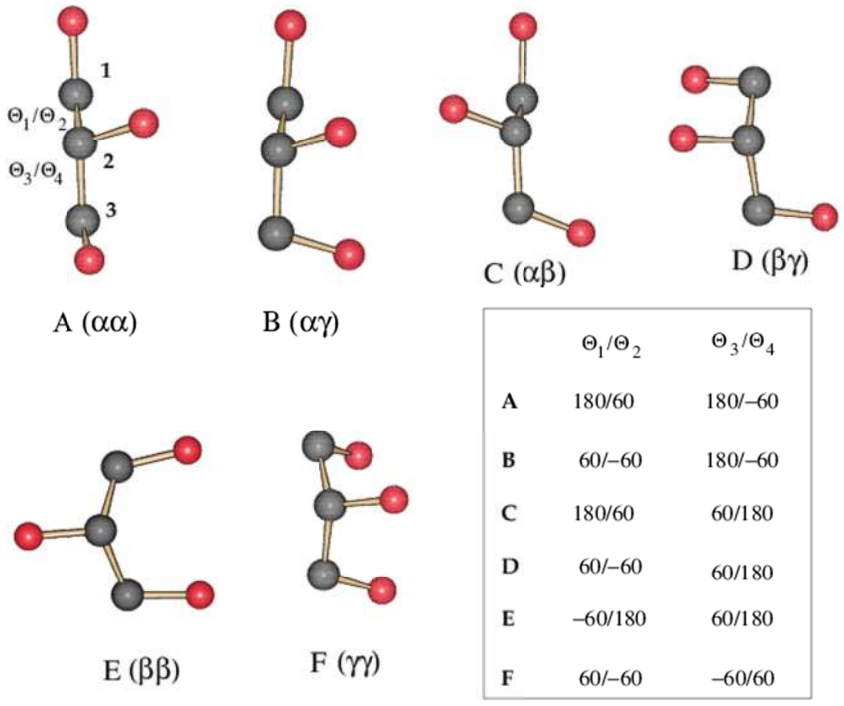
with different pressures for structural changes within about 1 ms to 1 s, allowing a realistic description of low frequency protein motions.

Underlining the importance of taking into account structural fluctuations, ranging from vibrations to conformational changes, NMR chemical shift calculations for molecules and solids have been combined recently with MD simulations for analyzing fast molecular motions, in particular those related to intramolecular changes. Although low frequency molecular motions are not yet within the reach of ab-inito BOMD, thus classical MD simulations have been used to study them [121]. In these recent studies, MD simulations are based on force field, semi-empirical or *ab initio* methods. Among them, one can distinguish two different approaches: (i) use MD to generate the various possible conformations of the system and compare the obtained structures with those which have been proposed on the basis of experimental chemical shifts or coupling constants [122-124]; (ii) perform MD trajectories at a chosen temperature, extract a series of configurations (snapshots) at given time intervals for which the desired NMR parameters are calculated [125-130]. If the number of snapshots is large enough, averaging the calculated NMR values leads to a good representation of the experimental NMR spectrum. In addition, this method allows a quantitative description of the main conformational contributions to the system dynamics.

The latter approach has been adopted to study the time evolution of the structure of glycerol and to analyze the contribution of its internal motions to the ^{13}C NMR spectrum [86]. Thus, BOMD has been applied with forces calculated from density functional theory (DFT) calculating multiple trajectories to sample the configuration space of an isolated glycerol molecule. As stated before, glycerol is the *in vivo* first precursor of phospholipids and triglycerides and its rich conformational space is a determining factor for its efficient conduction in the channels of transport proteins or phosphorylation enzymes where it has been located by X-ray crystallography [131-133].

The results were compared with experimental spectra of solutions of glycerol measured at room temperature as well as at 50 °C and 80 °C. In this range of temperatures, which can be explored by magnetic measurements, glycerol is in the liquid state. Large chemical shift fluctuations characterize the sampling of the

molecular vibrational states, showing their major contribution to this magnetic property, as has also been reported for small peptide models [127-134]. Combining these chemical shift results with a structural analysis of the BOMD trajectories in terms of minima and transition states of the potential energy surface (PES) leads to two main conclusions: (i) the non-random conformer interchange dynamics (successive rotations of the primary alcohol groups) reveals the importance of the kinetic stabilization which is not included in traditional Boltzmann distributions of the conformers; (ii) the molecular motions are essential to generate accurate NMR signals, particularly for this very flexible small molecule.



	Θ_1/Θ_2	Θ_3/Θ_4
A	180/60	180/–60
B	60/–60	180/–60
C	180/60	60/180
D	60/–60	60/180
E	–60/180	60/180
F	60/–60	–60/60

Figure 11: Structures of the six most stable glycerol backbone conformers, defined by the four torsion angles Q_1 (O-C1-C2-C3), Q_2 (O-C1-C2-O), Q_3 (C1-C2-C3-O) and Q_4 (O-C2-C3-O). The hydrogen atoms have been omitted for clarity. Torsion angles are in degrees.

4.1. General Computational Method

All calculations were performed within the framework of auxiliary density functional theory (ADFT) [27, 135] including the calculations of the NMR

shielding tensors [136], as implemented in the deMon2k program [29]. Optimizations and BOMD simulations were performed with the revised PBE exchange [137] and LYP correlation [138] functionals. An empirical correction for dispersion [139] was included. Double zeta and triple zeta plus valence polarization (DZVP and TZVP) basis sets optimized for DFT calculations [76] were employed. Automatically generated auxiliary function sets containing up to *d* functions were used to fit the density [78]. The exchange correlation potential was numerically integrated on an adaptive grid [28] setting the accuracy to 10^{-6} a.u.

4.2. Born-Oppenheimer Molecular Dynamics

In BOMD simulations the electronic structure is converged at every time step, which is necessary when snapshots are used to calculate NMR properties. To facilitate conformer sampling the BOMD simulations were performed at 400 K. This temperature is within the liquid phase of glycerol which melts at 291 K and boils at 563 K. Its magnetic properties in the gas phase have not been studied due to the high boiling temperature. The NMR spectrum of a dilute solution of glycerol in D_2O has been measured at 355 K, a temperature slightly lower than that used here (400 K). In our NVT simulations the temperature was controlled using the Nosé-Hoover thermostat [140, 141] with 3 chains and coupling frequency set to 100 cm^{-1}. The velocity Verlet algorithm was used with a time step of 1 fs to propagate the nuclear coordinates. Temperature equilibration was reached within 3 ps for all trajectories. In order to test the basis set effects on the BOMD and NMR results, two sets of trajectories have been recorded, using double zeta and triple zeta plus valence polarization basis sets, namely DZVP and TZVP, respectively. Total trajectories of 2.25 ns (DZVP) and 3.4 ns (TZVP) were obtained from multiple trajectories with different starting structures, *i.e.* the optimized A ($\alpha\alpha$), B ($\alpha\gamma$), C ($\alpha\beta$), D ($\beta\gamma$), E ($\beta\beta$) and F ($\gamma\gamma$) backbone conformers that are illustrated in Fig. **11**, omitting the hydrogen atoms for clarity. To be consistent with previous works we use a double nomenclature according to Refs. [95 and 101].

4.3. NMR Chemical Shift Calculations

The ^{13}C NMR shielding calculations of glycerol were performed in the scheme of the gauge including atomic orbitals (GIAO) [142]. Based on common experience

for NMR calculations, the PW91 functional [143] was used in combination with the extended basis set aug-cc-pVDZ [144] in order to ensure a sufficient number of *s* functions to correctly describe the electronic environment around the nuclei. It also includes polarization functions, as recommended for NMR calculations [145]. However, test calculations (see Table **2**) with a larger orbital basis, aug-cc-pVTZ, lead to a decrease of the ^{13}C nuclear shieldings by about 10-14 ppm, whereas ^{13}C shielding of the tetra-methylsilane (TMS) reference decreases by only 8.4 ppm. This effect yields a general increase in the ^{13}C chemical shifts by about 2- 6 ppm with respect to the aug-cc-pVDZ calculated values. This finding is in agreement with benchmark calculations comparing mean absolute errors of ^{13}C shieldings of organic molecules calculated with these two augmented DZ and TZ bases and different functionals [136]. Therefore, we chose to use the aug-cc-pVDZ basis for our NMR shielding calculations. It is important to note that choosing another functional or a smaller basis will only displace the absolute chemical shifts but will keep the relative shift values between the carbon atoms (σ_2-σ_1 and σ_2-σ_3) almost constant with deviations less than 1 ppm. The augmented Dunning basis set was used in combination with the (GEN-A2*) auxiliary function set that contains up to *g* functions.

Table 2: NMR isotropic ^{13}C shielding of the optimized conformers using aug-cc-pVDZ and aug-cc-pVTZ (PW91 method); the corresponding TMS σ_{ref} values are 187.6 and 179.4 ppm, respectively

Basis set	Property	A (αα)	B (αγ)	C (αβ)	D (βγ)	E (ββ)	F (γγ)
aug-cc-pVDZ	σ_1	120.1	121.4	117.4	123.6	112.1	112.2
	σ_2	109.7	110.5	106.2	110.5	114.3	114.6
	σ_3	121.2	121.2	119.6	122.4	112.8	116.9
	$\sigma_1-\sigma_2$	10.5	10.9	11.2	13.1	-2.2	-2.4
	$\sigma_3-\sigma_2$	10.7	10.7	13.4	11.9	-1.5	-2.3
aug-cc-pVTZ	σ_1	107.4	108.9	105.0	111.4	101.9	102.1
	σ_2	98.3	99.0	94.6	99.0	100.7	100.6
	σ_3	109.0	109.3	107.4	110.7	100.4	104.5
	$\sigma_1-\sigma_2$	9.1	9.9	10.4	12.4	1.2	1.5
	$\sigma_3-\sigma_2$	10.7	10.3	12.8	11.7	-0.3	3.9

The ^{13}C isotropic shieldings σ were calculated at every 210 time-steps and the values obtained for each glycerol carbon were averaged over the total time (2.25 ns and 3.4 ns). The corresponding averaged chemical shifts, δ_1 and δ_3 for the primary carbons and δ_2 for the secondary carbon, were obtained using the expression $\delta_i = \sigma_{ref} - \sigma_i$, where σ_{ref} is the shielding for the TMS reference compound. The σ_{ref} value (186.9 ppm) was also calculated from averaging snapshot values over a 400 K trajectory with a total time of 500 ps.

Dividing by two the DZVP simulation time (sampling over 1.12 ns) and keeping the same time interval between the selected snapshots leads to changes of about 0.5 ppm for the averaged shieldings. The averaged shielding accuracy is also dependent on the basis set used for the dynamics. Indeed, averaging the σ values of the TZVP trajectory over 2.25 ns leads to a $\Delta\sigma$ of 0.5 ppm between the two primary ^{13}C shieldings that are experimentally indistinguishable, instead of 0.1 ppm obtained for the DZVP trajectory. In fact, extension of the TZVP trajectory to 3.4 ns reduces $\Delta\sigma$ to 0.1 ppm.

4.4. Transition State Calculations

Because conformer interchanges have been observed along the BOMD trajectories the transition states for these rearrangements were located. To this end a saddle interpolation [146] between relevant pairs of the six minima backbone configurations were performed. Before starting the saddle interpolation the glycerol backbones (C and O atoms) were aligned to maximum coincidence [147]. Once the saddle interpolation converged, the resulting coordinates were used as the starting point for a local transition state optimization by the uphill trust region method [148, 149]. This optimization was initialized with a calculated starting Hessian to guarantee a correct eigenvalue spectrum. The transition states found were verified by frequency analyses. To ensure that these transition states really connect the desired backbone conformers the intrinsic reaction coordinate (IRC) were calculated [150]. The transition state structures of the six most relevant backbone rearrangements are illustrated in Fig. **12**, omitting hydrogen atoms for clarity. The corresponding energies are reported in Table **3**.

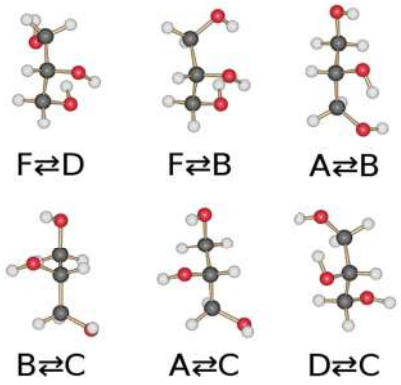


Figure 12: Calculated transition state (TS) structures of glycerol.

Table 3: Low-lying transition states activation energy (ΔE^*) calculated with DZVP and TZVP basis sets. All values are given in kcal/mol

Interchange reaction	ΔE^* (DZVP)	ΔE^* (TZVP)
F (γγ) → D (βγ)	1.2	1.8
D (βγ) → C (αβ)	6.0	5.6
B (αγ) → C (αβ)	3.6	2.4
A (αα) → C (αβ)	3.6	2.4
A (αα) → B (αγ)	3.7	3.9
F (γγ) → B (αγ)	3.7	3.7

4.5. Trajectory Mapping to Optimized Structures

Dynamics at a finite temperature allows one to follow the structural evolution of a system with time. In the present study, the glycerol dynamics shows structural changes related with interchanges between the different conformers. Analysis of these dynamical structural changes has been performed using the mapping of trajectory snapshots to optimized structures, employing the similarity index *h* [151],

$$h = 1 - \frac{d}{c}, \tag{14}$$

where *d* is proportional to the root mean square distance of the two sets of coordinates and is defined as:

$$d^2 = \sum_{i=1}^{N} \left| R'_i - R_i \right|^2 \tag{15}$$

and c is an upper bound of d which serves as a normalization constant and is defined as:

$$c^2 = 2\sum_{i=1}^{N} \left| R'_i \right|^2 + 2\sum_{i=1}^{N} \left| R_i \right|^2 \tag{16}$$

This definition ensures that h always lies between 0 and 1, with 1 denoting perfect matching of the two considered structures. For the similarity index calculations as well as for the automatic alignment procedure [147] only the coordinates of the six backbone atoms of the glycerol were considered. These coordinates are stored for the structural mapping in $\{R_i\}$ and $\{R'_i\}$. This procedure allows an automatic quantitative matching, more convenient than using dihedral angle values [152]. The similarity index values have been used to classify the six backbone conformers and assign all snapshots into six separate conformer sets (for DZVP and TZVP trajectories). In a further analysis of each of these conformer sets we identified transition state structures with a larger similarity index than the corresponding conformers. In this way, subsets of conformers are created that consist of F⇄D, F⇄B, A⇄B, B⇄C, A⇄C and D⇄C transition state structures (see Fig. **12**). Here the nomenclature X⇄Y indicates that the similarity index for the X⇄Y transition state structure is larger than for the X or Y conformer structure (PES minimum) for the given snapshot. Of course, the X⇄Y transition state label is valid for both, X to Y and Y to X, conformer interchanges.

4.6. Experimental NMR Spectra

In order to estimate the impact of the various assumptions made in this study, we have recorded the ^{13}C NMR spectrum of glycerol at different concentrations and temperatures using a Bruker DPX 400 spectrometer. Anhydrous glycerol (1,2,3 propanetriol) >99.5% from Sigma was used in D_2O solutions. Two very distinct molar concentrations were tested, namely 0.12 and 0.01, in order to verify if glycerol oligomers could be present, following indications of Raman spectra measured in these concentration conditions [91]. Since BOMD trajectories were

obtained at a temperature about 100 °C higher than standard room temperature, we have used the Bruker variable temperature setting. In order to be in keeping with our setting (spinner and magnetic probe) and below the D_2O boiling point, we have maintained a stable temperature of 80 °C.

4.7. Optimized and Dynamical Structures of Glycerol

The structural study of glycerol conformers has been approached in the literature through force field MD and *ab-initio* calculations of the potential energy surface (PES). So far only PES minima were considered using a variety of computational methods. Most *ab-initio* studies focus on the different backbone conformers (see Fig. **11**) [87, 98, 153]. As shown in this figure, four torsion angle values are necessary to define the six backbone conformers. Taking into account the possible O-H orientations will imply many more minima [154]. If one considers only the OC-CO-CO backbone (Fig. **11**), there are possible mirror images for the six conformers, due to the C-C-C mirror plane and the perpendicular C_2 axis. These mirror images are usually not distinguishable from each other and, therefore, conformer types will be discussed rather than conformers. The relative energies of these six conformer types with the methodology used here were already partially reported in [101] and are given in Table **4**.

Table 4: Relative energies ΔE and free energies ΔG (kcal/mol) of the six most stable glycerol conformers obtained with DZVP and TZVP basis sets, respectively. Corrections for zero point energy and thermal population are included in Δ**E**

Isomer	ΔE (DZVP)	ΔG (DZVP)	ΔE (TZVP)	ΔG (TZVP)
Isomer A (αα)	2.4	0.9	0.9	0.0
Isomer B (αγ)	1.9	1.0	1.1	0.7
Isomer C (αβ)	4.3	2.7	3.3	2.2
Isomer D (βγ)	3.5	1.9	2.2	1.4
Isomer E (ββ)	4.5	3.2	4.1	3.3
Isomer F (γγ)	0.0	0.0	0.0	0.5

The *ab-initio* electronic structure calculations lead to the general consensus that the backbone conformers of A (αα), B (αγ) and F (γγ) types are most stable. However, their relative energy differences are quite sensitive to the methodology

and, in particular, to the basis set. Based on *ab-initio* electronic structure calculations, two studies have reported Boltzmann population analyses of gaseous glycerol [153, 154]. It was shown that the use of an aqueous solvation model has no significant effect on these populations [154]. In the first study, based on DFT-B3LYP calculations of 13 glycerol conformers, Chelli *et al.* found a predominance of A ($\alpha\alpha$) (~ 40%) and B ($\alpha\gamma$) (~ 45%) conformers, accompanied by a small fraction of F ($\gamma\gamma$) (~ 11%), in the temperature range between 298 and 498 K [153]. A more extended Boltzmann population analysis that included 126 conformers of glycerol [154] led to a more spread out distribution among the six backbone conformer types. Only the E ($\beta\beta$) conformer type was negligible at around 2%. The A ($\alpha\alpha$), C ($\alpha\beta$), D ($\beta\gamma$) and F ($\gamma\gamma$) conformer types appeared with contributions between 10% and 20% whereas the dominant B ($\alpha\gamma$) conformer type reaches a contribution of 30% and more. These analyses were based on DFT and post Hartree-Fock electronic structure calculations and kept their qualitative form, *i.e.* negligible E ($\beta\beta$) contribution, dominant B ($\alpha\gamma$) contribution and a homogeneous distribution of the other four conformer types between 10% and 20%, independent of the electronic structure method used.

In addition, classical MD simulations have been performed at 300 K or 400 K for gaseous, pure liquid or hydrated glycerol, leading to diverse conclusions about the most populated conformer types [155-157]. The main difficulty of classical MD in this particular case is associated with the validity of the backbone torsion parameters which govern the conformer interchanges. This problem has also been reported for peptide backbone conformers [123, 158].

In order to combine the needed reliability of *ab-initio* glycerol calculations with the advantages of MD sampling we performed extensive BOMD calculations. To relate the BOMD trajectory snapshots to the six conformer types and also to the corresponding six most relevant backbone transition states (see Figs. **11** and **12**), the above described mapping algorithm was used. The barrier heights of the corresponding conformer type interchange reactions are given in Table **3**. The results of these analyses are presented in Table **5** for the DZVP and TZVP BOMD simulations. The conformer populations obtained from mapping all snapshots to every conformer type are gathered in Table **5**. This total conformer population is then analyzed in terms of subsets of snapshots mapping the conformer minimum

or/and possible transition state structures. Table **5** provides some very interesting insight into the dynamics of the glycerol molecule: (i) conformer types (Table **4**) B ($\alpha\gamma$) and F ($\gamma\gamma$) are the most populated along our BOMD trajectories as expected considering its relative energies; (ii) structures corresponding to the transition state type for the interchange between B ($\alpha\gamma$) and F ($\gamma\gamma$) show substantial population with both basis sets. This F-B transition state population decreases substantially the F ($\gamma\gamma$) minimum participation; (iii) The population of the A ($\alpha\alpha$) conformer type is surprisingly small in both BOMD trajectories considering that this conformer type is energetically similar to the B ($\alpha\gamma$) type; and (iv) increasing the basis set used for BOMD enlarges the spread of the conformer type populations but keeps the B ($\alpha\gamma$) and F ($\gamma\gamma$) types preponderant.

Table 5: Distribution of the BOMD trajectory snapshots (DZVP and TZVP) among the conformer sets and subsets (minimum and TS types). The reported populations are given in %. See text for detail and TS nomenclature

Conformer sets		**DZVP Polulations**		**TZVP Populations**	
	Subset	*Total*	*Subset*	*Total*	*Subset*
A	Minimum	7.74	6.93	9.10	7.74
	A⇄B		0.56		0.30
	A⇄C		0.25		1.06
B	Minimum	34.72	34.34	31.82	31.37
	A⇄B		0.38		0.43
	F⇄B		-		0.02
C	Minimum	2.68	2.11	12.38	7.36
	A⇄C		0.17		0.27
	B⇄C		0.12		0.85
	D⇄C		0.40		3.90
D	Minimum	5.84	4.71	17.61	16.05
	F⇄D		1.13		1.56
E	Minimum	0.89	0.89	1.41	1.41
F	Minimum	48.13	38.19	27.67	19.38
	F⇄B		9.94		8.09
	F⇄D		-		0.20

Obviously, these BOMD glycerol conformer type distributions are different from those given by Boltzmann population analyses. Neither the BOMD nor the Boltzmann distributions are exceptionally sensitive to the underlying electronic structure methods. It seems, thus, likely to attribute this difference to the finite

sampling time in our BOMD simulations. Through dynamics, kinetic stabilization that is strongly related to the pathway between conformers is included, even on the short ns time scale, whereas it is not present in the Boltzmann distribution. Thus, the question arises if these kinetic effects are only notable on very short time scales or if they also exist on longer macroscopic ones. To this end, it is informative to compare the BOMD and Boltzmann distributions to the corresponding experiments. We have found three experimental gas phase studies on glycerol that propose conformer type distributions in the literature [87, 92, 159]. The population analyses proposed by these experimental studies are based on fitting the contribution of every conformer, through the weighted superposition of calculated conformer spectra, to the experimental spectrum.

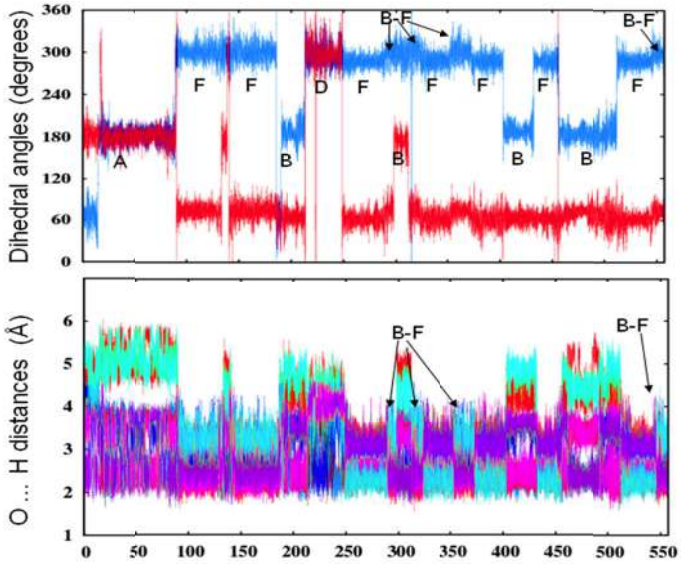


Figure 13: Analysis of a 550 ps trajectory cutout from the DZVP BOMD. Top: Variation of θ_1 (blue) and θ_3 (red) torsion angles. The assignment of trajectory regions to glycerol conformer types and B to F or F to B transition states (B-F) are also given. Bottom: Fluctuations of the 6 non-bonding interatomic O···H distances.

All three studies propose two dominant glycerol conformer types. The gas phase electron diffraction patterns were interpreted by Bastiansen assuming a mixture of A (αα) and B (αγ) conformer types [92]. The same mixture, 83% A (αα) and 15% B (αγ), results from the best fit of the gas phase IR spectrum of glycerol to corresponding DFT spectra of the individual conformer types [87]. On the other

hand, the gas phase microwave study of Maccaferri and co-workers [159] yields a mixture of B (αγ) and F (γγ) glycerol conformer types. This shows that the experimental results are sensitive to their interpretation. In fact, the structure assignment from the microwave study is based on *ab-initio* geometries and measured rotational transitions [159]. In microwave experiments, rotational constants are properties that quantify the rotational energies directly related to conformer interchanges, whereas electron diffraction and IR spectra concern predominantly higher energy spectrum domains. This may explain that the agreement of the microwave results with BOMD distributions is remarkable. Moreover, the energy splitting between the two conformer types of 0.62 kcal/mol from this microwave experiment (taking into account the most stable F (γγ) conformer) is in good agreement with the calculated Gibbs energy splitting of 1.00 and 0.20 kcal/mol with the DZVP and TZVP basis, respectively (Table **4**).

The agreement between the gas phase microwave and BOMD conformer distribution, in particular the very low contribution of the A conformer in contrast to electron diffraction and IR populations, suggests that the observed kinetic stabilization at the ns time scale in the BOMD simulations is also important on much longer time scales. A necessary prerequisite for this assumption is that all relevant structure rearrangements occur along our BOMD trajectories. To analyze this in more detail, Fig. **13** depicts the structural fluctuations along a 550 ps trajectory cutout of the DZVP BOMD. The trajectory cutout starts with a B (αγ) conformer type. The upper part of Fig. **13** shows the variation of the two O-C-C-C backbone torsion angles (θ_1 and θ_3 in Fig. **11**), whereas the lower part depicts the corresponding fluctuations in the six non-bonded O···H distances between the three OH groups. Fig. **13** shows that the non-bonded O···H distances are rapidly exchanging due to the rotation of the OH groups around the C-O bonds. As expected, these rotations are much faster than the backbone conformational changes as the comparison of the top and bottom parts of Fig. **13** immediately reveals. In the depicted trajectory cutout of 550 ps one finds around 15 backbone conformational changes which indicate a glycerol conformer type interchange roughly every 40 ps. The various conformer types along with the B to F transition state, as obtained from the trajectory mapping, are also marked in the top part of Fig. **13**. From this assignment and the fluctuations in the non-bonding

OH distances (see bottom of Fig. **13**) one finds that OH rotations are hindered, *i.e.* O···H exchanges occur less frequently, in the F (γγ) conformer type structure. On the other hand, conformer backbone rearrangements are always accompanied by large O···H exchanges. This suggests that the C-O-H degrees of freedom are coupled to the backbone torsion angles in order to lower energy barriers.

4.8. NMR Calculations

To further support the hypothesis that the BOMD samplings yield glycerol conformer type distributions similar to the ones at macroscopic time scales the ^{13}C chemical shielding was calculated from the BOMD. Because relaxation times in NMR spectroscopy are of the order of ms to s we only can expect a fair match with experiment if the sampling covers a similar configurational space. Averaging the chemical shieldings yields the following chemical shifts δ: for C_1, C_3 (primary carbons) 71.0, 71.1 ppm (DZVP trajectory) and 69.0, 69.0 (TZVP trajectory); for C_2 (secondary carbon): 79.3 ppm and 79.1 for the DZVP and TZVP trajectories, respectively. In Table **6** these values and the relative Δδ (δ_1-δ_2, δ_3-δ_2) chemical shifts are compared with experimental data reported in the literature [96, 97, 160-162].

Table 6: Measured and calculated glycerol chemical shifts in ppm

δ_1	δ_2	$\delta_1-\delta_2 = \delta_3-\delta_2$	Conditions	Refs.
Experimental values				
63.0	72.4	9.4	DMSO-d6, room temperature	[149]
64.0	73.5	9.5	near room temperature	[154]
66.9	76.4	9.5	D_2O, 50 °C	[150]
66.0	75.0	9.0	Solid state	[155]
62.0	71.7	9.7	Olive oil, 35 °C	[151]
62.34	71.90	9.56	conc. 0.12 molar in D_2O, 25 °C	[124]
62.34	71.93	9.59	conc. 0.01 molar in D_2O, 25 °C	[124]
63.54	72.83	9.29	conc. 0.12 molar in D_2O, 80 °C	[124]
63.57	72.92	9.35	conc. 0.01 molar in D_2O, 80 °C	[124]
Theoretical values				
71.0	79.3	8.3	DZVP Dynamics, 127 °C, NMR aug-cc-pvDZ/PW91	[124]
69.0	79.1	10.1	TZVP Dynamics, 127 °C, NMR aug-cc-pvDZ/PW91	[124]

It is found that the calculated relative chemical shifts are in good agreement with experimental values, whereas the absolute chemical shifts improve going from DZVP to TZVP BOMD simulations. It is well known that the methodology used for the calculation of nuclear magnetic shieldings is sensitive to the basis set extension, but this result shows that the method used for the description of the structures as a function of time is also important. The calculated absolute δ values are shifted by about 5 ppm with respect to experiment, which corresponds to the expected error range of the methodology used here for the molecular ^{13}C δ with an external reference [136].

The spread of the experimental chemical shifts is about 4.8 ppm, whereas that of the relative shift, Δδ, is only 0.3 ppm. The effect of a temperature increase from room temperature to 400 K on Δδ values can be estimated at about 0.5 ppm. The average of the Δδ values of the two dynamics, 9.2 ppm, is in close agreement with experiment. It is important to note that chemical shift values are not affected by the molar concentration of glycerol in D_2O. In fact, Raman studies of glycerol in diluted D_2O solutions show that D_2O breaks dimers and higher oligomers of glycerol and thus involves mainly glycerol monomers, in particular for the lowest dilution of 0.01 [91].

In Table **7**, the NMR shieldings calculated for the optimized structures of the glycerol conformers are compared with the corresponding BOMD average values associated with the reference structures (DZVP trajectory). This BOMD averaging has been performed using all the trajectory snapshots recognized as relevant to the different backbone conformer types and transition states by the mapping algorithm (Table **5**). The first interesting effect of the dynamics is to provide very similar σ values to the two primary carbons, σ_1 and σ_3, at least for the stable conformers, since their difference has become less than the width of the observed signal (0.5 ppm). This benefit from the dynamics also shows that our trajectory time-lengths are long enough to converge this NMR property. The coincidence of the two primary carbon signals, averaged over all dynamical structures, validates *per se* the calculated population statistics. The second interesting effect of dynamics observed in Table **7** is the general expected decrease of carbon shielding. This effect, particularly pronounced for C2, is due to the sampling of higher energy vibrational structures with fluctuations of bond

lengths as well as bond and torsion angles which decrease the shieldings of the carbon nuclei with respect to structures at T=0 K. However, even if the dynamical representation of the 6 PES minima improves the calculated ^{13}C shieldings with respect to the static values, it is necessary to include a sampling of all structures present in the dynamics, in particular, those gathered in the TS subsets (Table **5**) to get a realistic description of the spectrum. This remark is in agreement with conclusions from a previous dynamical study of ^{13}C chemical shifts in solid peptides that also reports very large ^{13}C chemical shift fluctuations and shows how much accuracy of the ^{13}C shifts can be improved when averaging these fluctuations over a large time scale (5μs) [127].

Table 7: Comparison of calculated NMR isotropic shielding (PW91/aug-cc-pvDZ) of optimized conformers (revPBE/DZVP) with averaged BOMD results (revPBE/DZVP). All values are in ppm

	Property	A (αα)	B (αγ)	C (αβ)	D (βγ)	E (ββ)	F (γγ)	F⇄B TS
Optimized Structures	σ_1	120.1	121.4	117.4	123.6	112.1	112.2	112.1
	σ_2	109.7	110.5	106.2	110.5	114.3	114.6	120.7
	σ_3	121.2	121.2	119.6	122.4	112.8	116.9	115.0
	$\sigma_1-\sigma_2$	10.5	10.9	11.2	13.1	-2.2	-2.4	8.6
	$\sigma_3-\sigma_2$	10.7	10.7	13.4	11.9	-1.5	-2.3	5.7
BOMD Averaged	$\langle\sigma_1\rangle$	118.6	118.8	116.1	120.8	116.9	113.3	114.0
	$\langle\sigma_2\rangle$	105.1	107.0	105.2	108.2	114.3	108.9	108.8
	$\langle\sigma_3\rangle$	118.4	118.6	116.0	120.8	116.5	113.0	113.2
	$\langle\sigma_2\rangle-\langle\sigma_1\rangle$	13.5	11.8	10.9	12.6	2.6	4.4	5.2
	$\langle\sigma_2\rangle-\langle\sigma_3\rangle$	13.3	11.6	10.8	12.6	2.2	4.1	4.4

CONCLUSION

A combined BOMD/NMR study of glycerol allows one to gain more insight into the conformer distribution of this molecule in the gas phase. The analysis of the BOMD trajectories shows that B (αγ) and F (γγ) conformer types are most populated. This is in qualitative accordance with experimental gas phase microwave studies [159]. This agreement can be interpreted such that the observed dynamical effects at the ns time scale in the BOMD simulations are also important on much longer time scales. To further support this hypothesis the NMR spectra of glycerol have been studied experimentally and theoretically. The

experimental NMR studies show that the molar concentrations of glycerol in D_2O have no effect on the chemical shift values. This is in perfect agreement with corresponding Raman studies [91] which show that D_2O breaks dimers and higher oligomers of glycerol and thus involves mainly glycerol monomers. Thus, one can safely assume that experimental NMR results indeed refer to the monomer. As a consequence, the good (absolute δ values) and very good ($\Delta\delta$ values) agreement between the experimental and theoretical NMR results can be seen as further evidence for our hypothesis that the ns BOMD distribution is indeed similar to the ones on much longer time scales (ms to s). Quite surprisingly, a considerable amount of F ($\gamma\gamma$) type conformer in the BOMD trajectory (see Table **5**) is compatible with very good $\Delta\delta$ values (see Table **6**) which is fundamentally different to $\Delta\delta$ values from optimized structures (see Table **7**). This, and the large amount of other minima structures that entered the BOMD NMR calculations indicate that accurate prediction of NMR spectra of flexible molecules like glycerol requires the incorporation of all dynamical configurations.

In conclusion, one finds from ns BOMD simulations a glycerol conformer distribution that is governed by kinetics and, therefore, different from the corresponding Boltzmann distributions. Comparison with experimental gas phase microwave results and D_2O chemical shifts suggests that this distribution is stable on much longer time scales (ms to s) than the one (ns) used in the BOMD simulations.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge the computational resources provided by Compute Canada WestGrid and CINVESTAV-IPN for the production of calculations and benchmarking. AAI and AMK acknowledge financial support from CONACyT project 179409 (Mexico) and access to the hybrid computacional cluster Xiuhcóatl in CINVESTAV (Mexico). PC and JMVP acknowledge financial support from CONACyT project 130726 (Mexico). AG, TM and BZG acknowledge access to the HPC resources of CCRT made by GENCI (France). DRS is grateful to NSERC (Canada) for continued support through Discovery Grants. CZGC, RGA and AV acknowledge the partial financial support from CONACyT project 128369 (Mexico) and for

computational resources provided by DGSCA-UNAM and LSVP-UAMI for providing computing time in their servers KanBalam and Aitzaloa, respectively (Mexico).

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

REFERENCES

[1] Szabo, A.; Ostlund, N.S. *Modern Quantum Chemistry*; Dover Publishing: Mineola, **1996**, pp. 108-230.
[2] Thomas, L.H. The calculation of atomic fields. *Proc. Cambridge Phil. Soc*., **1927**, *23*, 542-548.
[3] Fermi, E. Un metodo statistico per la determinazione di alcune prioprietà dell'atomo. *Rend. Accad. Naz. Lincei*, **1927**, *6*, 602-607.
[4] Teller, E. On the Stability of Molecules in the Thomas-Fermi Theory. *Rev. Mod. Phys.,* **1962**, *34*, 627-631.
[5] Balàzs, N.L. Formation of Stable Molecules within the Statistical Theory of Atoms. *Phys. Rev.,* **1967**, *156*, 42-47.
[6] Lieb, E.H.; Simon, B. The Thomas-Fermi Theory of Atoms, Molecules and Solids. *Adv. Math.,* **1977**, *23*, 22-116.
[7] Dirac, P.A.M. Note on exchange phenomena in the Thomas atom. *Math. Proc. Camb. Phil. Soc*., **1930**, *26*, 376-385.
[8] Wigner, E; Seitz, F. On the constitution of metallic sodium. *Phys. Rev*., **1933**, *43*, 804-810.
[9] Wigner, E; Seitz, F. On the constitution of metallic sodium. II *Phys. Rev.*, **1934**, *46*, 509-524.
[10] Slater, J.C. A simplification of the Hartree-Fock method. *Phys. Rev*., **1951**, *81*, 385-390.
[11] Gaspar, R. Über eine Approximation des Hartree-Fockschen Potentials durch eine Universelle Potentialfunktion. *Acta Phys. Acad. Scientiarum Hungaricae*, **1954**, *3*, 263-283.
[12] Slater, J.C.; Johnson, K. H. Self-consistent-field Xα cluster method for polyatomic molecules and solids. *Phys. Rev. B*, **1972**, *5*, 844-853.
[13] Slater, J.C. Wave functions in a periodic potential. *Phys. Rev*., **1937**, *51*, 846-851.
[14] Martin, R.M. *Electronic Structure: Basic Theory and Practical Methods*; Cambridge University Press: Cambridge, **2004**, pp. 323-329.
[15] Salahub, D.R.; Goursot, A.; Weber, J.; Köster, A.M.; Vela, A. Applied density functional theory and the deMon codes 1964-2004. In *Theory and applications of computational chemistry: the first 40 years. A volume of technical and historical perspectives*. Eds: Dykstra, C.; Frenking, G.; Kim, K.; Scuseria, G. Elsevier, Amsterdam, **2005**.
[16] Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev*., **1964**, *136*, B864-B871.
[17] Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev*., **1965**, *140*, A1133-A1138.

[18] Roothaan, C.C.J. New developments in molecular orbital theory. *Rev. Mod. Phys*., **1951**, *23*, 69-89.
[19] Hall, G.G. The molecular orbital theory of chemical valency. VIII. a method of calculating ionization potentials. *Proc. Roy. Soc. Lond. A*, **1951**, *205*, 541-552.
[20] Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley: Chichester, **2000**, pp. 365-371.
[21] Baerends, E.J.; Ellis, D.E.; Ros, P. Self-consistent molecular Hartree-Fock-Slater calculations I. The computational procedure. *Chem. Phys*., **1973**, *2*, 41-51.
[22] Sambe, H.; Felton, R.H. A new computational approach to Slater's SCF-Xα equation. *J. Chem. Phys*., **1975**, *62*, 1122-1126.
[23] Dunlap, B.I.; Connolly, J.W.; Sabin, J.R. On first-row diatomic molecules and local density models. *J. Chem. Phys*., **1979**, *71*, 4993-4999.
[24] Mintmire, J.W.; Dunlap, B.I. Fitting the Coulomb potential variationally in linear-combination-of-atomic-orbitals density-functional calculations. *Phys. Rev. A*, **1982**, *25*, 88-95.
[25] Trickey, S. B.; Alford, J. A.; Boettger, J. C. Methods and Implementation of Robust, High-Precision Gaussian Basis DFT Calculations for Periodic Systems: the GTOFF Code. In *Computational Materials Science: Theoretical and Computational Chemistry, Vol. 15*; Ed: Leszcynski, J.; Elsevier, Amsterdam, **2004**, pp 171-228.
[26] Köster, A.M. Hermite Gaussian auxiliary functions for the variational fitting of the Coulomb potential in density functional methods. *J. Chem. Phys*., **2003**, *118*, 9943-9951.
[27] Köster, A.M.; Reveles, J.U.; del Campo, J.M. Calculation of exchange-correlation potentials with auxiliary function densities. *J. Chem. Phys*., **2004**, *121*, 3417-3424.
[28] Köster, A.M.; Flores-Moreno, R.; Reveles, J.U. Efficient and reliable numerical integration of exchange-correlation energies and potentials. *J. Chem. Phys*., **2004**, *121*, 681-690.
[29] Köster, A.M.; Geudtner, G.; Calaminici, P.; Casida, M.E.; Dominguez, V. D.; Flores-Moreno, R.; Gamboa, G.U.; Goursot, A.; Heine, T.; Ipatov, A.; Janetzko, F.; del Campo, J.M.; Reveles, J.U.; Vela, A.; Zuniga-Gutierrez B.; Salahub, D.R. deMon2k, Version 3, The deMon Developers, CINVESTAV, Mexico City, **2011**.
[30] Boys, S.F. Electronic wave functions. I. A general method of calculation for the stationary states of any molecular system. *Proc. R. Soc. Lond. A*, **1950**, *200*, 542-554.
[31] Köster, A.M. Efficient recursive computation of molecular integrals for density functional methods. *J. Chem. Phys*., **1996**, *104*, 4114-4124.
[32] Almlöf, J.; Faegri, K.; Korsell, K. Principles for a direct SCF approach to LICAO-MO ab-initio calculations. *J. Comp. Chem*., **1982**, *3*, 385-399.
[33] Greengard, L.; Rokhlin, V. A fast algorithm for particle simulations. *J. Comp. Phys*., **1987**, *73*, 325-348.
[34] White, C.A.; Head-Gordon, M. Derivation and efficient implementation of the fast multipole method. *J. Chem. Phys*., **1994**, *101*, 6593-6605.
[35] Petersen, H.G.; Soelvason, D.; Perram, J.W.; Smith, E.R. The very fast multipole method. *J. Chem. Phys*., **1994**, *101*, 8870-8876.
[36] Kudin, K.N.; Scuseria, G.E. A fast multipole method for periodic systems with arbitrary unit cell geometries. *Chem. Phys. Lett*., **1998**, *283*, 61-68.
[37] White, C.A.; Johnson, B.G.; Gill, P.M.W.; Head-Gordon, M. The continuous fast multipole method. *Chem. Phys. Lett*., **1995**, *230*, 8-16.

[38] Strain, M.C.; Scuseria, G.E.; Frisch, M.J. Achieving linear scaling for the electronic quantum Coulomb problem. *Science*, **1996**, *271*, 51-53.
[39] Kudin, K.N.; Scuseria, G.E. Linear-scaling density-functional theory with Gaussian orbitals and periodic boundary conditions: efficient evaluation of energy and forces *via* the fast multipole method. *Phys. Rev. B*, **2000**, *61*, 16440-16453.
[40] Alvarez-Ibarra, A.; Köster, A.M. Double asymptotic expansion of three-center electronic repulsion integrals. *J. Chem. Phys*., **2013**, *139*, 024102-1 - 024102-10.
[41] Alvarez-Ibarra, A. Asymptotic expansion of molecular integrals in self-consistent auxiliary density functional methods. PhD Thesis, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional: Mexico City, August **2013**.
[42] Obara, S.; Saika, A. Efficient recursive computation of molecular integrals over cartesian Gaussian functions. *J. Chem. Phys*., **1986**, *84*, 3963-3974.
[43] E. Clementi, Analytical self-consistent field functions for positive ions. I. Isoelectronic series with 2 to 10 electrons. *J. Chem. Phys*., **1963**, *38*, 996-1000.
[44] Quintanar, L.; Rivillas-Acevedo, L.; Grande-Aztatzi, R.; Gómez-Castro, C. Z.; Arcos-López, T.; Vela, A. Copper coordination to the prion protein: Insights from theoretical studies. *Coord. Chem. Rev.*, **2013**, *257*, 429-444.
[45] Prusiner, S. B. Prions. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 13363-13383.
[46] Prusiner, S. B. Novel proteinaceous infectious particles cause scrapie. *Science*, **1982**, *216*, 136-144.
[47] Prusiner, S. B. Molecular biology of prion diseases. *Science*, **1991**, *252*, 1515-1522.
[48] Bessen, R. A.; Marsh, R. F. Distinct PrP properties suggest the molecular-basis of strain variation in transmissible mink encephalopathy. *J. Virol.*, **1994**, *68*, 7859-7868.
[49] Prusiner, S. B., Prion diseases and the BSE crisis. *Science*, **1997**, *278*, 245-251.
[50] Brown, D. R.; Qin, K. F.; Herms, J. W.; Madlung, A.; Manson, J.; Strome, R.; Fraser, P. E.; Kruck, T.; vonBohlen, A.; Schulz-Schaeffer, W.; Giese, A.; Westaway, D.; Kretzschmar, H. The cellular prion protein binds copper *in vivo*. *Nature*, **1997**, *390*, 684-687.
[51] Davies, P.; Brown, D. R. The chemistry of copper binding to PrP: Is there sufficient evidence to elucidate a role for copper in protein function? *Biochem. J.*, **2008**, *410*, 237-244.
[52] Pauly, P. C.; Harris, D. A. Copper stimulates endocytosis of the prion protein. *J. Biol. Chem.*, **1998**, *273*, 33107-33110.
[53] Sumudhu, W.; Perera, S.; Hooper, N. M. Ablation of the metal ion-induced endocytosis of the prion Protein by disease-associated mutation of the octarepeat region. *Curr. Biol.*, **2001**, *11*, 519-523.
[54] Bounhar, Y.; Zhang, Y.; Goodyer, C. G.; LeBlanc, A. Prion protein protects human neurons against bax-mediated apoptosis. *J. Biol. Chem.*, **2001**, *276*, 39145-39149.
[55] Klamt, F.; Dal-Pizzol, F.; da Frota, M. L. C.; Walz, R.; Andrades, M. E.; da Silva, E. G.; Brentani, R. R.; Izquierdo, I.; Moreira, J. C. F., Imbalance of antioxidant defense in mice lacking cellular prion protein. *Free Radic. Biol. Med.*, **2001**, *30*, 1137-1144.
[56] Mange, A.; Milhavet, O.; Umlauf, D.; Harris, D.; Lehmann, S. PrP-dependent cell adhesion in N2a neuroblastoma Cells. *FEBS Lett.*, **2002**, *514*, 159-162.
[57] Mouillet-Richard, S.; Ermonval, M.; Chebassier, C.; Laplanche, J. L.; Lehmann, S.; Launay, J. M.; Kellermann, O. Signal transduction through prion protein. *Science*, **2000**, *289*, 1925-1928.

[58] Nadal, R. C.; Abdelraheim, S. R.; Brazier, M. W.; Rigby, S. E. J.; Brown, D. R.; Viles, J. H. Prion protein does not redox-silence Cu^{2+}, but is a sacrificial quencher of hydroxyl radicals. *Free Radic. Biol. Med.*, **2007**, *42*, 79-89.

[59] You, H.; Tsutsui, S.; Hameed, S.; Kannanayakal, T. J.; Chen, L.; Xia, P.; Engbers, J. D. T.; Lipton, S. A.; Stys, P. K.; Zamponi, G. W. Aβ neurotoxicity depends on interactions between copper ions, prion protein, and *N*-methyl-D-aspartate receptors. *Proc. Natl. Acad. Sci. USA*, **2012**, *109*, 1737-1742.

[60] Donne, D. G.; Viles, J. H.; Groth, D.; Mehlhorn, I.; James, T. L.; Cohen, F. E.; Prusiner, S. B.; Wright, P. E.; Dyson, H. J. Structure of the recombinant full-length hamster prion protein PrP(29-231): The N terminus is highly flexible. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 13452-13457.

[61] Aronoff-Spencer, E.; Burns, C. S.; Avdievich, N. I.; Gerfen, G. J.; Peisach, J.; Antholine, W. E.; Ball, H. L.; Cohen, F. E.; Prusiner, S. B.; Millhauser, G. L. Identification of the Cu^{2+} binding sites in the N-terminal domain of the prion protein by EPR and CD spectroscopy. *Biochemistry*, **2000**, *39*, 13760-13771.

[62] Burns, C. S.; Aronoff-Spencer, E.; Legname, G.; Prusiner, S. B.; Antholine, W. E.; Gerfen, G. J.; Peisach, J.; Millhauser, G. L. Copper coordination in the full-length, recombinant prion protein. *Biochemistry*, **2003**, *42*, 6794-6803.

[63] Jackson, G. S.; Murray, I.; Hosszu, L. L. P.; Gibbs, N.; Waltho, J. P.; Clarke, A. R.; Collinge, J. Location and properties of metal-binding sites on the human prion protein. *Proc. Natl. Acad. Sci. USA*, **2001**, *98*, 8531-8535.

[64] Klewpatinond, M.; Davies, P.; Bowen, S.; Brown, D. R.; Viles, J. H. Deconvoluting the Cu^{2+} binding modes of full-length prion protein. *J. Biol. Chem.*, **2008**, *283*, 1870-1881.

[65] Qin, K. F.; Yang, Y.; Mastrangelo, P.; Westaway, D. Mapping Cu(II) binding sites in prion proteins by diethyl pyrocarbonate modification and matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometric footprinting. *J. Biol. Chem.*, **2002**, *277*, 1981-1990.

[66] Berti, F.; Gaggelli, E.; Guerrini, R.; Janicka, A.; Kozlowski, H.; Legowska, A.; Miecznikowska, H.; Migliorini, C.; Pogni, R.; Remelli, M.; Rolka, K.; Valensin, D.; Valensin, G. Structural and dynamic characterization of copper(11) binding of the human prion protein outside the octarepeat region. *Chem. Eur. J.*, **2007**, *13*, 1991-2001.

[67] Chattopadhyay, M.; Walter, E. D.; Newell, D. J.; Jackson, P. J.; Aronoff-Spencer, E.; Peisach, J.; Gerfen, G. J.; Bennett, B.; Antholine, W. E.; Millhauser, G. L. The octarepeat domain of the prion protein binds Cu(II) with three distinct coordination modes at pH 7.4. *J. Am. Chem. Soc.*, **2005**, *127*, 12647-12656.

[68] Burns, C. S.; Aronoff-Spencer, E.; Dunham, C. M.; Lario, P.; Avdievich, N. I.; Antholine, W. E.; Olmstead, M. M.; Vrielink, A.; Gerfen, G. J.; Peisach, J.; Scott, W. G.; Millhauser, G. L. Molecular features of the copper binding sites in the octarepeat domain of the prion protein. *Biochemistry*, **2002**, *41*, 3991-4001.

[69] Barry, S. D.; Rickard, G. A.; Pushie, M. J.; Rauk, A. The affinity of HGGG, GHGG, GGHG, and GGGH peptides for copper(II) and the structures of their complexes - An *ab initio* study. *Can. J. Chem.*, **2009**, *87*, 942-953.

[70] Pushie, M. J.; Vogel, H. J. A Potential Mechanism for Cu^{2+} Reduction, beta-Cleavage, and beta-Sheet Initiation Within The N-Terminal Domain of the prion Protein: Insights from Density Functional Theory and Molecular Dynamics Calculations. *J. Toxicol. Environ. Health A*, **2009**, *72*, 1040-1059.

[71] Shearer, J.; Soh, P.; Lentz, S. Both Met(109) and Met(112) are utilized for Cu(II) coordination by the amyloidogenic fragment of the human prion protein at physiological pH. *J. Inorg. Biochem.*, **2008**, *102*, 2103-2113.

[72] Yamamoto, N.; Kuwata, K. Difference in redox behaviors between copper-binding octarepeat and nonoctarepeat sites in prion protein. *J. Biol. Inorg. Chem.*, **2009**, *14*, 1209-1218.

[73] Grande-Aztatzi, R.; Rivillas-Acevedo, L.; Quintanar, L.; Vela, A. Structural models for Cu(II) bound to the fragment 92−96 of the human prion protein. *J. Phys. Chem. B*, **2013**, *117*, 789-799.

[74] Rivillas-Acevedo, L.; Grande-Aztatzi, R.; Lomeli, I.; Garcia, J. E.; Barrios, E.; Teloxa, S.; Vela, A.; Quintanar, L. Spectroscopic and electronic structure studies of copper(II) binding to His111 in the human prion protein fragment 106-115: Evaluating the role of protons and methionine residues. *Inorg. Chem.*, **2011**, *50*, 1956-1972.

[75] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, **1996**, *77*, 3865-3868.

[76] Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. Optimization of Gaussian-type basis-sets for local spin-density functional calculations. Part 1. Boron through neon, optimization technique and validation. *Can. J. Chem.*, **1992**, *70*, 560-571.

[77] Dunlap, B. I.; Connolly, J. W. D.; Sabin, J. R. Some approximations in applications of X-alpha theory. *J. Chem. Phys.*, **1979**, *71*, 3396-3402.

[78] Calaminici, P.; Janetzko, F.; Köster, A.M.; Mejia-Olvera, R.; Zuniga-Gutierrez, B. Density functional theory optimized basis sets for gradient corrected functionals: 3d transition metal systems. *J. Chem. Phys.,* **2007**, *126*, 044108-044300.

[79] Neese, F. ORCA - An ab-initio, density functional and semiempirical program package, 2.6, rev. 35; University of Bonn: **2008**.

[80] Hureau, C.; Charlet, L.; Dorlet, P.; Gonnet, F.; Spadini, L.; Anxolabehere-Mallart, E.; Girerd, J. J. A spectroscopic and voltammetric study of the pH-dependent Cu(II) coordination to the peptide GGGTH: relevance to the fifth Cu(II) site in the prion protein. *J. Biol. Inorg. Chem.*, **2006**, *11*, 735-744.

[81] Hureau, C.; Mathe, C.; Faller, P.; Mattioli, T. A.; Dorlet, P. Folding of the prion peptide GGGTHSQW around the copper(II) ion: identifying the oxygen donor ligand at neutral pH and probing the proximity of the tryptophan residue to the copper ion. *J. Biol. Inorg. Chem.*, **2008**, *13*, 1055-1064.

[82] Di Natale, G.; Grasso, G.; Impellizzeri, G.; La Mendola, D.; Micera, G.; Mihala, N.; Nagy, Z.; Osz, K.; Pappalardo, G.; Rigo, V.; Rizzarelli, E.; Sanna, D.; Sovago, I. Copper(II) interaction with unstructured prion domain outside the octarepeat region: Speciation, stability, and binding details of copper(II) complexes with PrP106-126 peptides. *Inorg. Chem.*, **2005**, *44*, 7214-7225.

[83] Jobling, M. F.; Huang, X. D.; Stewart, L. R.; Barnham, K. J.; Curtain, C.; Volitakis, I.; Perugini, M.; White, A. R.; Cherny, R. A.; Masters, C. L.; Barrow, C. J.; Collins, S. J.; Bush, A. I.; Cappai, R. Copper and zinc binding modulates the aggregation and neurotoxic properties of the prion peptide PrP106-126. *Biochemistry*, **2001**, *40*, 8073-8084.

[84] Klewpatinond, M.; Viles, J. H. Fragment length influences affinity for Cu^{2+} and Ni^{2+} binding to His(96) or His(111) of the prion protein and spectroscopic evidence for a multiple histidine binding only at low pH. *Biochem. J.*, **2007**, *404*, 393-402.

[85] Remelli, M.; Donatoni, M.; Guerrini, R.; Janicka, A.; Pretegiani, P.; Kozlowski, H. Copper-ion interaction with the 106-113 domain of the prion protein: a solution-equilibria study on model peptides. *Dalton Trans.*, **2005**, 2876-2885.

[86] Goursot, A.; Mineva, T.; Vasquez-Perez, J. M.; Calaminici, P.; Köster, A.M. and Salahub, D.R. Contribution of high-energy conformations to NMR chemical shifts, a DFT-BOMD study. *Phys. Chem. Chem. Phys.*, **2013**, *15*, 860-867.

[87] Chelli, R.; Gervasio, F. L.; Gellini, C.; Procacci, P.; Cardini, G.; Schettino, V. Density functional calculation of structural and vibrational properties of glycerol. *J. Phys. Chem. A*, 2000, **104**, 5351-5357.

[88] Perova, T. S.; Christensen, D. H.; Rasmussen, U.; Vij, J. K.; Nielsen, O. F. Far-infrared spectra of highly viscous liquids: glycerol and triacetin (glycerol triacetate). *Vib. Spectrosc.,* **1998**, *18*, 149-156.

[89] Zelent, B.; Nucci, N. V.; Vanderkoi, J. M. Liquid and ice water and glycerol/water glasses compared by infrared spectroscopy from 295 to 12 K. *J. Phys. Chem. A,* **2004**, *108*, 11141-11150.

[90] Mendelovici, E.; Frost, R. L.; Kloproge, T. Cryogenic Raman spectroscopy of glycerol. *J. Raman Spectrosc*. **2000**, *31*, 1121-1126.

[91] Mudalige, A.; Pemberton, J. E. Raman spectroscopy of glycerol/D_2O solutions. *Vib. Spectrosc.,* **2007**, *45*, 27-35.

[92] Bastiansen, O. Intra-Molecular hydrogen bonds in ethylene glycol, glycerol, and ethylene chlorohydrin. *Acta Chem. Scand*., **1949**, *3*, 415-421.

[93] Garavi, M.; Dore, J. C.; Champeney, D. C. Structural studies of liquid D-glycerol II. Molecular conformation and long range correlations. *Mol. Phys.*, **1987**, *62*, 475-487.

[94] van Koningsveld, H. A conformational study on glycerol in a D_2O solution by means of 220 Mc PMR data. *Recl. Trav. Chim. Pays-Bas,* **1970**, *89*, 801-812.

[95] Soltwisch, M.; Steffen, B. Z. The x-ray structure factor of liquid glycerol. *Naturforsch., A: Phys. Sci.* **1981**, *36*, 1045-1051.

[96] Kalinowski, H. Berger, S. Braun, S. *C-13 NMR Spectroscopy*; J. Wiley: New York, NY, **1988**.

[97] Moire, L.; Schmutz, A.; Buchala, A.; Yan, B.; Stark, R.E.; Ryser U. Glycerol Is a Suberin Monomer. New experimental evidence for an old hypothesis. *Plant Physiol*., **1999**, *119*, 1137-1146.

[98] Goursot, A.; Mineva, T.; Bissig, C.; Gruenberg, J.; Salahub, D.R. Structure, dynamics, and energetics of lysobisphosphatidic acid (LBPA) isomers. *J. Phys. Chem. B*, **2010,** *114*, 15712-15720.

[99] Krishnamurty, S.; Stefanov, M.; Mineva, T.; Begu, S.; Devoisselle, J.M.; Goursot, A.; Zhu, R.; Salahub, D.R. Density functional theory-based conformational analysis of a phospholipid molecule (dimyristoyl phosphatidylcholine). *J. Phys. Chem. B,* **2008**, *112*, 13433-13442.

[100] Goursot, A.; Mineva, T.; Krishnamurty, S.; Salahub, D. R. Structural analysis of phosphatidyl choline lipids and glycerol precursors. *Can. J. Chem.,* **2009**, *87*, 1261-1267.

[101] Kobayashi, T.; Beuchat, M. L.; Chevallier, J.; Makino, A.; Mayran, N.; Escola, J. M.; Lebrand, C.; Kobayashi, T.; Gruenber, J. Membrane transport structure function and biogenesis. *J. Biol. Chem.*, **2002**, *277*, 32157-32164.

[102] Kobayashi, T.; Stang, E.; Fang, K. S.; de Moerloose, P.; Parton, R. G.; Gruenberg, J. A lipid associated with the antiphospholipid syndrome regulates endosome structure and function. *Nature,* **1998**, *392*, 193-197.

[103] Bissig, C.; Lenoir, M.; Velluz, M.C.; Kufareva, I.; Abagyan, R.; Overduin, M.; Gruenberg, J. Viral Infection Controlled by a Calcium-Dependent Lipid-Binding Module in ALIX. *Developmental Cell,* **2013**, *25*, 364-373.

[104] Matsuo, M.; Chevallier, J.; Mayran, N.; Le Blanc, I.; Ferguson, C.; Faure, J.; Sartorie Blanc, N.; Matile, S.; Dubochet, J.; Sadoul, R.; Parton, R. G.; Vilbois, F.; Gruenberg, J. Role of LBPA and alix in multivesicular liposome formation and endosome organization. *Science,* **2004**, *303*, 531-534.

[105] IUPAC The Nomenclature of Lipids. *Eur. J. Biochem.,* **1977**, *79*, 11-21.

[106] Kaupp, M.; Bühl M.; Malkin, V.G., Eds. *Calculations of NMR and EPR parameters: Theory and Applications*; Wiley-VCH Verlag GmbH & Co.: Weinheim, **2004**.

[107] Wüthrich K. *NMR of Proteins and Nucleic Acids*; Wiley, New York, **1986**.

[108] Hennel, J.W.; Klinowski, J. Magic-angle spinning: a historical perspective. In *New Techniques in Solid-State NMR*; Klinowski, J., Ed.; Springer-Verlag: Berlin, **2005**; Vol. *246*, pp. 1-14.

[109] McDermott, A.E.; Polenova, T., Eds. *Solid-State NMR Studies of Biopolymers*, John Wiley & Sons: Singapore, **2010**.

[110] Mackenzie, K.J.D.; Smith, M.E. Multinuclear solid-state NMR of inorganic materials. In Pergamon Materials Series; Cahn, R.W., Ed.; Pergamon Elsevier Science, Amsterdam, **2002**.

[111] Asakawa, N.; Kuroki, S.; Kurosu, H.; Ando, I.; Shoji, A.; Osaki, T. Hydrogen-bonding effect on carbon-13 NMR chemical shifts of L-alanine residue carbonyl carbons of peptides in the solid state. *J. Am. Chem. Soc.*, **1992**, *114*, 3261-3265.

[112] Deetz, M.J.; Fahey, J.E.; Smith, B.D. NMR studies of hydrogen bonding interactions with secondary amide and urea groups. *J. Phys. Org. Chem*., **2001**, *14*, 463-467.

[113] Kameda, T.; Ando, I. The relationship between the helical conformation and ^{13}C NMR chemical shift of amino acid residue carbonyl carbons of polypeptides in the solid state. *J. Molec. Struc*., **1997**, *412*, 197-203.

[114] Alexandrescu, A.T.A.; Synder, D.R.; Abildgaard, F. NMR of hydrogen bonding in cold-shock protein A and an analysis of the influence of crystallographic resolution on comparisons of hydrogen bond lengths. *Protein Sci*., **2001**, *10*, 1856-1868.

[115] Aliev, A.E.; Law, R.V. Solid State NMR Spectroscopy. In *Nuclear Magnetic Resonance*; Webb, G.A., Ed.; The Royal Society of Chemistry: Cambridge, **2008**.

[116] Heller, D.M.; Giorgetti, A. NMR Constraints Analyser: a web-server for the graphical analysis of NMR experimental constraints. *Nucleic Acids Research*, **2010**, *38*, W628-W632.

[117] Hauser, H.; Pascher, I.; Sundell, S. Preferred conformation and dynamics of the glycerol backbone in phospholipids. An NMR and x-ray single-crystal analysis. *Biochemistry,* **1988**, *27*, 9166-9174.

[118] Hong, M.; Schmidt-Rohr, K.; Zimmermann, H. Conformational Constraints on the headgroup and sn-2 chain of bilayer DMPC from NMR dipolar couplings. *Biochemistry,* **1996**, *35*, 8335-8341.

[119] Kitahara, R.; Yokoyama, S.; Akasaka, K. NMR Snapshots of a fluctuating protein structure: Ubiquitin at 30 bar-3 kbar. *J. Mol. Biol*., **2005**, *347*, 277-285.

[120] Li, H.; Akasaka, K. Conformational fluctuations of proteins revealed by variable pressure NMR. *Biochim. Biophys. Acta*, **2006**, *1764*, 331-345.

[121] Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Shaw, D.E. How fast-folding proteins fold. *Science*, ***2011****, 334, 517-520.*

[122] Woolf, T.B.; Malkin, V.G.; Malkina, O.L.; Salahub, D.R.; Roux B. The backbone ^{15}N chemical shift tensor of the gramicidin channel. A molecular dynamics and density functional study. *Chem. Phys. Lett.*, **1995**, *239*, 186-194.

[123] Aliev, A.E.; Courtier-Murias, D.; Bhandal, S.; Zhou, S. A combined NMR/MD/QM approach for structure and dynamics elucidations in the solution state: pilot studies using tetrapeptides. *Chem. Commun*., **2010**, *46*, 695-697.

[124] Aliev, A.E.; Courtier-Murias, D. Experimental verification of force fields for molecular dynamics simulations using Gly-Pro-Gly-Gly. *J. Phys. Chem. B*, **2010**, *114*, 12358-12375.

[125] Krishnamurty, S.; Heine, T.; Goursot, A. Influence of dynamics on the structure and NMR chemical shift of a zeolite precursor. *J. Phys. Chem B*, **2003**, *107*, 5728-5732.

[126] Heine, T.; Vietze, K.; Seifert, G. ^{13}C NMR fingerprint characterizes long time-scale structure of Sc3N@C80 endohedral fullerene. *Magn. Reson. Chem*., **2004**, *42***,** S199-S201.

[127] De Gortari, I.; Portella, G.; Salvatella, X.; Bajaj, V.S.; van der Wel, P.C.A.; Yates, J.R.; Segall, M.D.; Pickard, C. J.; Payne, M. C.; Vendruscolo, M. Time averaging of NMR chemical shifts in the MLF peptide in the solid state. *J. Am. Chem. Soc.,* **2010**, *132*, 5993-6000.

[128] Özarslan, E; Shemesh, N; Basser, P.J A general framework to quantify the effect of restricted diffusion on the NMR signal with applications to double pulsed field gradient NMR experiments. *J. Chem. Phys*., **2009**, *130*, 104702-1 - 104702-9.

[129] Robinson, M.; Haynes, P.D. Dynamical effects in *ab initio* NMR calculations: Classical force fields fitted to quantum forces. *J. Chem. Phys.,* **2010***, 133,* 084109-1-084109-9.

[130] Mineva, T.; Gaveau, P.; Galarneau*,* A.; Massiot, D*.;* Alonso B. ^{14}N: A sensitive NMR probe for the study of surfactant-oxide interfaces. *J. Phys. Chem. C*, **2011**, *115*, 19293-19302.

[131] Mao, C.; Oser, Z.; Zhou, M.; Uckun F.M. X-Ray Structure of glycerol kinase complexed with an ATP analog implies a novel mechanism for the ATP-Dependent glycerol phosphorylation by glycerol kinase. *Biochem. Biophys. Res. Commun*., **1999**, *259*, 640-644.

[132] Bystrom, C.E.; Pettigrew, D.W.; Branchaud, B.P.; O'Brien, P.; Remington, S.J. Crystal structures of Escherichia coli glycerol kinase variant S58→W in complex with nonhydrolyzable ATP analogues reveal a putative active conformation of the enzyme as a result of domain motion. *Biochemistry,* **1999**, *38*, 3508-3518.

[133] Fu, D.; Libson, A.; Miercke, L.J.W.; Weitzman, C.; Nollert, P.; Krucinski, J.; Stroud, R.M. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **2000**, *290*, 481-486.

[134] Tuttle, T. Averaging Semiempirical NMR chemical shifts: Dynamic effects on the subpicosecond time scale. *J. Phys. Chem. A*, **2009**, *113*, 11723-11733.

[135] Koster, A. M.; del Campo, J. M.; Janetzko, F.; Zuniga-Gutierrez, B. A MinMax self-consistent-field approach for auxiliary density functional theory. *J. Chem. Phys.*, **2009**, *130*, 114106-1 - 114106-8.

[136] Zuniga-Gutierrez, B.; Geudtner, G.; Köster, A.M. NMR shielding tensors from auxiliary density functional theory. *J. Chem. Phys*., **2011**, *134*, 124108.

[137] Zhang, Y.; Yang W. Comment on "Generalized gradient approximation made simple". *Phys. Rev. Lett*., **1998**, *80*, 890-890.

[138] Lee, C.; Yang, W.; Parr, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, **1988**, *37*, 785-789.

[139] Goursot, A.; Mineva, T.; Kevorkyants, R.; Talbi D. Interaction between n-alkane chains: Applicability of the empirically corrected density functional theory for van der Waals complexes. *J. Chem. Theory Comput*., **2007**, *3*, 755-763.

[140] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys*., **1984**, *81*, 511.

[141] Hoover, W. Enhanced interaction energy shifts in the x-ray spectrum of ^{163}Ho. *Phys. Rev. A*, **1985**, *31*, 1965- 1967.

[142] Dichtfield, R. Molecular orbital theory of magnetic shielding and magnetic susceptibility. *J. Chem. Phys*., **1972**, *56*, 5688.

[143] Perdew, J.P.; Chevary, J.A.; Vosko, S.H.; Jackson, K.A.; Pederson, M.R.; Singh, D.J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B*, **1992**, *46*, 6671-6687. Erratum in: *Phys. Rev. B*, **1993**, *48*, 4978-4978

[144] Dunning Jr., T.H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys*., **1989**, *90*, 1007-1023.

[145] Helgaker, T.; Jaszunski, M.; Ruud, K. *Ab initio* methods for the calculation of NMR shielding and indirect spin−spin coupling constants. *Chem. Rev.*, **1999**, *99*, 293-352.

[146] Del Campo, J.M.; Köster A.M. A hierarchical transition state search algorithm. *J. Chem. Phys*., **2008**, *129*, 024107.

[147] Vásquez-Pérez, J.M.; Gamboa Martínez, G.U.; Köster, A.M.; Calaminici P. The discovery of unexpected isomers in sodium heptamers by Born-Oppenheimer molecular dynamics. *J. Chem. Phys*., **2009**, *131*, 124126.

[148] Culot, P.; Dive, G.; Nguyen, V.H.; Ghuysen, J.M. A quasi-Newton algorithm for first-order saddle-point location. *Theor. Chim. Acta*, **1992**, *82*, 189-205.

[149] Bofill, J.M. Updated Hessian matrix and the restricted step method for locating transition structures. *J. Comput. Chem*., **1994**, *15*, 1-11.

[150] Gonzalez, C.; Schlegel, H.B. An improved algorithm for reaction path following. *J. Chem. Phys*., **1989**, *90*, 2154.

[151] Vásquez-Pérez, J.M. Propiedades termodinámicas de sistemas finitos a partir de dinámica molecular de Born-Oppenheimer. PhD, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional: Mexico City, October **2011**.

[152] London, R.E.; Wingrad, B.D.; Mueller, G.E. Dependence of amino acid side chain ^{13}C shifts on dihedral angle: Application to conformational analysis. *J. Amer. Chem. Soc*., **2008**, *130*, 11097-11105.

[153] Chelli, R.; Gervasio, F. L.; Gellini, C.; Procacci, P.; Cardini, G.; Schettino, V. Conformational distribution of gas-phase glycerol. *J. Phys. Chem. A,* **2000**, *104*, 11220-11222.

[154] Callam, C. S.; Singer, S. J.; Lowary, T. L.; Hadad, C. M. Computational analysis of the potential energy surfaces of glycerol in the gas and aqueous phases: effects of level of theory, basis set, and solvation on strongly intramolecularly hydrogen-bonded systems. *J. Amer. Chem. Soc.,* **2001**, *123*, 11743-11754.

[155] Chelli, R.; Procacci, P.; Cardini, D.; Valle, R. G.; Califano, S. Glycerol condensed phases Part I. A molecular dynamics study. *Phys. Chem. Chem. Phys.,* **1999**, *1*, 871-877.

[156] Chelli, R.; Procacci, P.; Cardini, D.; Valle, R. G.; Califano, S. Glycerol condensed phases Part II. A molecular dynamics study of the conformational structure and hydrogen bonding. *Phys. Chem. Chem. Phys.,* **1999**, *1*, 879-885.

[157] Yongye, A. B.; Foley, B. L.; Woods, R. J. On achieving experimental accuracy from molecular dynamics simulations of flexible molecules: Aqueous glycerol. *J. Phys. Chem. A,* **2008**, *112*, 2634-2639.

[158] Kaminski, G.A.; Friesner, R.A.; Tirado-Rieves, J.; Jörgensen, W.L. Evaluation and reparametrization of the OPLS-AA force field for proteins *via* comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, **2001**, *105*, 6474.

[159] Maccaferri, G.; Caminati, W.; Favero, P.G. Free jet investigation of the rotational spectrum of glycerol. *J. Chem. Soc., Faraday Trans*., **1997**, *93*, 4115-4117.

[160] Spectral Database for Organic Compounds. http://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi.

[161] Voelter, W.; Breitmaier, E.; Jung, G.; Keller, T.; Hiss D. Fourier transform ^{13}C-NMR spectroscopy of biologically active cysteine peptides. *Ang. Chemie*, **1970**, *82*, 812.

[162] Zhang, B.L.; Buddrus, S.; Trierweiler, M.; Martin, G.J. Characterization of glycerol from different origins by ^{2}H- and ^{13}C-NMR studies of site-specific natural isotope fractionation. *J. Agric. Food Chem*., **1998**, *45*, 1374-1380.

CHAPTER 8

Recent Advances in Computational Simulations of Lipid Bilayer Based Molecular Systems

R. Galeazzi*, E. Laudadio and L. Massaccesi

Dipartimento di Scienze della Vita e dell'Ambiente (DISVA) - Università Politecnica delle Marche- via Brecce bianche, 60128 Ancona, Italy

Abstract: Computer simulations in lipid bilayers research has become prominent for the last couple of decades. As computational resources became more available to the scientific community, simulations play an increasingly important role in understanding the processes that take place in and across cell membranes. The scientific interest is strictly related to the Biological importance of the Biomembranes, which act as barriers separating cell's internal environment from the external one. Membranes are selectively permeable, and thus they actively participate in the movement control of compounds into and outside cells. These membranes have an heterogeneous complex composition and they include many different lipids together with proteins, steroids, carbohydrates and other membrane-associated molecules. Each of these compounds are involved in a great number of cellular processes and thus, membranes exist as dynamic structures. As a consequence, the understanding of biomembrane functioning requires the knowledge of chemical-physical behavior of lipid bilayers and it represents a great challenge in biophysical and medical sciences.

In the last decades, molecular dynamics (MD) simulations have become one of the most useful tool in the *in silico* investigations of molecular structures; in fact, such computations provide structural dynamical information which is essential and hardly obtained by experimental methods; furthermore, it furnishes a system real-time imaging at atomistic-level resolution. In this chapter, we want to point out the recent advances in computer simulations in the field of lipid bilayers and proteins-lipid bilayers systems during the last few years, by covering several selected subjects such as state of art in *ad hoc* force fields' development, Cholesterol induced effects on structure and properties of the bilayer, mixed composition lipid matrix, and biomolecular application of coarse-grained models.

Keywords: Atomistic Molecular Dynamics, Coarse grained simulations, membrane simulations, lipid bilayers, lipid bilayers force fields, liposomial nanovectors, membrane proteins, membrane properties predictions.

***Corresponding author Roberta Galeazzi:** Dipartimento di Scienze della Vita e dell'Ambiente, Università Politecnica delle Marche, *via* Brecce bianche, 60128 Ancona, Italy; Tel: +390712204724;
E-mail: r.galeazzi@univpm.it

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

INTRODUCTION

Molecular processes regulated by Biomembranes are critical for living cells' normal function. The features of a cell membrane are mainly determined by a lipid bilayers. These membranes have an heterogeneous complex composition and they include many different lipids together with proteins, steroids, carbohydrates and other membrane-associated molecules, involved in a great number of cellular processes. Thus, Membranes exist as dynamic structures: they are selectively permeable, and they actively participate in the movement control of compounds into and outside cells. Indeed, observing the great diversity in the structure and composition of biological membranes, the importance of lipids' nature in many cellular processes is coming out.

More in general, main differences in the lipid molecular structure reside in the hydrophilic polar head-groups and in the lipophilic hydrocarbon chains of the fatty acids (FAs). The most abundant phospholipid in animal and plants is phosphatidyl-choline (PC), which is a key building block of membrane bilayers. Beside it, cholesterol (Chol) can be found as an essential component of mammalian cell membrane. This steroid plays an important role in formation of heterogeneities (known also as *rafts*) which are supposed to be responsible for cell signaling. Thus, it is clear that the understanding of biomembrane functioning requires the knowledge of physical-chemical properties of lipid bilayers and it is one of the greatest challenging problems in biophysical and medical sciences.

Moreover, the elucidation of the complex interactions between proteins and membrane, represents a current intriguing challenge in the field of structural biology.

Indeed, the spectacular growth of membrane simulations in the last 10-15 years had led to a better overall picture of membrane systems at atomic resolution, when the only employment of experimental methods is often insufficient. Thus, the usage of high-performance computing (HPC) allows investigation of complex membrane molecular systems using powerful molecular modeling techniques such as molecular dynamics (MD) (both atomistic MD and coarse grained CG-MD). In particular, full atoms MD simulations reproduce the motions of each atom in the simulation, by using an empirical potential energy function and it

provides molecular atomistic interactions and energetic details which are generally hard to obtain from experiments. Thus, it represents a critical additive information for the full comprehension of membrane macroscopic behavior.

The timeline for the first attempts of computer simulations of model bilayers with atomistic resolution might begin in the 1980's with the early simulation of a solvent-free decanoate lipid bilayers [1-3] and with water between rigid lipid headgroups [4] which was followed in the 1990's by simulations of fully hydrated bilayers formed by lipids generally found in biological membranes, with both phospholipids and water represented with atomic details [5, 6]. Later on, the development of HPC power has made feasible simulations of more huge complex systems, with increased size. At present, simulation of hundred hydrated lipids for a time length of 50-100 ns, is considered routinely and can be easily extended for a much longer simulation time (1 μs-1ms) [7, 8]. These achievements led to the construction of a large set of model bilayers and micelles differing in lipid composition. Such models are used to study different classes of membrane-active proteins (ion channels, GPCR receptors) and peptides (such as toxins, antimicrobial and fusion peptides) [9].

Of course, since the amount of works on simulations of lipid membrane systems has grew up significantly, many reviews related to this topic appeared in the past decade [10-29] and more recently [30-48]. In this chapter, we want to point out the recent advances of *in silico* simulations of lipids bilayer and proteins-lipid bilayers systems, in different selected subjects such as development of the force fields for lipid bilayer simulations, application of both atomistic simulations and coarse-grained models, also providing *state-of-art* current computational protocols to settle out the simulations. More in detail, we highlight those cases where simulations are used in conjunction with experimental techniques, covering topics such as simulations including the Cholesterol effect, membrane protein simulations and simulation involving relevant changes in basic bilayers structure (mixed composition bilayers or salt effect prediction) [49-51].

FORCE FIELDS FOR LIPID BILAYERS SIMULATIONS

As papers describing simulations of lipid bilayers emerged [1-6], it was immediately clear that much work needed to be done to reach the desiderate

accuracy. A critical component for this aspect has been the development of a reliable force field (FF) potential energy function. A simulation can't proceed without it, and the availability of the appropriate FF is among the first issues that a researcher must confront when considering a new system. Furthermore, proper parameterizations in the force field developed is necessary. This is a technical aspect which needs continuous attention in molecular simulations since a force field is good if it provides agreement with all available experimental data, taking into account for the simulation and experimental uncertainty. It is true that experimental techniques are improving and thus, even if a force field have always provided satisfactory agreement with experimental data, it may later begin to show discrepancies. The only way to solve this problem is further and continuous improvements of the force field which can lead to a better description of the molecular interactions and a better agreement with the experimental data.

The three main force fields, which have been tested and used in recent years in lipid bilayers simulations, are GROMOS [52-54], AMBER [55] and CHARMM [56, 57].

GROMOS uses united atoms approach representing each of apolar CH_3, CH_2, and CH groups of hydrocarbons as a single particle, allowing in this way to reach about 3-fold speedup comparing to other atomistic simulations. The GROMOS force field are developed in many versions which can be divided into two main groups: one with original GROMOS non-bonded parameters (for example, 45A3 and similar parameter sets [58]), and the second with the Berger modification [59] of parameters, which is actually the most frequently used. In the latter one, the Ryckaert-Bellemans potential is implemented to describe torsional rotations of the lipidic hydrocarbon chains. The GROMOS force field is native and continuously implemented in GROMACS simulation software package [60].

Beside, AMBER has been implemented to treat at best lipid systems only in recent years. The release of Amber 14 [61] includes Lipid14 [62], a lipid force field suitable for the dynamics of phospholipids. Lipid14 derives directly from Lipid 11 which was originally developed to be fully compatible with the other pairwise-additive AMBER-based force fields. Currently, it includes parameters from the General Amber Force Field (GAFF), a novel charge derivation, as well as ongoing

refinements of parameters for phospholipids. Furthermore, in this implementation within the AMBER MD package, it allows tensionless simulation of a wide number of lipid types. In addition, the modular nature of this force field provides a large number of combinations of head and tail groups to create different lipid types, permitting in this way the easy insertion of new lipid species. It is totally compatible with the standard AMBER force fields built for protein, nucleic acid, carbohydrate, thus it enables simulation of hybrid molecular systems.

Finally CHARMM force field [56] describes all hydrogens explicitly and within this FF, a more detailed description of intramolecular interactions is provided. Thus, the Urey-Bradley term for covalent angles and a richer variety of parameters for dihedral angles are included. Concerning the force field (FF) parametrization, CHARMM parameters for lipids were introduced first by Feller *et al.* [63] (Charmm22 parameter set or referred as C22) and then were updated in Charmm27 (c27) parameter set [12, 64] and also in its extension C27r. The CHARMM force field was firstly born in the original CHARMM software, and it is originally present in NAMD simulation package [65]. It is also implemented in a number of other simulation packages such as GROMACS. Successively, the CHARMM36 lipid force-field have been developed which have led to a quantitative accuracy for many membrane-protein properties predicitons [66]. The limit is the membrane dipole potential which is incorrectly predicted unless some form of polarization is included within the force-field. Indeed, non-polarizable FFs attempt to reproduce many-body polarization effects in an averaged way, using partial atomic charges that are invariant to their electrostatic environment. To solve this problem, recently there has been some efforts to introduce polarization for the common lipids found in cell membranes [67-69]. However, there still remains only non-polarizable FFs for the treatment of other membrane molecules as sphingomyelin (SM) and Cholesterol (Chol) [70, 71].

Recent applications and developments of such FFs in classical MD simulations are discussed in reviews [72, 73]. Among them, very recently, Robinson *et al.* developed a new polarizable force-field which could be successfully used to treat the major components of membrane raft domains and it is consistent with previous force fields of the same class [74]. Using this novel approach, the properties computed show good agreement with the available experimental data

for a wide variety of properties and also represent an improvement over the non-polarizable FFs for the above mentioned critical properties (*i.e.* membrane dipole potential and induced dipole moments). In one of its applications, it has been used to perform simulations on membrane raft-like domains using the polarizable Drude oscillator model and it provided qualitative and quantitative agreement with experimental details.

RUNNING MOLECULAR DYNAMICS SIMULATION FOR LIPID BILAYERS SYSTEMS

Possible Pathways to Generate a Mixed Lipid Matrix

There are many different pathways to generate an homogenous or heterogeneous lipid matrix. Most of them are extensively and clearly reported in some recent tutorials, as reported hereby:

- *http://www.gromacs.org/Documentation/Tutorials#Membrane_Simulations*
- *http://ambermd.org/tutorials/advanced/tutorial16/*
- *http://www.ks.uiuc.edu/Training/Tutorials/#specializedtopics*

Other methods are reported in recent papers [75-78]. In a typical approach, to build an homogenous membrane lipid, the reference lipid molecule oriented along molecular axis z (perpendicularly to the membrane face, the xy plane), is randomly rotated and copied several times on both sides of the lipid bilayers [75]. Water molecules are added to fill in the gaps above and below the phospholipids leaflet to hydrate the lipid head groups. In addition molecules such as SM (sphingomyelin) and Chol can be inserted in order to correctly represent a cell membrane model, which must be inherently asymmetric and heterogeneous for composition. This procedure keep intact the overall number of phospholipids in both the leaflets matrix [75]. Indeed, the asymmetry of lipid distribution can be generated by substitution of certain number of phospholipids in one of the two leaflets by same number of another type of membrane lipid [76-78].

Other pathways to obtain the lipid bilayers structure is the spontaneous aggregation runs generated using the GROMACS tool *genbox* and it consists in repeatedly place

one of eight different phospholipids' conformations from a *vacuum* ensemble with random orientations into a simulation box of a given size. Water can be added using the same tool to complete the model. This procedure must be repeated with different seeds for the random number generator and for different sets of lipid configurations in order to generate independent starting structures.

Often, Na^+ or Ca^{2+} ions are added substituting randomly a number of water molecules if it is necessary to balance the net charge associated with the presence of certain type of phospholipids [76]. At this point, the excess of energy associated to the eventual overlaps of neighboring atoms must be removed throught a full minimization process.

Very recently, free web servers have been developed in order to automate and simplify the building process of heterogeneous lipid bilayers. They also provide molecular topologies for most common lipids which can be used on both united and all-atom force fields. Some examples include MemBuilder (www.membuilder.org), a web-based graphical user interface have been implemented for using with GROMACS [79] and other software packages such as PACKMOL [80], CHARMM GUI (http://www.charmm-gui.org/input/membrane), web-based graphical user interface supporting up to 32 different lipid types [81] and the most recent LIPID Builder (http://lipidbuilder.epfl.ch/builder) [82].

Settling out Force Field and Simulation Conditions

As already reported in the previous section, many force fields have been developed to treat lipid bilayers systems. However, the proper choice of force field potential and parameters is always critical for accurate simulation results, since it must be directly related to the model and the system under test. As a first choice, the basic strategy to describe phospholipids is based on the force field developed by Berger *et al.* [59]. Other ways, the *cis* double bond in the oleoyl chains at position C-9 can be modeled successfully using torsional parameters of GROMOS96 FF. Within the same force field, the parameters for the lysine ammonium group are used to model the phosphatidyl-ethanolamine (PE) headgroup, following Tieleman and Berendsen's model for POPE [83] and considering that a small repulsive potential must be put on the ammonium

hydrogens. In all the simulations it must be settled the time step (for example 2 fs), the *cutoff* (for example 1.0 nm) for the Lennard-Jones interactions and the Ewald Particle Mesh method in order to evaluate the electrostatic interaction. All the MD simulations should be performed using an anisotropic-coupling pressure, keeping the pressure about 1 atm, value which allow the independent fluctuation of each axis of the computer box under periodic boundary conditions (PBC). The SPC (Berendsen *et al.*) [84] or the better TIP3P [85] water model is adequate for this kind of simulation. In these conditions, the systems achieves an equilibrated state after a certain number of ns of simulation (usually 30-50 ns), which can be stated observing the average surface area per lipid for a specific type of membrane lipid leaflet. The data analysis must be performed on trajectories only after that the average lipid area of the bilayer had converged and the box dimensions had stabilized, thus collecting data from the last ns (for example 20-25, chosen for most systems) of the trajectories. Most analysis can be performed using standard GROMACS tools or other software packages such as VMD [86].

The volume per lipid V_l is calculated from the box volume V_{box} by subtracting the volume of pure SPC or TIP3P water molecules under the same conditions and dividing by the number of lipids in the box. The area per lipid of the bilayer is calculated from the lateral box dimensions and it refers to its projection onto the lateral plane. More details can be found in the reference papers [78, 87]. Other type of systems use the force field parameters for standard lipids (*i.e.* POPC, DPPC) as included in the validated united-atom description of Tielman and Berendsen [83], whereas parameters from works of Niemela *et al.* [88] are used for SM and Chol.

Mostly, GROMACS [60] or AMBER [61] software are generally used to integrate the equations of motion and all bilayer systems are initially minimized using steepest descent algorithm to avoid bad contacts arising from steric constraints [89]. Other choose of force field and parameterization is described below.

Most frequently the recent version of the CHARMM force field (CHARMM27) [66] is used, since it has been shown to have brilliant performances for saturated lipids and for lipids with a low degree of unsaturation.

An example of a well tested protocol is reported herein. In detail, the minimized bilayers are simulated first in the NVT ensemble using a Langevin thermostat, and subsequently in the NPT ensemble, using a Berendsen thermostat and semi-isotropic pressure coupling. Often, when a high degree of unsaturation is present, a long equilibration period (such as 2.0 ns) is needed to achieve a convergence of the dimensions of the system. After the equilibrium reached, the dimensions of the lipid bilayers can be monitored over an additional 1 ns were the area per lipid and lamellar spacing, have been defined accurately [90].

Finally, only in recent years, computer simulation approaches have been used to unravel the molecular details of membrane fusion. Over the past decade, fusion between apposed membranes and vesicles has been studied using a large variety of simulation methods and systems. Initially the full fusion of two flat membranes was studied by means of Monte Carlo (MC). These computational approach, however, have an implicit problem associated to the boundary conditions which may influence the results of simulations; in fact, the volume of solvent between the two fusing membranes must remain constant until a fusion pore opens to allow the solvent to flow into the spaces behind them. However, the final technique used, which at the expense of higher computational cost can provide a more accurate chemical representation, is once again molecular dynamics (MD). Because of the computational cost coarse-grained (CG) models are introduced into the fusion field (see in next CG section). The use of MD allowed, for instance, to discriminate between different types of lipids and in addition interesting differences are found in the fusion pathways between, for example, PC and PE lipids. More recently also simulations at atomistic detail have been reported of full vesicle fusion. The first such study was performed by Knecht and Marrink on a highly fusogenic mixture of DPPC and palmitic acid [91]. Because of the high computational cost of such atomistic MD simulations, it is not possible to considerer the self-fusion of a very large vesicles (not over 15 nm) with its periodic image. Noteworthy is also the introduction of distributed computing to the fusion arena by Kasson *et al.* [92] and indeed the availability of thousands of processor nodes allows for systematic studies on fusion pathway statistics. Although simulations show that the choice of fusion pathway is stochastic, that is, the same system can follow multiple pathways, membrane composition does play a role in the prevalence of one pathway over the other [91, 92].

FULL ATOMS MOLECULAR DYNAMICS OF LIPID BILAYERS: STATE OF ART AND CASE STUDIES

Over the last few decades, among the other computational techniques that have emerged in Science, the atomistic Molecular Dynamics (MD) simulation have been extensively implemented and used from scientists all over the globe to obtain insight with atomic detail of steady and dynamic properties of lipid bilayers based supramolecular systems. In this regard, a critical aspect that must be identified in all the MD simulations is related to the time and length scale of process observed, the ensamble used, the pressure and temperature control.

However, the amount of works on simulations of lipid membranes has achieved higher accuracy and prediction capabilities; the scale of systems that can be studied continues to increase with the increasing of the computer power [4-50] and with the improvement of the MD simulation software.

At present it can be identified a general limit on simulations which can be up to 5000 lipids in full atoms MD simulations and up to 50000 in coarse grained models. Concerning simulation time length, it must be pointed out that the equilibrium motions in lipid bilayers range from picosecond time scales (rattling of individual lipid hydrocarbon tails) to over many microseconds (collective motion of all lipids, μm size dimension). Those motions can be investigated experimentally by means of spectroscopy methods (the former) and microscopy (the latter). Actually, hundreds of ns have became routinely for full atoms MD in bilayers, but if a further extension of the simulation time lengths is required, it must be necessary to refer to the coarse grained approach (up to ms-s time length). Usually, periodic boundary conditions (PBC) are able to avoid strong artifacts from presence of boundary planes, and this way a stack of bilayers with infinite dimension is simulated.

Finally, a particular attention in simulations must be taken to identify the equilibration time to reach a steady state. To this purpose, Porasso *et al.* [93] studied lipid bilayers with different composition in their liquid crystalline phase in order to establish a general criterion to identify the reaching of such a state. In

particular they studies the dynamical temporal evolution of some lipids properties (*i.e.* area per lipid, the deuterium order parameter, the lipid hydration and lipid-salt coordination) and they observed that the time required depends strictly on the single property studied. It has been found out the following order from faster equilibration property to slower: coordination of ions = deuterium order parameter > area per lipid=hydration. As a consequence, when the mean area per lipid or hydration of lipids are stable, we can ensure that the lipid membrane has equilibrated to the steady state.

Live cell membranes can be constituted possibly by thousands of unique lipid species [94], and the complete *in silico* reproduction of such complex systems is quite impossible. For this purpose, *ad hoc* model systems have been built to gain a deep insight into the way the lipids mix and these kind of simulations have developed in parallel with the study of Cholesterol containing bilayers. At the start, such simulations explored interactions between steroids and other lipids, in particular local interactions often reproduced in terms of highly stylized umbrella or other models. More systematic studies which may be extended to a length scale where domains might form, have been out of reach until recently. Beside, the condensing effect of Cholesterol on phospholipid bilayers can be reproduced both with all-atoms and Coarse Grained models [95]. However, atomistic simulations have not been the primary choice to directly observe lipid domain formation in membranes, due to its intrinsic computational constraints. In fact, the dimension of a small domain is considered around 10 nm in diameter and this still represents a rather large simulation by today's standards.

Cholesterol's Inclusion Effects on Properties of the Bilayers: From Early to Recent Studies

Cholesterol amounts up to 50 mol% of the membrane lipids in eukaryotic cell membranes and thus represent an important constituent [96]. Its biological roles involve the maintenance of proper fluidity [97], the reduction of passive permeability [98], and the increase of the mechanical strength [99] of the membrane. In particular, the Cholesterol's ability to condense fluid phospholipid membranes has been known for almost a century [100]. Because of all these

important biological roles, phospholipid-Chol interactions in the membrane have been studied widely [101], revealing that the steroid increases both the order of the hydrocarbon chains (*ordering effect*) [102, 103] and the surface density of the membrane (*condensing effect*) [104, 105]. Both effects are important since they have been observed in biological and in model membranes [106]. Several studies concerning molecular dynamics (MD) simulation of dimyristoylphosphatidyl-choline-cholesterol (DMPC-Chol) bilayers, focused on the effect of Chol on the organization of the membrane-water interface [107] and the ordering of DMPC alkyl chains [108]. However, at present the ways cholesterol induces phospholipid condensation is poorly understood, despite the numerous experimental and theoretical investigations which have been carried out. The umbrella mechanism have been proposed and according this model, cholesterol and lipid acyl chains become more tightly packed as the steroid content increases, since they share limited space under phospholipid headgroups (Fig. **1**) [109]. In an alternate proposed mechanism, the flexible acyl chains of the phospholipid complement perfectly cholesterol's planar nucleus to produce a high number of close hydrophobic contacts and tight packing [110]. In a recent work, Regen *et al.* distinguishes between these two mechanisms on the basis of an experimental evidence and shed light into the origin of Chol's condensing effect [111].

Beside the pure mechanicistic studies, Chol's preference for specific fatty acid chains was investigated in MD computer simulations [112] of a lipid bilayers membrane consisting of Chol and 18:0/22:6(n-3)*cis*-phosphatydylcoline (PC) in a 1:3 ratio. Three bilayer compositions were investigated [*i.e.* 18:0/22:6(n-3)*cis*-PC, 18:0/22:5(n-6)*cis*-PC, and 18:0/22:6(n-3)*cis*-PC with 25 mol % Chol] and it came out that the distribution of lateral stress within the hydrophobic core of the membrane is sensitively dependent on the degree of chain unsaturation and on the presence of the steroid. Moreover, the bond order parameters in the phospholipid tails showed an increase with addition of the steroid which is proportional also to the Chol content [112-116]. Furthermore, it has been observed that the steroid promotes the formation of highly ordered *raftlike* membrane domains in presence of saturated lipids, whereas domains rich in unsaturated lipids with a double bond in the middle remain highly fluid despite the presence of Chol [117].

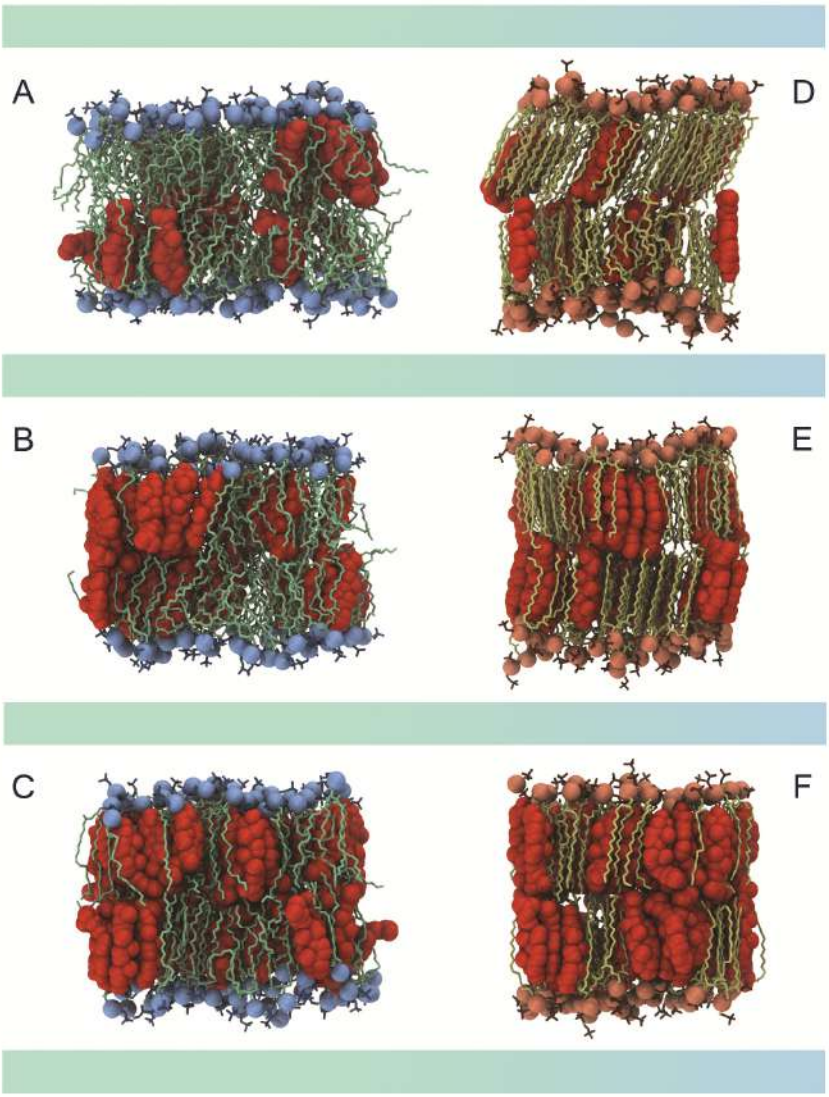


Figure 1: DOPC and DPPC bilayers arrangements at increasing concentration of Chol. A) DOPC + Chol 20% B) DOPC + Chol 30% C) DOPC + Chol 50% D) DPPC + Chol 20% E) DPPC + Chol 30% F) DPPC + Chol 50%

Furthermore, cholesterol's effect on the distribution of the terminal methyl groups of DOPC has been investigated combining neutron scattering with molecular dynamics simulations [118]. The results showed a higher disorder in the DOPC

bilayer than was previously thought based on X-ray scattering profiles, with methyl groups reaching the head group region. The extent of the broad methyl density dramatically decreased when a higher Chol amount have been incorporated (33 mol%), and this behavior can be ascribed to the ordering of the bilayer and the straightening of the lipid tails. Full atoms simulations reproduced both the extent of the DOPC tail's disorder and the effect of the steroid on reducing the disorder.

Finally, Olsen and Baker investigated the difference between the same percentage of cholesterol and of oxysterol (40% mol) on a lipid membrane, showing dramatic differences between them [119]. They extended this work to study other oxysterol and Chol mixtures [120]. Other recent *in silico* simulations are related to the identification of interactions involving glycolipids in model membranes [121-123], phosphoinositides [124], sphingomyelin [125], diacylglycerol [126], and other sterols [127, 128], model yeast membrane [129], bacterial model membranes [130], asymmetric cholesterol and sphingomyelin mixtures [131].

Cholesterol and Nature of Lipid Rafts

The role of the steroid in regulating raft stability, and organization in animal cells, was highlighted in several studies carried out on phospholipid monolayers containing cholesterol. Indeed, many groups studied membrane domains with high-Chol concentrations [132-136], which are often referred to as rafts. These domains are supposed to be important in cellular signaling, although final experimental evidence *in vitro* is not yet attained [137].

The existence of lipid domains had been already suggested on earlier work, but the lipid raft hypothesis was conceived as an explanation of the difference in membrane sorting between the apical and basal sides of epithelial cells (Fig. **2**) [132]. During last, the nature of lipid raft domains in model membranes have been studied extensively during last few years, by means of long term molecular dynamics simulations [138, 139]. These simulations were performed on binary mixtures of Cholesterol-phospholipids and also on ternary mixtures containing the steroid, phospholipids with a low transition temperature Tm, and phospholipids with a high main transition temperature Tm. In this way, precise information

about Chol-lipid interactions are provided and the nature of lipid rafts has been pointed out in model membranes [138].

The hypotesis of membrane sorting induced by cholesterol-sphingomyelin interactions has changed the traditional viewing of lipid membranes, and an enhanced bioactive role for specific lipids has been finally recognized [133]. However, in the current prevailing representation, rafts are small, dynamic domains in membranes enriched in sphingomyelin or other saturated lipids, Cholesterol, and specific membrane proteins [134].

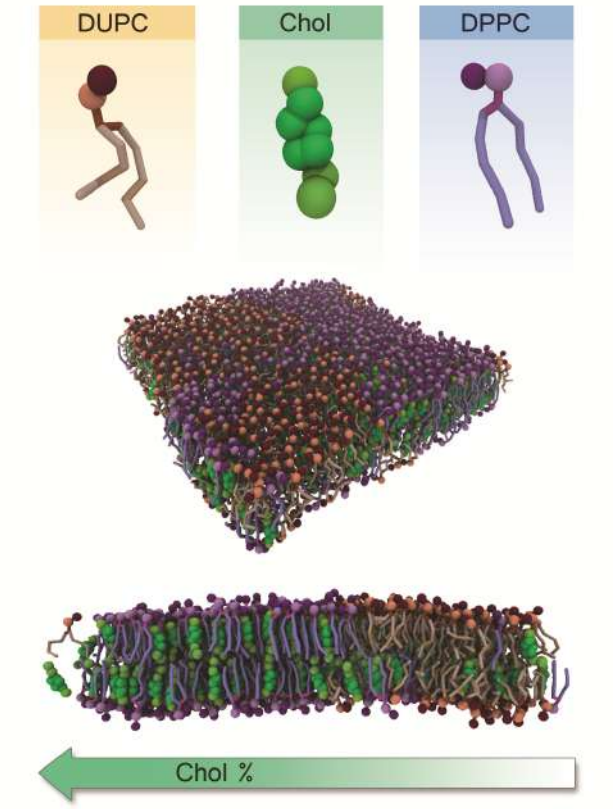


Figure 2*: Fluid Mixed Lipid Bilayer Domain Formation.* On top, the Coarse Grained model for DUPC, DPPC and Chol. Fully saturated (DPPC) and doubly unsaturated (DUPC) lipids, tend to segregate in large domains when mixed in the same bilayer. At the bottom we can see the preferential packing of Cholesterol within the fully saturated fatty acid chains of DPPC. *Systems coordinates pdb Files have been taken unmodified from http://md.chem.rug.nl/cgmartini/index.php/example-applications2/lipid-membranes Authors: Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH.*

The above mentioned compounds show an increased tendency to pack more tightly than in the surrounding membrane. This behavior gives rise to a liquid-ordered region(Lo), which coexist with the surrounding *non-raft* liquid-disordered regions(Ld). The origin of the tighter packing is not completely clear: it has been ascribed or to strong hydrogen bonding between sphingolipids and Cholesterol, or also to stronger van der Waals' interactions occurring between Chol and SM with respect those founds for other classes of lipids such as DPPC [135]; such packing can be described with qualitative and quantitative accuracy by means of molecular dynamics only if polarization effects are included into the simulations [139].

Beside these intermolecular interactions, another important fact to be considered for the lipid raft hypothesis is the cholesterol's localization. It is well documented in literature that phospholipids have an asymmetric distribution: phosphatidylethanolamine (PE) and phosphatidylserine (PS) are concentrated on the inner leaflet while the external leaflet is enriched in phosphatidyl-choline (PC) and sphingomyelin [140].

The understanding of domain coupling across these asymmetric membrane leaflets since it is supposed to be critical for the existence of lipid rafts and one of the most discussed topic in this research area is represented by cholesterol's distribution across the asymmetric plasma membrane. In fact, the knowledge of the exchange velocity of Chol between leaflets is an important aspect to clarify for assessing its effect on domain formation. Among the proposed models, Garg *et al.* proposed a slow cholesterol *flip-flop* [141], which is associated to an active transport of the steroid, which could easily maintain a non-equilibrium distribution across the bilayers leaflets. These *flip-flop* motions have been studied on the basis of atomistic and coarse-grained methods: they start on a poly-unsaturated DAPC bilayer and at the end describe a rigid ordered DPPC bilayer containing a 40 mol% concentration of cholesterol together with a model '*raft*' bilayer with a composition of (1:1:1 ratio for Chol: PSM: POPC) [142-144].

Beside, other studies were oriented towards the full understanding of the chemical-physical properties for both lipid domain structure and their functional roles. In particular, it was observed that cholesterol broadens the gel-liquid phase transition of the lipids by preventing packing at low temperatures, as reported for

instance for DPPC [145]. Furthermore, after its incorporation a clear *condensing effect* is induced into phospholipid membranes: the lipid tails become more ordered, the bilayer becomes thicker with a reduction of the area per lipid; as a consequence the bilayer becomes more rigid, thus preventing membrane deformation and bending [146-149].

Cholesterol Containing Membranes and Bilayers as Probes

Fluorescent steroids are an important class of membrane probes [150] used to identify the eventual localization of any non-homogeneous distribution of the bilayer's lipids, particularly during the formation of Cholesterol-rich domains [151].

Recently the phase diagrams of complex lipid mixtures gained a renewed interest, in particular three-component mixtures of cholesterol or other functionalized steroids with two phospholipids species have been investigated, well knowing that the development of these phase diagrams for such complex mixtures is notoriously demanding.

As an example, fluorescent cholesterol analogues which are commonly used in membrane biophysics, are Nitrobenzoxadiazole (NBD)-labeled steroids. In a recent study, it has been reported the design of a new fluorescent probe (NBD-R595) aimed to explore the phase behaviors of large unilamellar vesicles containing one, two-, and three-components [152]. The investigations have gone further and 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoCholine (POPC) bilayers, containing either Cholesterol or one of its two fluorescent analogues, 22-NBD-Cholesterol or 25-NBD-Cholesterol, have been carried out by means of atomistic molecular dynamics [153]. From the data collected it has been identified peculiar adopted conformations of these probes in which their tail-labeled fluorophore is directly oriented toward the lipid/water interface; this location is quite similar to that observed in previous molecular dynamics simulations of other NBD probes. However, from a detailed analysis it is clear that nor 22- or 25-NBD-Chol are unable to mimic the most important features of cholesterol's behavior in lipid bilayers. Beside, the similarity between the behaviors of other compounds belonging to this class of fluorescent probes and Cholesterol itself have been underlined in a recent MD [154].

Cholesterol as Stabilizer in Liposome Based Vaccines

The properties of Cholesterol as an artificial membrane stabilizer have been extensively studied in a wide range of phospholipid liposome systems [155, 156]. In a recent study taken from literature, Kaur *et al.* shed a light into a possible role of t steroids in cationic liposomal adjuvants containing both trehalose 6,6′-dibehenate (TDB) and dimethyl dioctadecyl ammonium (DDA), already used as strong adjuvant system for vaccines against a wide range of diseases [157].

The steroid packaging within TDB:DDA liposomes was investigated and it have been observed that the incorporation of Chol into liposomal membranes induces the formation of a liquid-condensed monolayer; furthermore it modifies the main phase transition temperature of the system, with an increase of the bilayer fluidity and a reduced antigen retention *in vitro*. Beside, *in vivo* distribution studies have been carried out and it resulted that this increased membrane fluidity did not interfere with liposomes and antigen deposition at the site of injection. Concerning the immune responses, some differences were pointed out: in fact, early IgG responses were reduced by the steroid inclusion but there were no differences in antibody (IgG, IgG1, IgG2b) responses promoted by TDB: DDA liposomes both in presence and in absence of the steroid. Moreover, these liposomes induced significantly higher levels of IFN-gamma, and also the liposome uptake by macrophages *in vitro* was higher for pure TDB:DDA liposomes compared to their Chol-containing counterparts. All these observations suggests both cellular interactions and immune responses can be altered by small changes in bilayers mechanics [157]; thus, these physicochemical properties must be balanced to gain the desired immunological outcome.

Cholesterol and Neurological and Immunological Diseases

Recent studies were directed towards the explanation the Cholesterol role in cell membrane in relation to the amyloid toxicity, associated with Alzheimer's disease [158]. In fact, structure and physical properties of the membrane is strongly affected by its chemical composition and by the incorporation of small molecules, such as melatonin and cholesterol; thereby its interaction with amyloid peptides can be affected. Recently, both compounds have been put into relation with amyloid toxicity and the melatonin showed a protective role. Instead, the

cholesterol's role remains controversial since the mechanism of this protection is still not well understood at molecular level. However, many studies have shown clearly that the lipid membrane is critical in enabling amyloid fibril formation and its ensuing toxicity [159, 160].

In their research work, Drolle *et al.* [158] used small-angle neutron diffraction (SAND) from oriented lipid multi-layers, small-angle neutron scattering (SANS) from unilamellar vesicles experiments in conjuction with Molecular Dynamics simulations to evaluate the interactions of melatonin and cholesterol with 1,2-dioleoyl-sn-glycero-3-phosphoCholine (DOPC) and 1,2-dipalmitoyl-sn-glycero-3-phosphoCholine (DPPC) model membranes. It resulted that melatonin decreases the thickness of both model membranes inducing a disorder in the lipid hydrocarbon chains; this behavior leads to an increase of membrane fluidity, showing an opposite effect to the well-known condensation induced by cholesterol. On the basis of these results, melatonin is supposed to counteract cholesterol's membrane ordering effect, thus affecting amyloid binding [161], and leading to membrane damage.

Beside, other studies reported that Chol can hinder the insertion of amyloid peptides into lipid membranes but it depends strongly on the type of the bilayers lipid and involved amyloid peptide [162]. Moreover, recently it has been observed that a non-homogeneous binding of the amyloid peptide to the lipid membrane is induce by cholesterol, and it can result in the formation of defects in membrane [163, 164]. At the cellular level, the steroid proposed effect is to reduce the toxic effect of amyloid plaques on neuroblastoma cells [165]. Beside, to gain a deeper understanding of the processes involved, phospholipid membranes with melatonin have been modeled using full atoms molecular dynamics [166-168]. As an example, the simulated system is composed of 512 lipid molecules, approximately 28,000 water molecules corresponding to full hydration, together with 72 molecules of melatonin (12 mol% overall concentration) [166]. The GROMOS 53a6 force-field inside GROMACS software package has been used in the simulations, since it has been tested and validated *ad hoc* for lipid and peptide systems [167, 168].The corrections by Bachar *et al.* were included for acyl double bonds for proper modeling of sterol-lipid interactions [169]. The protocol used

has been already found to be reliable for membranes and other related systems [167].

Other computational studies [170] in the immunological field were aimed to establish the Cholesterol role with the activity of some HIV Peptide fusion inhibitors (FI). In particular, T-20, also known as Enfuvirtide or Fuzeon, [171] and T-1249 [172] have been shown to interfere with human immunodeficiency virus fusion of the virus envelope with the immune system cell. The inhibition of such a fusion process takes place by binding to proteins responsible for recognition and fusion, such as the gp41 protein [173-175], protein responsible for the viral pore formation and membrane fusion [176, 177]. More in details, T-20 is a synthetic peptide containing 36 amino acids, whose sequence is homologous to the *C*-terminal of HR2 (Heptad Repeat 2) of gp41 [178]. In particular, the activity of T-20 is strictly linked to 10 residues from the membrane proximal external domain (MPER), also known as tryptophan-rich domain (TRD), which are also involved in the peptide binding to lipids [179]. From the *in silico* results, T-20 and T-1249 fusion inhibitor (FI) peptides were shown to interact with 1-palmitoyl-2-oleyl-phosphatidylCholine (POPC) (liquid disordered ld model) and POPC/Cholesterol (1:1) (POPC/Chol) bilayers (liquid ordered lo model), and the two peptides establish this interaction with different extents [170]. This different behavior can be ascribed mainly to a pocket binding domain (PBD), which is lacking in T-20 and present in T-1249. In fact, it has been suggested that the PBD domain enhances the interaction of FI peptides with HIV gp41 protein and with model membranes [180]. This FI Interaction with both the cell membrane and the viral envelope membrane is critical for the fusion process inhibition. Thus, molecular dynamics were carried out with the aim of assessing the consequences and the differences in the interaction of these FI peptides with ld and lo model membranes. As a result, T-20 and T-1249 induced different effects on bilayer structural and dynamical parameters [180, 181] and a stronger influence in the measured parameters is determined by T-1249's adsorption to the membrane surface. The presence of both binding domains in T-1249 are supposed to be critical to explain its stronger interaction, and they can be responsible of membrane properties changes upon peptide adsorption. Beside, 100 ns molecular dynamics simulations of both solvated bilayers (POPC-in ld and POPC/Chol 1:1-

in lo) were performed, and several parameters where determined, both for validation purposes and comparison with the corresponding peptide-bilayers simulations. The results show that the two bilayers behave differently upon peptide adsorption, and undergo sizeable structural modifications. Furthermore, the computational MD analysis shows an atomistic level of the modulation through H bond formation which involves the steroid in the POPC/Chol lo system; this further confirm the hypothesis that to gain an enhanced effectiveness, both in the binding to gp41 and membranes, the novel designed fusion inhibitor must hold both TRD and PBD [180, 182, 183]. Moreover, a stable adsorption to bilayers in both phases is required in order allow for higher concentration of the HIV fusion inhibitors in proximity of both the cell membrane and the viral envelop, thus resulting in a more effective action [184].

Mixed Composition Lipid Matrix Systems in Nanomedicine

Many efforts have been made by scientists to develop new safe and efficient supramolecular vectors for transporting a pharmaceutical compound in the body to achieve the desired therapeutic effect (Drug Delivery). In fact, if a drug is not able to reach its site of action, it is essentially useless; moreover, its delivery is affected by its physico-chemical properties, and also by the interplay of these factors with binding, transport, and metabolism of the drug in the body. Thus, the choice of the delivery methods to efficiently transport *hard-to-deliver* compounds to the appropriate target sites, is of great importance. Beside Drug Delivery, another major challenge in pharmaceutical science is represented by the transport of gene-based drugs (DNA, oligonucleotides, siRNA) (Gene Delivery). The recent progress in the field of Nanotechnology can help to improve the delivery in these difficult cases and thus, nanocarriers are widely investigated as potential solution. Among nanocarries for DNA delivery, liposomes of mixed composition and with different nature (cationic, anionic, neutral) have been the most studies ones which are still gaining much attention from the scientific world due to their efficiency and manageability [185, 186]. In these field of research, molecular systems containing charged lipids, such as cationic bilayers, have been studied throught a great number of molecular dynamics (MD) simulations, employing all-atom or coarse-grained molecular models (better described in next chapter session). These mixed lipid systems have a composition in line with the cell lipid

composition and thus they represent a good choice to design safe and efficient nanovectors for Drug or Gene Delivery. The use of *in silico* models allows the lipid to target drug design allowing the saving of time and money, features that make the *in vitro* experiments very tiring. Computational models also allow a more detailed study of the interactions, which allows the development of vehicles for drugs even better directed towards their biological target [187]. Some recent examples of studies in this field are here reported [188, 189]. The first concerns a nonphysiological, cationic lipid, diC(14)-amidine (amidine), which is considered a promising vector for DNA, proteins, and also drugs due to its capability to arrange in stable liposomes under physiological pH. Indeed, the design of new cationic lipids is particularly worth of importance since their liposomes have been supposed to be mediators for the cellular uptake of nucleic acids or proteins [188]. This new cationic lipid gain both anti-inflammatory and inflammatory properties that makes it a particularly interesting target for possible therapeutic applications.

Thus, to gain inside into its mechanism of internalization, the structure and dynamics of a model amidine bilayer together with mixed amidine: DMPC bilayers, atomistic molecular dynamics simulations have been performed within the GROMACS package, using OPLS all-atom force field together with Berger united-atom parameters for the lipid tails, and the Ryckaert-Bellemans torsion angles potential. It was observed from the computational data collected a considerable fluidity in the hydrophobic bilayer core together with the tendency for strong surface curvature, in agreement with the relatively small size of experimentally formed liposomes. From the cluster analysis of the trajectories frames a thermally accessible population of V-shaped lipids, indicative of fusion capabilities with the plasma membrane was showed. This hypothesis was later confirmed by fusion experiments and a comparison with the data obtained for the zwitterionic dimyristoylphospho-choline (DMPC), which also carries two saturated C(14) tails, was done [188]. Beside, several atomistic MD simulations have been performed to explain the polyethylene-imine (PEI) mediated DNA aggregation [189]. The aim was the understanding of the molecular mechanism of this process for an optimal design of new gene delivery carriers. All the simulations have been conducted using CHARMM General Force Field for PEI, and CHARMM 27 force field for all the other molecules. The FF parametes used

for PEI have been carefully validated through high level *quantum-mechanical* calculations, since the outcome of MD results are sensible to the chosen torsional parameters. The NAMD package, TIP3P water model under periodic boundary condition, and Particle-Mesh Ewald electrostatics method were used for all the simulations. From the computational results, it has been shown that PEIs can contribute to DNA aggregation involving two different mechanisms: (i) it can forms poly-ions bridges between DNA segments and/or (ii) it can screen the negative DNA charges at a short distance from the surface of DNA molecules. According to the latter mechanism, the PEI/DNA charge ratio needs to be above certain value in order to maintain a stable aggregation. Indeed, it has been proved that PEIs are more capable of neutralizing the DNAs at close distance compared with monovalent ions; moreover, they provide a full neutralization at about 12 Å from the DNA-C1′ atoms, when the PEI/DNA charge ratio is above 1. Furthermore, in these DNA/PEI aggregates the DNAs spacing leis always between 21.4 and 29.0 Å and the presence of an excess of PEIs favours the binding to the already positively charged aggregate, since it increase its overall charge. The molecules in excess may also replace the PEIs previously bound to the DNAs in the aggregate, showing a dynamical evolution of the process; however, this excess do not change the spacing of the DNAs in the aggregates. Overall, these interesting *in silico* studies contributes to give insights into polycation-mediated DNA aggregation and condensation [189].

The same group went further in their work considering the modification of polyethylene-imine with lipid substitution with the aim to improve the efficiency of polycationic gene carriers [190]. The elucidation of the role of the lipid substitution in DNA binding and aggregation was the major goal, since it still unclear and it needs to be probed at the molecular level. Thus, they carried out a series of atomistic MD simulations on DNA aggregation considering a possible lipid modified polyethylene-imine (lmPEI) mediation. Indeed, it can be observed a significant association among lipid tails of different lmPEIs, which can promote DNA aggregation and stabilize the formed lmPEI/DNA polyplex. In additon, the lipid tails on the lmPEIs are located at the periphery of the lmPEI/DNA polyplex and thus contributing favorably to the interaction of the polyplex with cell membrane.

The same work was carried out considering siRNA aggregation and it was observed that for PEIs modified with short lipids, the increase of the lipid substitution level results in more compact and stable siRNA structure [190]. On the contrary, functionalizing PEIs with long lipids, the amount of PEI linkage *via* lipid association does not change; instead, it can be observed a reverse effect on compacting siRNA structure due to increased steric hindrance during the lipid association on individual PEIs. The high correlation between *in silico* results and the experimental data underline the importance of the computational approach in designing and evaluating the strength of polycation-based siRNA carriers [189, 190].

Another important application in nanomedicine is related to the understanding of membrane-mediated pharmacological effects of drugs. In fact, many drugs of both synthetic and natural origin, have a wide range of useful pharmacological properties that cannot be explained with the exclusive binding to the protein target. Biophysical studies of the interactions of these bioactive compounds can shed some light onto these drugs' effects when used in combination with Molecular Dynamics simulations, since they can access time and length scales not accessible experimentally and furthermore they can obtain a quantitative thermodynamic description of the molecular interactions [191]. Thus, a complete biophysical description of possible routes of their membrane-mediated pharmacological effects can be provided carrying out simulations of small drugs with model membranes.

SIMULATION OF PROTEINS AND PEPTIDES IN MEMBRANE BILAYERS

In Silico Dynamic Simulation of Membrane Proteins

Proteins play a major role on Membrane Transport, which represents a fundamental process for the living cells. They can act as channels, providing highly selective diffusive pathways gated by environmental factors, but they can also act as transporters, assisting the movement of substances inside or outside the cells by diffusion or active transport. Almost 25 % of proteins in eukaryotic genomes are Membrane proteins, and they also constitute more than 50% of current drug targets. Most of them are formed by bundles of hydrophobic transmembrane (TM) α-

helices. Being the lipid bilayer environment a complex molecular system, a detailed description of protein-membrane interactions using only experimental techniques is very difficult. For this reason they are good targets for computer simulations, which in this field of research are growing up in rapid progress due to both the availability of high resolution protein structures and advances in atomistic simulation algorithms [192]. As major goal, atomistic MD simulations can actually reveal the molecular mechanisms at the basis of channel and transporter function, extending simulations to physiologically relevant time scales such as several micros. Furthermore, methodological improvements in coarse-grained (CG) models has potentially lead to multi-microsecond dynamics of extremely large systems (see next section for example case studies).

However, the approach in this field of research needs to put a particular attention in the quality and accuracy of simulations which are related to several factors.

Firstly, the treatment of the surrounding lipid bilayer can affect deeply the dynamics of the simulated macromolecule. The lack of inclusion of polarizability and low precision force field parameters, can result in erroneous description of geometrical and dynamical properties of the simulated lipid structures. As an example, the area per lipid for membranes can be underestimated and too much order in the lipid tail region can be observed. Overall, inaccuracies may affect the dynamics of the embedded protein, and large scale protein motions, such as those happening during the transport cycle of membrane transporters or gating of channels can be hindered. The use of polarizable force fields, already discussed in a previous section of this chapter, can partially solve these problems. More difficulties still remains in the accurate prevision of all membrane properties affecting the protein embedding, such as the dipole potential, even if a first successful test have been carried out for a DPPC monolayer [193]. Other key attributes in membrane channels that might be affected by force field approximations are ion conduction and selectivity. In fact, electronic polarization effects, such as those induced by a permeating ion in the channel protein, are not negligible, and, therefore, an exhaustive description of such processes can be achieved only if these effects are included in the simulation.

Another important aspect to be considered is the lipid composition. In fact, biological membranes are generally composed of various types of lipids, and even exhibit asymmetry with respect to the lipid composition of their two leaflets. However, due to the short duration and finite size of the membrane used in simulations, a homogeneous single lipid bilayers composition is used in most *in silico* studies. This problem of lipid composition is less of concern, despite the fact that the majority of the simulated membrane proteins conserves its main function after inclusion in model bilayers. However, in some cases, the biological function can be altered by the physical-chemical nature of lipids: for example the presence of anionic lipids affects the binding and activation of coagulation factors, and signaling proteins. This problem can be overcome including a closer representation of the protein natural environment in the simulation.

Indeed, during last years the greatest part of simulations concerning membrane protein focused on ion channels [194-201], on G-protein coupled receptors (GPCRs) [202, 203], on protein-conducting channels [204], on Receptor tyrosine kinases [205-208], as well as on various transporters [209, 210]. In this section we discuss some example studies in this field of research.

Most studied for their physiological importance are Membrane channels, responsible of selectivity, and sensitivity for a cellular optimal conductivity. During conduction events, these proteins are usually involved in small motions, easily reproduced over time scales shorter than microseconds and thus readily studied by MD methods.

A system of considerable biophysical interest is the Voltage Sensor Domain (VSD) of potassium (Kv) channels, which has been extensively studied as Kv ion channel tetramers as well as isolated voltage sensors by means of atomistic simulations [194, 212]. These membrane proteins are highly selective for the cellular conduction of K^+ ions and they are responsible for maintaining the membrane resting potential. These channels open in response to a variety of *stimuli*, such as binding of ligands or changes in the transmembrane potential and intracellular pH. The conduction pore is regulated by the selectivity filter of the channel and four binding sites for K^+ ions are provided. Conformational changes in the selectivity filter have been generally considered as responsible for the K^+ channels inactivation.

In order to investigate this behavior [194], the KcsA potassium channel in its inactivated state has been characterized using a combination of experiment and simulation techniques. The *in silico* results showed that the conductive conformation of the filter is intrinsically unstable and can assume different conformations. On this basis, it has been suggested that during the opening of the channel gate, the selectivity filter fluctuates between alternative conformations, thus assuming the role of a secondary (fast) gate for the conduction of the K^+ ions. Other non-conductive conformations of the filter have also been identified in the simulations of KirBac1.1 channel (Fig. **3**).

More recent computational studies have been conducted on KcsA channels and the research have gone forward with the aim to better clarify the basis of ions selectivity [198-200]. In fact, the high number of theories on the origin of the K^+ channel selectivity depends on the intrinsic difficulty in relating these simulations to the experimentally observed activity. One of the most direct measures of K^+ selectivity is provided by blockade experiments in the presence of Ba^{2+} ions. Recently, Piasta *et al.* carried out Ba^{2+} blockade experiments on the KcsA channel [213], which have been particularly useful due to the availability of a high-resolution X-ray structure of this channel. Later on, Roux *et al.* [200] reported the results of umbrella sampling and free energy perturbation (FEP) simulations aimed to provide an explanation of the previous conducted Ba^{2+} blockade experiments [211]. The attention was focused on some of the TM domains of voltage-gated potassium channels which are known as voltage-sensor domains (VSDs). In these domains, there are highly charged segments, which are formed by multiple positively charged residues, known as gating charges. These aminoacids are positioned uniformly at every fourth position on one of the TM helices. Changes in the electrostatic potential induce the VSDs movements, thus coupling the transmembrane voltage to the conformational change of the conduction gate. From results collected, it appears that the charged residues, mainly represented by arginines, are fully solvated inside the membrane and they are further stabilized either by interaction with the lipid headgroups or by salt bridge interactions involving other protein's charged residues [195]. In addition, the membrane electric field is concentrated within the VSDs, thus resulting in a sharp voltage-gradient across the membrane. Finally, in 2012, Shaw *et al*. [197]

reached the final goal in proposing a general mechanistic model for the Kv voltage-gated ion channel superfamily in order to explain experimental data apparently in conflict. In their study, they used atomistic MD simulations, and obtained a representation of how voltage-gated potassium channel (Kv) switches between activated and deactivated states [197].

Beside all these studies, once again by means of full atom MD, another research group investigated the structural and dynamical properties of the voltage-gated potassium channel Kv1.2 embedded in bilayers with modified upper or lower leaflet compositions corresponding to more realistic biological environments. Thus, they considered the effects of sphingomyelinase, an enzyme that modifies the composition of lipids of the outer membrane leaflets, and secondly the effect of the presence of a small fraction of a highly negatively charged lipid, phosphatidylinositol 4,5-bisphosphate (PIP2), which is known to modulate voltage-gated channel function. The molecular dynamics simulations did not exclude the global effect mechanism in the first case, instead for the latter, however, local interactions between the ion channel and the lipid head-groups were shown to be key elements of the modulation [201, 211].

Other groups proposed hypothesis for ion gating mechanisms on the basis of simulations' results of the mechanosensitive channels MscS and MscL. In fact, it is not clear how these proteins change conformation in response to different surface tensions in the membrane [214, 215]. The mechanosensitive channels are proteins found in prokaryotic and eukaryotic cell membranes, which open a conductance pore in response to mechanical stress. An important role is played by the bacterial large conductance MS channel (MscL) which regulates turgor pressure around the cell. Since permanent leak in the membrane could lead to the bacterium death, a complete understanding of the gating mechanism of bacterial MscL could improve the development of novel antimicrobial agents. Following these purposes, the biophysical mechanisms of the gating of the bacterial MS channel have been fully investigated starting from its 3D protein structure in the closed state and using a large amount of electrophysiological data on its gating kinetics [216]. In particular, the initial process of MscL opening in response to an increasing tension in the lipid bilayers have been focused by atomistic molecular dynamics (MD) simulations. As a result, the interaction energy between

membrane lipids and facing amino acids have also been calculated in order to identify the sites sensible to tension in the channel protein.

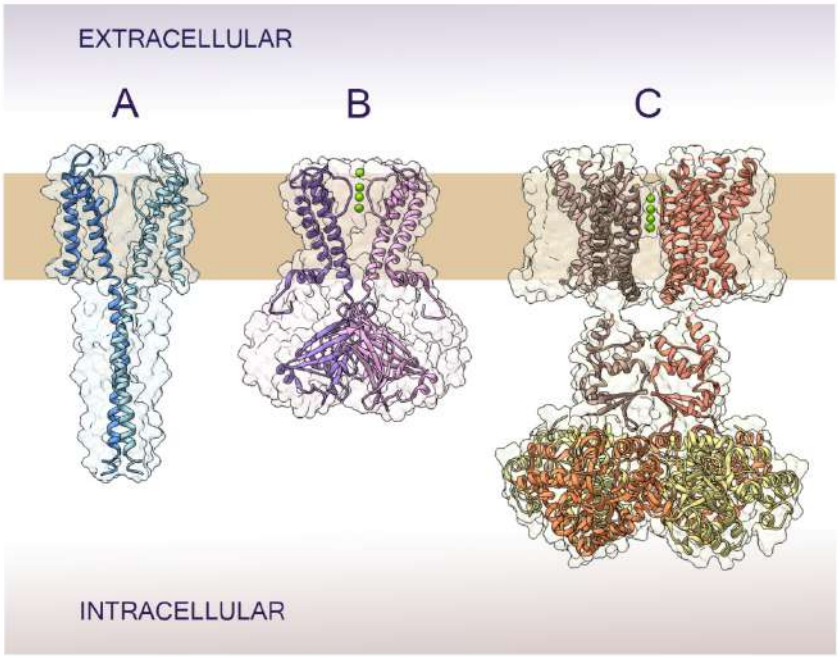


Figure 3: Crystal structure of different potassium channel proteins orientation inside through the cell membrane. The latter is shown as a simplified stripe for clarity. A) Kcsa channel in its closed conformation (PDB code 3EFF], no K+ ions where crystallized inside the channel domain; B) Kirbac1.1 channel (PDB code 1P7B) with K+ ions show in green; C) voltage-dependent potassium channel with K+ ions show in green.

An aspect worth of notice is that spontaneous motions related to function have begun to be reproduced by simulations. Some relevant examples are the concerted global twisting motion and helix tilting proposed for nAChR activation [217], or initial conformational changes responsible for the transport mechanisms of the lactose permease (LacY) protein [218]. At present, however, simulations are able to efficiently reproduce the position of a membrane protein in a bilayer, and also its molecular interactions with the surrounding lipids.

The identification of amino acidic residues involved in such interactions are relevant to gain information about residue distributions within membranes [219,

220]. In particular, solvation of the involved amino acid side chains as a function of their position along the bilayers normal [221] and also partition free energies [222, 223] have been calculated. In a more recent research study, the mean force Potentials (PMFs) between ionizable amino acid side chains (Arg, Lys, His, Glu) among the headgroup area of a palmitoyl oleoyl phosphatidylcholine (POPC) lipid bilayer have been reported. The data were obtained from both atomistic MD and adaptive biasing force method in the context of metadynamics [224]. As a result, it has been identified the attractive interactions between charged amino acids which are mostly stronger in the lipid interface; the strongest interaction involves a pair of Arginine in stacked orientations. However, the involved free energies of interaction for ionizable side chains cannot be related to simple dielectric factors; furthermore the collected results shed some light in membrane protein stability and its oligomerization and this is particularly important since the principles of both stability and association are currently poorly understood.

Indeed, if it is necessary to explore more specific interactions of lipids with membrane proteins, an increasing duration of MD simulations is needed. Thus, it has been observed that selective interactions with anionic lipids are present for the potassium channel KcsA [225], well correlating with the available experimental data.

Another case study refers to the identification of protein specific interactions with lipid tails, as can be seen between polyunsaturated fatty acids and rhodopsin [226]. This study was achieved as a part of [26 ns-100 ns] set of simulation runs. However, the need for very long-large-scale atomistic simulations to reach convergence in lipid and protein dynamics has been stressed by statistical analysis [227]. For this purpose, there is an increasing interest in exploiting the longest accessible timescales through the association of atomistic MD with a more extensive CG simulation; only this combination can provide sufficient sampling for more quantitative exploration of protein-lipid interactions.

Among Receptor tyrosine kinases, the study of the activation mechanism of the intracellular kinase domains for the epidermal growth factor receptor (EGFR) is particularly worth of note [205-208]. As it is well known, Receptor tyrosine kinases are key regulators of normal metabolism of the cells [228], but they have

also a critical role both in the development and progression of many types of cancer [229-231]. The trigger for EGFR signaling is an extracellular ligand binding, which promotes EGFR first dimerization then activation. This dimerization-driven activation of the EGFR intracellular kinase domains is fundamental for cellular pathways regulating differentiation and proliferation.

However, inactive state of EGFR can be both in monomeric and dimeric form; this fact suggests that the EGFR activity is regulated by a more complicated mechanism. For sure, a key role is played by the surrounding membrane but it is difficult to study it by experimental structural studies.

Thus, in order to elucidate the molecular basis of the EGFR activation mechanism, it has been carried out extensive molecular dynamics simulations on EGFR embedded in membrane [205]. From the data collected, it has been suggested that in ligand-bound dimeric form, the extracellular domains can assume conformations favoring the association of the TM helices near their N-termini, and determines formation of asymmetric kinase dimers (*active form*). Instead, in ligand-free dimers, the N-termini of the TM helices is holded apart, and in this conformation, the extracellular domains favor instead C-terminal association, thus inducing the formation of symmetric kinase dimers (*inactive form*). Furthermore, electrostatic interactions, which are established between EGFR's intracellular domains and the membrane, are critical in maintaining this coupling.

Other studies conducted by the same group pointed out other aspects of EGFR activation and inhibition [206-208]. In detailed, atomic-level computer simulations, however, they found out that the Receptor lies down on the membrane, placing its ligand (EGF)-binding site adjacent to the membrane surface. They also showed that the epidermal growth factor (EGF) may interact with its receptor EGFR in two distinct ways: with or without the involvement of the membrane. This may explain the experimental finding that an EGF molecule binds to EGFR more weakly at higher EGF concentration. This phenomenon, which is a manifestation of an underlying negative cooperativity, is an important but poorly understood characteristic of EGFR activity. Thus, further long term Molecular dynamics (MD) simulations investigated the structural basis of the negative cooperativity in the ligand binding of human EGFR. The long term simulations (tens of μs) on such huge

number of atoms (62,000-315,000 total atoms range) were performed on a supercomputer developed specifically *i.e.* Anton, [232], using the Amber ff99SB-ILDN [233] parameters set, together with the ff99SB* backbone correction [234] for proteins, the CHARMM C36 FF [66] for lipids, and TIP3P [85] for the water model. In particular, it has been suggested that interaction of receptor bound ligands with membrane gives a negative binding cooperativity of the EGF Receptor. This is particularly important since this cooperativity is widely believed to be central to the effects of ligand concentration on EGFR mediated intracellular signaling and thus implicated in cancer proliferation.

Proteins and Peptides Bound to Bilayer Surface

In addition to the research fields described in the previous section, the location and interactions of a peripheral membrane protein is more complex to predict than for integral membrane proteins.

Replica Exchange Molecular Dynamics (REMD) are versatile computational techniques particularly suitable to simulate how some peptides partition spontaneously into the bilayer, and then fold to helices either during or after insertion. Many application studies of this powerful conformational sampling techniques have been carried out [235-239].

The typical membrane bilayers, in which a protein or a peptide must be in, is in general not a pure lipid bilayer, but may contain steroids, such as cholesterol, and two or three phospholipids. Computing resources are certainly at a stage where researchers can study more realistic membranes containing five or more phospholipids. However, the complexity of building such realistic and heterogeneous membranes makes this a considerable challenge. To simplify the building process of protein-membrane systems for MD simulations, the graphical user interface (GUI), MemBuilder [79] or the CHARMM-GUI website [81] can be successfully used. As already stated, it is well-known that more than one kind of lipid molecule forms a lipid membrane, and the organization and composition of lipid molecules are very important for a membrane's biological functions, which range from membrane trafficking to signal transduction. Moreover, it was reported that for some proteins, the heterogeneity of a membrane is particularly important for protein function: for this reason it must be used the extension

of MemBuilder GUI to incorporate the generation option of heterogeneous lipid bilayers. The lipids available for bilayer generation are numerous, and the interface allows a user to select more than one kind of lipid to build membranes of different compositions. As an application example of this methodological approach, we can report here a recent work which presents the benefits of atomistic Molecular Dynamics (MD) to study the dynamical properties of biological membranes and small peptides such as the Myelin Basic Protein (MBP) C-terminal α-helical peptide, an important protein in the central nervous system. This protein is found in various isoforms with a predominant splice isoform of 18.5 kDa in an adult brain [240]. Recent studies have demonstrated that the severity of multiple sclerosis (MS) is correlated with post-translational modifications of MBP, such as citrullination [241]. Due to its central role, MBP is thought to be connected with myelin degradation when MS attacks the myelin-wrapped nerves of the central nervous system. Molecular Dynamics (MD) provides a nice, quick way to study the behavior and interaction patterns of MBP with lipid membranes that could provide insights into molecular details of myelin structure, and pathogenic mechanisms in MS. This study is particularly important since, in addition to bare membrane simulations, the simulation of a C-terminal peptide in and mixed bilayers was done for the first time. Interesting trends of the two systems related to membrane composition and protein secondary structure preservation were observed over the course of the 200 ns simulation. The depth of penetration and α-helical structure stability were successfully measured. Overall, in this case, the electrostatic environment of the two membranes might partially dictate stability of the peptide amongst other factors (*i.e.*, localized pH, lipid-peptide thermodynamics, lipid density). These findings are significant in their implication that membrane composition affects the behavior of MBP, providing further insights into myelin structure. The obtained results suggest that local changes in membrane composition (*e.g.* enrichment in DMPE molecules), as well as, electrostatic nature of primary amino acid sequence could cause localized denaturation/instability of external MBP α-helices possibly augmenting the degradation of myelin in multiple sclerosis (MS), resulting in a subsequent decrease of nerve impulse propagation efficiency. Other studies worth of notice have been conducted on voltage sensitive pheripheal proteins aimed to point out the membrane composition influence in inducing response [242]. In particular, interactions of PTENs (Phosphatase and Tensin Homologue Proteins) with Phosphatidylinositol Phosphates

have investigated. The simulations performed have allowed to increase knowledge on the nature of the PTEN-membrane association process. The predicted PTEN final orientation in membrane resulted in good agreement with experimental and computational data collected previously [242 and references therein]. Moreover, as a result an increased concentration of PIP3 gave more frequent association of PTEN with the model bilayer, and PTEN could achieve easily its optimal orientation. Another interesting investigation have been conducted by McCammon *et al.* [243] on antimicrobial peptide CM15. They carried out 3-μs molecular dynamics simulation studying the peptide interactions with different composition membranes, in particular with two models: pure POPC and mixed POPG:POPC (1:2) bilayers. From the data collected, it was observed that CM15 has significantly reduced interactions with lipids in a pre-folded α-helix, compared with its random-coil conformation, thus suggesting that the peptide initial structures can affect the simulation results on the 100-ns timescale.

Finally, even if large membrane proteins cannot yet be folded *ex-novo i.e.* from their primary aminoacid sequence, structures of membrane proteins, obtained from both X-ray diffraction studies or by comparative molecular modeling, can be successfully refined by *in silico* computations. It is also possible to efficiently reproduce the insertion of small peptides, and even to evaluate the related free energy costs.

Technical Tips for Molecular Dynamics Simulations of Membrane Proteins

In this section we report briefly common simulation protocols used in molecular dynamics studies of membrane proteins and peptides; in particular we focus on a practical approach to set up and run simulations of such systems. Before embarking in such simulations, many questions and issues must be posed such as the current applicability of atomistic Molecular dynamics techniques, strongly related to the overall system dimensions, and the accuracy of the force field chosen for energetic optimization considering the existence of an accurate parametrization of reference molecules.

In their paper, Tielman and coworkers [244] introduce two new methodological approaches to embed a protein in a lipid bilayer. These methods have the great advantages to reduce equilibration time and to be almost completely automated.

Of course, before approaching any simulation, the choice of adequate force field in describing both protein-peptide and lipids must be done. As already discussed in previous sections [245], currently there are only four widely used FFs for simulating biological macromolecules: GROMOS [60], AMBER [61], CHARMM [63], and OPLS [246], which are subjected to continuous development and reproduce well many protein characteristics [246, 247]. The correct specification of the chosen FF requires the report of the exact revision number.

However, despite the fact that they have some different strengths and weaknesses, they share a number of the same limitations. This is particularly referred to electrostatics' treatment which is simplified accounting for electronic polarizability only in an average way. The methods improvements aimed to resolve these deficiencies are still ongoing [74, 245, 247-249].

An important aspect, which must be considered ongoing with dynamic simulations of molecules moving between two very different environments (lipids and water), is the parameters' choice for their description which have to be accurate in both environments. Fortunately, the free energies of transfer between water and hydrophobic environments can now be directly tested computationally [250-255]. Another important factor to consider in evaluating a cofactor or a ligand, is its partial charge distribution, or more generally, how the force field treats the non-bonded interactions. In fact, point derived charges and Leonard-Jones interactions give the proper behavior for a molecule in water, but they may not be adequate to describe behavior inside the bilayer or in a hydrophobic protein pocket. Overall, charges must be preferably derived from high level *ab initio* QM and DFT methods, and they should be totally compatible with the protein FF in order to not overestimate charges which may artificially and adversely affect the balance of forces controlling the protein-ligand interactions. Generally, it is advisable to use the same methodology used in parametrizing the protein and/or lipid FF.

Finally, the proper assembly of the starting lipid bilayer and protein system represents the major difficulty in beginning a membrane protein simulation [256-258]. There are slightly different possible approaches to this problem: (a) Inserting the protein into a pre-equilibrated bilayers [259, 260]; (b) building a

bilayer around the protein to be embedded; (c) Tieleman suggested methods (Fig. **4**) [244] and finally (d) GUI related methods [79, 81]. At the end of the section, we put some link to useful recent tutorial for most used software packages, which actually enable to build this kind of simulating systems. Finally, common molecular graphic softwares which can be of great help in that system setting are MolMol [261], VMD [86] or the GROMACS tool *editconf* [262, 263].

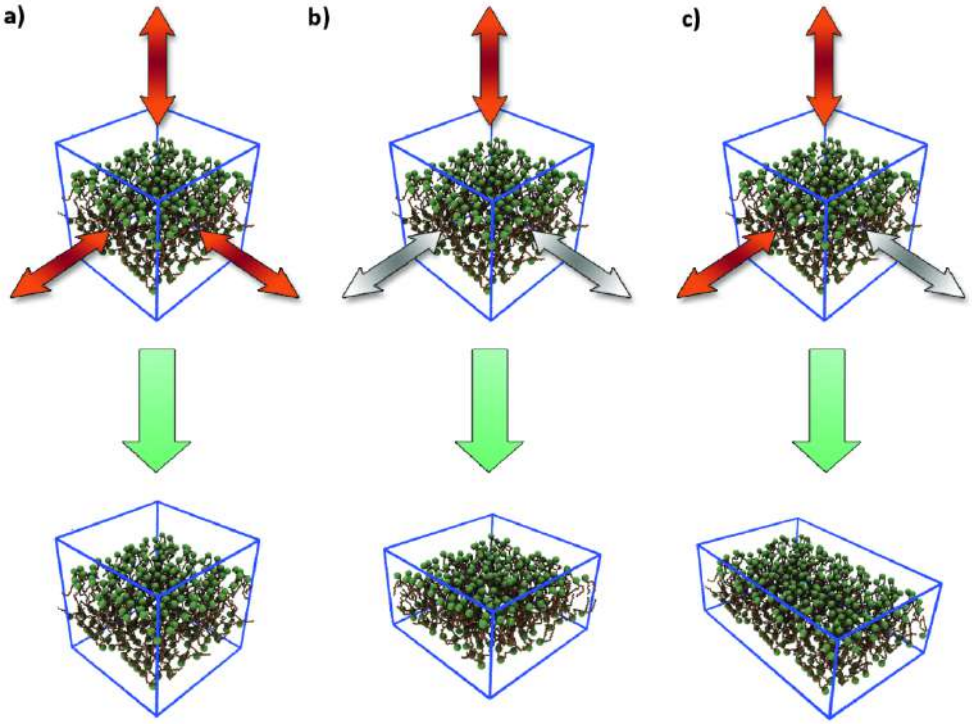


Figure 4: Pressure coupling types: (a) isotropic pressure (b) Semi-isotropic pressure (c) Anisotropic pressure

Tutorial Links

- ACEMD: http://multiscalelab.org/acemd/protocols/ACETK.BLDMEMBPROT
- NAMD:http://www.ks.uiuc.edu/Training/Tutorials/science/membrane/mem-tutorial.pdf

- CHARMM: http://www.charmm-gui.org/input/membrane
- GROMACS:http://www.gromacs.org/Documentation/Tutorials#Membrane_Simulations
- AMBER: http://ambermd.org/tutorials/advanced/tutorial16/

COARSE-GRAINED MODELS: USEFULNESS AND APPLICATION

Extending full atom simulation of very large biological system up to the micro or even millisecond timescale, is nowadays a challenge beyond our possibilities. The main difficulty is due to computer resources, in terms of both speeding up the calculation and analyzing such huge trajectories generated during the simulation. However, whether we can simulate or not biological events occurring over timescales of thousands of nanoseconds, it's only a matter of how much approximation the system can undergo without introducing virtual artifacts, thus obtaining good agreement with experimental values. Reducing the number of particles, and consequently the "resolution", without influencing the chemical-physical properties associated to the molecular aggregates, means acting on the granularity of the system thus approximating a fine-grained model (full atom) to a coarse-grained one (CG) [264]. This way, multiscale modeling addresses challenges related to spatial and temporal scale, opening the way to the mesoscale level. Thus, in the CG approach, the system is represented by a reduced number of degrees of freedom with respect to an atomistic treatment. Due to this reduction and elimination of fine interaction details, the simulation of a coarse-grained (CG) system goes faster with respect to the corresponding atomistic simulation, and it requires less resources. As a result, orders of magnitude in the simulated time and length scales can be considerably increased.

Two CG approaches have been recently developed to address various time-scales in biomolecular simulations, namely the *Residue-based CG* [265] and the *Shape-based CG* [266].

Concerning the first, based on the Marrink's *et al.,* [267] coarse-grained (CG) model for lipids, it has been proposed a protein-lipid CG model. Clusters of ~10 atoms (including hydrogens) are substituted by a single CG bead: four water

molecules become one "water" bead, beside an ion with its hydration shell becomes an "ion" bead, chemical building blocks of lipids are reduced to single CG beads, and each amino-acid is represented by two CG beads - one for the backbone and one for the side-chain; glycine is the only exception, represented by a single backbone CG bead. After building the CG structure, one needs to define rules determining its dynamics. Following a common approach in molecular modeling, we assume that CG beads are point-like masses obeying to Newtonian mechanics and interacting through effective potentials. Thus, bonded beads are connected by harmonic springs, and harmonic angular potentials help to maintain shape of molecular chains. Long-range interactions are represented by the Lennard-Jones 6-12 potential, and by the Coulomb potential. The bond lengths and angles for the CG model are usually derived from the averaging of corresponding distances and angles over representative all-atom structures.

Beside, the *Shape-based coarse-graining* is designed to model large-scale motions of macromolecular assemblies, representing proteins and other biomolecules with the fewest point-like particles as possible. Often a single protein is formed by compact domains, disordered ones, together with elongated linkers or tails; this CG approach method enables one to model both compact domains and tails with equal accuracy, taking advantage of an efficient topology conserving algorithm originally developed for neural computations [268]. The placement of coarse-grain (CG) beads is performed using a self-organizing neural network, which adapts to the shape of the molecule to be represented [269]. Once CG beads are placed, they inherit the mass and charge of the groups of atoms they represent. Neighboring beads are connected by harmonic springs, while separate molecules interact through non-bonded forces which are once again described by Coulomb and Lennard-Jones potentials. These nb-interactions are parameterized on the basis of atomistic simulations and available experimental data. The solvent is modeled implicitly, using the Langevin equation to describe motions of beads, which allows one to introduce solvent viscosity. All-atom simulations or experimental data are used to parameterize the solvent model.

Among the recent achievements in the force field (FF) in Coarse-Grained Modeling, the MARTINI FF [266, 269] for biomolecular simulations, arises as

one of the most widespread and solid tool to explore the mesoscale level, and its features will be more extensively described in next subsection.

Moreover, the *Multiscale model* is capable to give a *backmapping* of the Coarse Grained system to a Fine Grained one, thus allowing for a consecutive step of relaxation and refinement of the CG simulation. Below we resume the main features of MARTINI coarse grained model and force field.

Features of MARTINI Force Field

The main feature of the most used Coarse Grained Force field, MARTINI, are presented in this section focusing in details to the principal critical points to manage before starting a simulation.

The first aspect to consider is strictly linked to the definition of the *Interaction Sites*. In fact, a four-to-one mapping is used in the model: this mean that a single interaction center represents four heavy atoms. In addition, four main types of interaction sites can be defined, namely polar (P), apolar (C), nonpolar (N), and charged (Q). For each particle type, a number of subtypes can be found, thus allowing a more accurate representation of the real chemical nature of the underlying atomic structure. Overall, 18 subtypes can be put within a main type, and they are distinguished by a single letter code, denoting its hydrogen-bonding capability [(d) donor, (a) acceptor, (da) both, (0) none] or by a number indicating the degree of polarity (from 1, low polarity, to 5, high polarity).

Ring moieties are treated with a special mapping since the classical four-to-one procedure is inadequate for preserving the geometry of small cyclic compounds. Thus the strategy include as many CG sites as necessary in order to keep the ring geometry, and this often results in a 2 or 3 to 1 mapping for each CG beads. Using this more detailed mapping, small cyclic compounds such as cyclohexane and steroids such as Cholesterol are well represented in their geometry. Hovewer, the CG beads have a large density, and thus the interaction parameters for ring beads need special treatment.

First to proceed with simulations, many important features of the CG models must be considered. The first concerns the introduction of *Antifreeze Particles*. In fact,

in the CG approach, water gains a freezing temperature that is too high compared to real one. In order to solve this problem and thus prevent the unwanted premature freezing of the CG water, an antifreeze agent can be added, similar to what is done in experimental setups performed at temperatures below the freezing point of water. This agents are called antifreeze particles, which correspond to special particle type denoted BP4 (Big P4), able to disturb the lattice packing of the uniformly sized solvent particles.

The second aspect to consider is the choice of the *Solvent* model. In acqueous environment, it generally represents four water molecules as a single type P4 CG site. In addition, other solvents are represented in other proper models which reproduce their own chemical-physical properties.

Another important treatment regards *Ions*. Commonly, the CG ions are represented by Q type particles and for monoatomic ions (*e.g.*, sodium, chloride) the first hydration shell is included in the CG representation. A correct balance must be taken in considering their with full charges and their increased hydration strength which could produce a short-range repulsion for positive-negative ion pairs. Finally a correct treatment of *Phospholipids* must be undertaken. In the CG representation, the phosphatidylcholine (PC) headgroup are formed by two hydrophilic groups: the Choline (type Q0) bearing a positive charge and the phosphate group (Qa) with a negative one. Beside, the PE lipids (phosphatidylethanolamine) have the positively charged group modeled as Qd, since it reflects better the hydrogen-bonding capacities of the amine moiety. The glycerol ester moiety is described by other two sites of intermediate hydrophilicity (Na). Each of the lipid tails is modeled by four hydrophobic particles (C1) (16 methylene/methyl moyeties) while Oleoyl tails are represented by five particles, four of C1 type and a central slightly more polar one (C3) to account for the polarizable nature of the double bond.

A particular attention need the *Cholesterol* representation, which is modeled by eight particles, among which six represents the steroid body and two the short tail. With this mapping the predominantly planar structure is well reproduced, and it still preserves the asymmetry between the two steroid faces. The cholesterol body is mapped on a 3:1 basis, using the special class of ring type particles. The ring

structure is kept rigid by the usage of a peculiar combination of bond, angle, and dihedral potentials.

Application Systems of CG Models: Case Studies

Improved Prediction of the Bilayer Properties

Many properties of lipid bilayers can be reproduced on a semi-quantitative level; these includes the area *per* headgroup for both saturated and unsaturated PC and PE lipids and the distribution of groups across the membrane

The use of a CG approach presents significant advantages given the current state of atomistic simulations. Among these, CG model allows many independent simulations in which state conditions can be systematically varied, and thus it can be applied to the *in silico* design of robust membranes, or to correlate lipid type and membrane properties. Another strength of the CG Martini model is the correct representation of lipid phase behavior, which require both considering very large systems due to the collective nature of phase transitions, and long simulation times in order to observe the critical nucleation events [270]. Some examples include transitions between micelles, bicelles, and vesicles and the formation of gel phases and inverted hexagonal and cubic phases. Moreover, the CG approach have been demonstrated to be efficient in predicitong the protein-lipid interactions of membrane embedded proteins. The most recent applications are grouped in several categories and concerns:

1) The binding modes prediction of proteins with membranes;

2) Simulating the preferential protein sorting, such as the propensity of TM peptides or membrane proteins to partition in either Ld or Lo domains;

3) The identification of specific lipid binding sites on membrane proteins;

4) The simulation of protein mediated membrane remodeling;

5) Membrane-mediated protein clustering and

6) Gating.

Some representative cases are reported below.

(a) *Membrane fusion* has been a very first applications of the Martini model and it represents still a great field of application, since only simulation studies can point out details that are not easily probed experimentally [271]. As a result, possible fusion pathways between lamellar membranes and between vesicles are now quite well described, and at present the efforts are directed towards calculating the energies of the various intermediates and elucidating the role of fusion promoting molecules such as PEG [272]. An important new direction is represented by the peptide and protein mediated fusion [273].

(b) Another research topic refers to the study of *Low-density and high-density lipoproteins* (LDL/HDL), which are involved in the Cholesterol transport in the bloodstream and thus play an important role in the development of atherosclerosis. Thus, the structure and dynamics of these particles have been studied with success, starting with the self-assembly of HDL model using the Martini variant developed by the Schulten group [265]. Subsequently, the structure and dynamics of HDL have been modeled with realistic size and lipid composition within the framework of standard Martini [274-276].

(c) Another case study regards the behavior of *lipid monolayers*, considering both pure lipid systems and a lipid combination with peptides and proteins. Although quantitative reproduction of the pressure-area isotherms at low surface pressure remains a challenge [277], the monolayers' properties in the liquid-expanded and liquid-condensed phases have been correctly describe by the Martini approach, which successfully reproduced also the collapse of lipid monolayers at highsurface pressure [278].

(d) Another application is related to the correct representation of *Membrane induced curvature.* A very important part of the living cells of biological

systems is the lipid membrane and the mechanical properties of this membrane play an important role in biophysical studies. Investigation of how the insertion of additional phospholipids in one leaflet of a bilayer affects the physical properties of the obtained asymmetric lipid membrane is of recent practical interest. Martini has proved to be capable of highlighting new insights on mechanical properties such as the computing of the pressure tensor, the lateral pressure, the surface tension and the first moment of lateral pressure in each leaflet of such a bilayer. Simulations show that adding more phospholipids into one monolayer results in asymmetrical changes in the lateral pressure of the individual bilayer leaflets. Interestingly, it has been observed that a change in phospholipid density in one leaflet affects the physical properties of unperturbed leaflet as well. Dissimilar lateral pressure in the bilayer leaflets results from changes in the contributing internal forces while the higher pressure in the top leaflet implies that this layer thermodynamically has a greater tendency to decrease the area with respect to the upper leaflet. As a consequence, spontaneous curvature is anticipated to emerge in the membrane [279].

(e) In conclusion, it must point out that Martini has a number of limitations as it can be found in any other model. However, the knowledge of the specific weaknesses are important in order to use the construct an appropriate model and to further improve it. Most of its limitations are shared with all CG models at a fundamental level. This regards in particular the chemical and spatial resolution, which are both limited compared to atomistic models, and the non correct balance between entropy and enthalpy (due to the reduced number of degrees of freedom). Other limitations are more specific for Martini and thus it could be improved in the next future with the level of coarse-graining approach which could accept a more detailed descriptions for parts of the studied system. In this way, adding back details in such a model can improve its accuracy, but this prevent some of its computational advantages such as the computation velocity [264]. This aspect will be focused in more details in next section.

Hybrid Atomistic–Coarse-Grained Biomolecular Dynamics

A new frontier for Coarse Grained model is oriented on speeding up calculation on solutes treated in full atoms (AA) embedded in coarse-grained environment (CG), *via* Hybrid Molecular Dynamics simulations [280]. The critical challenge of this hybrid modeling is the coupling of the different levels of resolution, which means how to describe the AA/CG interactions. Shi and coworkers re-parametrized the AA/CG interactions through force matching with the aim to simulate an atomistic TM peptide in a CG environment [281]; a different approach was followed, instead, by Michael and co-workers [282]. This regards in particular, the treatment of hybrid AA/CG Lennard-Jones interactions which are derived using standard mixing rules whereas the mixed charge-charge and charge-dipole interactions are calculated using the standard formulas employed in the pure models. Additional scaling parameters must be included for both Lennard-Jones and electrostatic AA/CG interactions, using an optimal scaling parameter which depends on the system size. Moreover, Masella and co-workers combined a polarizable pseudoparticle solvent with a polarizable atomistic force field while Wassenaar *et al.* [280] included explicit dielectric screening of the AA-AA interactions in presence of charged CG particles, which requires direct Coulomb interactions between AA and CG particles.

As an application, hybrid simulations are mainly focused on membrane-protein, where the possibility of speeding up the simulation with a coarse grained environment, maintaining the full atom resolution for the protein, could lead to the observation of conformational changes events for the protein backbone, occurring over long timescales.

AKNOWLEDGEMENTS

Declared none.

CONFLICT OF INTEREST

The authors confirm that this chapter content has no conflict of interest.

REFERENCES

[1] Kox, A. J.; Michels, J. P. J.; Wiegel, F. W. Simulation of a lipid monolayer using molecular dynamics. *Nature*, **1980**, *287*, 317-319.

[2] Van der Ploeg, P.; Berendsen, H. J. C. Molecular dynamics simulation of a bilayer membrane. *J. Chem. Phys*., **1982**, *76*, 3271-3276.

[3] Van der Ploeg, P.; Berendsen, H. J. C. Molecular dynamics of a bilayer membrane. *Mol. Phys*., **1983**, *49*, 233-248.

[4] Kjelander, R.; Marcelja, S. Polarization of water between molecular-surfaces—A molecular dynamics study, *Chem. Scr.*, **1985**, *25* (1), 73-80.

[5] (a) Heller, H.; Schaefer, M.; Schulten, K. Molecular dynamics simulation of a bilayer of 200 lipids in the gel and in the liquid-crystal phases. *J. Phys. Chem*., **1993**, *97*, 8343-49. (b) Marrink, S. J.; Berendsen, H. J. C. Simulation of Water Transport through a Lipid Membrane. *J. Phys. Chem.*, **1994**, *98*, 4155-4168.

[6] Tieleman, D. P.; Berendsen, H. J. C. Molecular dynamics simulations of a fully hydrated dipalmitoylphosphatidylcholine bilayer with different macroscopic boundary conditions and parameters. *J. Chem. Phys*., **1996**, *105*, 4871-4880.

[7] Wohlert, J.; Edholm, O. Dynamics in atomistic simulations of phospholipid membranes: Nuclear magnetic resonance relaxation rates and lateral diffusion. *J. Chem. Phys*., **2006**, *125*(20), 204703.

[8] Bjelkmar, P.; Niemel, P. S.; Vattulainen, I.; Lindahl, E. Conformational changes and slow dynamics through microsecond polarized atomistic molecular simulation of an integral Kv1.2 ion channel. *PLoSComp.Biol*. **2009**, *5*(2), 1000289.

[9] (a) Kuhner, M.; Tampe, R.; Sackmann, E. Lipid mono- and bilayer supported on polymer films: composite polymer-lipid films on solid substrates. *Biophys J*., **1994**, *67*(1), 217-226; (b)Castellana, E. T.; Cremer, P. S. Solid supported lipid bilayers: From biophysical studies to sensor design. *Surface Science Reports*, **2006**, *61*(10), 429-444.

[10] Merz, M. Molecular dynamics simulations of lipid bilayers. *Curr. Opin. Struct. Biol*., **1997**, *7*(4), 511-517.

[11] Berendsen, H. J. C.; Tieleman D. P. Molecular dynamics: Studies of lipid bilayers. In: *Encyclopedia of Computational Chemistry*, P. von Rague Schleyer Ed.; John Wiley & Sons, **1998**, Vol. *3*, pp.1639–1650.

[12] Feller, S. E.; MacKerell J. A. D. An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids. *J. Phys. Chem. B*, **2000**, *104*, 7510-7515.

[13] Feller, S. E. Molecular dynamics simulations of lipid bilayers. *Curr. Opin. Colloid Interf. Sci*., **2000**, *5*, 217-223.

[14] Forrest, L. R.; Sansom, M. S. P. Membrane simulations: bigger and better? *Curr. Opin. Struct. Biol,* **2000**, *10*(2), 174-81.

[15] Damodaran, K. V.; Merz, K. M. Computer Simulation of Lipid Systems. In: *Review in Computational Chemistry*, Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, **2005** Vol. *5*, 269-298.

[16] Mouritsen, O. G.; Jorgensen, K. Dynamical order and disorder in lipid bilayers. *Chem. Phys. Lipids*, **1994**, *73*(1-2), 3-25.

[17] Pastor, R. W. Molecular dynamics and Monte Carlo simulations of lipid bilayers. *Curr. Opin. Struct. Biol*., **1994**, *4*, 486-492.

[18] Mouritsen, O. G.; Sperotto, M. M.; Risbo, J.; Zhang, Z.; Zuckermann, M. J. Computational approach to lipid-protein interactions in membranes. *Adv. Comput. Biol*., **1996**, *2*, 15-64.

[19] Jacobsson, E. Computer simulation studies of biological membranes: progress, promise and pitfalls. *Trends Biochem. Sci*., **1997**, *22*, 339-344.

[20] Tieleman, D. P.; Marrink S. J.; BerendsenH. J. C. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *Biochim Biophys Acta,* **1997**, *1331*(3), 235-270 .

[21] Tobias, D. J.; Tu, K.; Klein, M. L. Atomic-scale molecular dynamics simulations of lipid membranes, *Curr Opin Colloid Interface Sci*., **1997**, *2*, 15-26.

[22] Huang, P.; Perez, J.J.; Loew, G.H. Molecular dynamics simulations of phospholipid bilayers. *J Biomol Struct Dyn*, **1994**, *11*, 927–956.

[23] Egberts, E. ; Marrink, S.J.; Berendsen, H.J.C. Molecular dynamics simulation of a phospholipid bilayers, *Eur Biophys J,* **1994,** *22*, 423–436.

[24] Feller, S.E**.**; Pastor R.W. On simulating lipid bilayers with an applied surface tension: periodic boundary conditions and undulations. *Biophys J*., **1996**, *71*(3), 1350-1355.

[25] Pastor, R.W.; Venable, R.M., Feller, S.E. Lipid bilayers, NMR relaxation, and computer simulations. *Acc Chem Res*., **2002**, *35*(6), 438-46.

[26] Saiz, L.; Bandyopadhyay, S.; Klein, M.L. Towards an understanding of complex biological membranes from atomistic molecular dynamics simulations. *Biosci Rep*., **2002,** *22*(2), 151-73.

[27] Katsaras, J.; Gutberlet, T. Lipid Bilayers In: *Structure and Interactions*, Springer, **2001**, 89-105.

[28] *Computational Biochemistry and Biophysics.* Becker, O. H.; MacKerell, A.D.; Roux, B.; Watanabe, M. Eds., Marcel Dekker Inc., New York, **2001**, pp.1-465.

[29] Balali-Mood, K., Harroun, T.A., Bradshaw, JP. Molecular dynamics simulations of a mixed DOPC/DOPG bilayer. *Eur Phys J E Soft Matter*. **2003**, *12*(1), S135-40.

[30] Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular dynamics simulations. *Curr. Opin. Struct.Biol*., **2002**, *12*, 190-196.

[31] Vigh, L.; Escrib'a P.; Sonnleitner, A.; Sonnleitner, M.; Piotto, S.; Maresca, B.; Horv´ath, I.; Harwood, J. L. The significance of lipid composition for membrane activity: new concepts and ways of assessing function. *Prog. Lipid Res*., **2005**, *44*(5), 303-344.

[32] Chan, Y. H. M.; Boxer, S. G. Model membrane systems and their applications. *Curr. Opin. Chem. Biol.,* **2007**, *11*(6), 581-587.

[33] Computational Modeling of Membrane Bilayers. In: *Current Topics in Membranes*, Feller S.E. Ed, Academic Press, Elsevier Inc., San Diego, **2008**, *60*, pp 1-466.

[34] Lyubartsev, A. P.; Rabinovich, A. L. Recent development in computer simulations of lipid bilayers, *Soft Matter*, **2011**, *7*, 25-39.

[35] Scott, H.L. Modeling the lipid component of membranes, *Curr. Opin. Struct. Biol*., **2002**, *12*(4), 495-502 .

[36] Klein, M.L.; Shinoda, W. Large-scale molecular dynamics simulations of self-assembling systems. *Science*, **2008**, *321*(5890), 798-800.

[37] Saiz, L.; Klein, M. L. Computer simulation studies of model biological membranes. *Acc. Chem. Res.,* **2002**, *35*(6), 482-9 .

[38] Orsi, M., Michel, J.; Essex, J.W. Coarse-grain modeling of DMPC and DOPC lipid bilayers. *J Phys Condens Matter*., **2010**, *22*(15), 155106.

[39] Hirtz, M.; Kumar, N.; Chi, L. Simulation modeling of supported lipid membranes - a review. *Curr Top Med Chem*., **2014**, *14*(5), 617-23.

[40] Sezgin, E.; Shwille, P. Model membrane platforms to study protein-membrane interactions. *Mol Membr Biol*, **2012,** *29*(5), 144-154.

[41] Vermeer, L. S.; De Groot, B. L.; Reat, V.; Milon, A.; Czaplicki, J. Acyl chain order parameter profiles in phospholipid bilayers: computation from molecular dynamics simulations and comparison with 2H NMR experiments. *J. Eur. Biophys. J.,* **2007**, *36*(8), 919-931.

[42] Pasqua, A.; Maibaum, L.; Oster, G.; Fletcher, D. A.Geissler P L. Large-scale simulations of fluctuating biological membranes. *J Chem Phys.,* **2010**, *132*(15), 154107.

[43] Marrink, S. J.; De Vries, A. H.L; Tieleman D. P. Lipids on the move: simulations of membrane pores, domains, stalks and curves. *Biochim. Biophys. Acta,* **2009**, *1788*(1), 149-68.

[44] Pandit, S. A.; Scott, H. L. Multiscale simulations of heterogeneous model membranes. *Biochim. Biophys. Acta*, **2009**, *1788*(1), 136-148.

[45] Gurtovenko, A. A.; Anwar, J.; Vattulainen, I. Defect-mediated trafficking across cell membranes: insights from in silico modeling. *Chem Rev*., **2010**, *110*(10), 6077-6103.

[46] Lyubartsev, A. P.; Rabinovich, A. L. Recent development in computer simulations of lipid bilayers. *Soft Matter*, **2011**, *7*, 25-39.

[47] Yanga, K.; Ma, Y.Q. Computer simulations of fusion, fission and shape deformation in lipid membranes. *Soft Matter*, **2012**, *8*, 606-618.

[48] Molecular Simulation and Biomembranes: from biophysics to function. Samson, M. S. P.; Biggins, P.C. Eds., *RSC Biomolecular Science* n.*20*, RSC Publishing, Cambridge, UK, **2010**, pp.1-330.

[49] Ash, W.L.; Zlomislic, M.R.; Oloo, E.O.; Tieleman, D.P. Computer simulations of membrane proteins. *Bioch Biophys Acta*, **2004**, *1666*(1-2), 158-89.

[50] (a) Lindhal, E.; Samson, M.S. Membrane proteins: molecular dynamics simulations. *Curr Opin Struct Biol.*, **2008**, *18*(4), 425-31. (b) Roux, B.; Shoulten, K. Computational studies of membrane channels. *Structure,* **2004**, *128*(8), 1343-51.

[51] Bond, P.J.; Samson, M.S.P. The simulation approach to bacterial outer membrane proteins, *Mol Membr Biol*., **2004**, *21*(3), 151-61.

[52] Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P.M. A consistent empirical potential for water–protein interactions. *Biopolymers*, **1984**, *23*(8), 1513-1518.

[53] Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase. *J. Comp. Chem*., **2001**, *22*(11), 1205-1218.

[54] Chandrasekhar, I.; Kastenholz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, D. P.; van Gunsteren, W. F. A consistent potential energy parameter set for lipids: dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field. *Eur. Biophys. J*., **2003**, *32*, 67-77.

[55] Skjevik, A.A.; Madej, B. D.; Walker, R.; Teigen, C. LIPID11: a modular framework for lipid simulations using AMBER. *J Phys Chem B*, **2012**, *116* (36), 11124-36.

[56] Mackerell, A. D.; Wi´orkiewicz-Kuczera, J.; Karplus, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *J.Am. Chem. Soc.,* **1995**, *117*, 11946-75.

[57] MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fisher, S.; Gao, J.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.;

Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wi´orkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B,* **1998**, *102*, 3586-616.

[58] Chandrasekhar, I.; Kastenholz, M.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, D. P.; van Gunsteren, W. F. A consistent potential energy parameter set for lipids: dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field. *Eur. Biophys. J.,* **2003**, *32* (1), 67-77.

[59] Berger, O.; Edholm, O.; J¨ahnig, F. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature, *Biophys. J.,* **1997**, *72* (5), 2002-2013.

[60] Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model*, **2001**, *7* (8), 306-17.

[61] Case, D.A.; Babin, V.; Berryman, J.T.; Betz, R.M.; Cai, Q. ; Cerutti, D.S.; Cheatham III, T.E.; Darden, T.A.; Duke, R.E.; , Gohlke, H.; Goetz, A.W. ; Gusarov, S. ; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A. ; Lee, T.S.; LeGrand, S. ; Luchko, T.; Luo, R.; Madej, B.; Merz, K.M.; Paesani, F. ; Roe, D.R.; Roitberg, A. ; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C.L. ; Smith, W. ; Swails, J. ; Walker, R.C; . Wang, J.; Wolf, R.M. ; Wu; X.; Kollman, P.A. The AMBER molecular modeling package AMBER 14, University of California, San Francisco **2014**.

[62] Dickson, C. J.; Madej, B. D.; Skjevik, Å. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker R. C. Lipid14: the Amber Lipid Force Field, *J. Chem. Theory Comput.*, 2014, *10* (2), 865-879.

[63] Feller, S. E.; Yin, D.; Pastor, R.W.; MacKerrel, Jr.A. D.Jr Molecular dynamics simulation of unsaturated lipid bilayers at low hydration: parameterization and comparison with diffraction studies. *Biophys. J.*, **1997**, *73* (5), 2269-79.

[64] Prakash, P˙, Sankararamakrishnan, R. Force field dependence of phospholipid headgroup and acyl chain properties: comparative molecular dynamics simulations of DMPC bilayers. *J Comput Chem.*, **2010**, *31*(2):266-77.

[65] Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. NAMD2: greater scalability for parallel molecular dynamics, *J. Comp.Phys*., **1999**, *151*, 283

[66] Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D., Jr.; Pastor, R. W. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types, *J. Phys. Chem. B*, **2010**, *114* (23), 7830-43.

[67] Harder, E.; MacKerell, A. D. Jr. Roux, B. Many-body polarization effects and the membrane dipole potential. *J. Am. Chem. Soc.*, **2009**, *131* (8), 2760-1.

[68] Vorobyov, I.; Allen, T. W. The electrostatics of solvent and membrane interfaces and the role of electronic polarizability, *J. Chem. Phys*, **2010**, *132*, 185101.

[69] Bauer, B. A.; Lucas, T. R.; Meninger, D. J.; Patel S. Water Permeation Through DMPC Lipid Bilayers using Polarizable Charge Equilibration Force Fields. *Chem. Phys. Lett*., **2011**, *508*(4-6), 289-294.

[70] Hyvönen, M .T.; Kovanen, P. T. Nuclear magnetic resonance study of sphingomyelin bilayers, *J. Phys. Chem. B*, **2003**, *107* (34), 9102-8.

[71] Pitman, M. C.; Suits, F.; MacKerell, A. D. Jr.; Feller, S. E. Molecular-level organization of saturated and polyunsaturated fatty acids in a phosphatidylcholine bilayer containing cholesterol. *Biochemistry*, **2004**, *43* (49), 15318-28.

[72] Bauer, B. A.; Patel, S. Recent applications and developments of charge equilibration force fields for modeling dynamical charges in classical molecular dynamics simulations. *Theor. Chem. Acc*., **2012**, 131, 1153-1159.

[73] Rabinovich, A. L.; Lyubartsev, A. P. Computer simulation of lipid membranes: Methodology and achievements. *Polymer Science, Ser. C,* **2013**, *55*, 162-180.

[74] Robinson, D. A polarizable force-field for membrane raft components. *J. Chem. Theory Comput.*, **2013**, *9*, 2498-2503.

[75] Polley, A.; Vemparala, S.; Rao, M. Intra- and Intermolecular Charge Fluxes Induced by the OH Stretching Mode of Water and Their Effects on the Infrared Intensities and Intermolecular Vibrational Coupling. *J. Phys. Chem. B*, **2012**, *10*(42), 13403-13409.

[76] Porasso, R. D.; López Cascales, J.J. Study of the effect of Na^+ and Ca^{2+} ion concentration on the structure of an asymmetric DPPC/DPPC + DPPS lipid bilayer by molecular dynamics simulation. *Colloids Surf. B Biointerface*, **2009**, *73*(1), 42-50.

[77] Leekumjorn, S.; Sum, A. K. Molecular simulation study of structural and dynamic properties of mixed DPPC/DPPE. *Bilayers Biophys J*., **2006**, *90*(11), 3951-3965.

[78] de Vries, A. H.; Mark, A. E.; Marrink, S. J. The Binary Mixing Behavior of Phospholipids in a Bilayer: A Molecular Dynamics Study. *J. Phys. Chem. B*., **2004**, *108*(7), 2454-2463.

[79] Ghahremanpour, M. M.; Arab, S. S.; Aghazadeh, S. B.; Zhang, J.; van der Spoel, D. MemBuilder: a web-based graphical interface to build heterogeneously mixed membrane bilayers for the GROMACS biomolecular simulation program. *Bioinformatics*, **2014**, *30*(3), 439-441.

[80] Martínez, L.; Andrade, R.; Birgin, E.; Martínez, J. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J.Comput. Chem*., **2009**, *30*(13), 2157-2164.

[81] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM . *J.Comput. Chem*., **2008**, *29*(11), 1859-1865.

[82] Lemmin, T.; Bovigny, C.; Lancon, D. ; Dal Peraro, M. Cardiolipin Models for Molecular Simulations of Bacterial and Mitochondrial Membranes. *J Chem Theory Comp*, **2013,** *9* (1), 670-678.

[83] Tieleman, D. P.; Forrest, L. R.; Sansom, M. S.; Berendsen, H. J. Lipid properties and the orientation of aromatic residues in OmpF, influenza M2, and alamethicin systems: molecular dynamics simulations. *Biochemistry*, 1998, *37*(50), 17554-61.

[84] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem*., **1987**, *91* (24), 6269-6271.

[85] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys*., **1983**, *79*, 926-935.

[86] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Molec. Graphics*, **1996**, *14* (1), 33-8, 27-28.

[87] Anézo, C.; de Vries, A. H.; Höltje, H.-D.; Tieleman, D. P.; Marrink, S. J.. Methodological Issues in Lipid Bilayer Simulations. *J. Phys. Chem. B*, **2003**, *107* (35), 9424-9433.

[88] (a) Niemelä, P. S.; Hyvönen, M. T.; Vattulainen I. Influence of Chain Length and Unsaturation on Sphingomyelin. *Bilayers Biophys J*, .**2006**, *90* (3), 851-63. (b)Niemelä, P.;

Hyvönen, M. T.; Vattulainen, I. Structure and Dynamics of Sphingomyelin Bilayer: Insight Gained through Systematic Comparison to Phosphatidylcholine. *Biophys. J.*, **2004**, *87* (5), 2976-2989.

[89] Polley, A.; Vemparala, S.; Rao, M. Atomistic Simulations of a Multicomponent Asymmetric Lipid Bilayer. *J. Phys. Chem. B*, **2012**, *116*, 13403-13410.

[90] Saiz, L.; Klein, M. L. Structural properties of a highly polyunsaturated lipid bilayer from molecular dynamics simulations. *Biophys. J.*, **2001**, *81*(1), 204-16.

[91] Markvoort, A. J.; Marrink, S. *Lipid acrobatics in the membrane fusion arena. J. Curr. Top. Membr*, **2011**, *68*, 259-294.

[92] (a) Kasson, P. M. ; Pande, V.S. A bundling of viral fusion mechanisms. *Proceedings of the National Academy of Sciences*, **2011,** *108* (10), 3827-28; (b) Kasson, P.M.; Lindahl, E.; Pande, V.S. Water ordering at membrane interfaces controls fusion dynamics. *J Am Chem Soc*., **2011,** *133*, 3812-15; (c) Kasson, P.M.; Lindahl, E.; Pande, V.S. Atomic-resolution simulations predict a transition state for vesicle fusion defined by contact of a few lipid tails. *PLoS Computational Biology*, **2010**, *6*, 24 (6), e1000829.

[93] Porasso, R.D.; Lopez Cascales, J.J.; A criterion to identify the equilibration time in lipid bilayer simulations. *Papers in Physiscs*, **2012**, *4*, 040005.

[94] van Meer, G. Cellular lipidomics. *EMBO J*., **2005**, *24* (18), 3159-65.

[95] Berkowitz, M.L. Detailed molecular dynamics simulations of model biological membranes containing cholesterol. *Biochim. Biophys. Acta Biomembr*, **2009**, *1788*(1), 86-96.

[96] Lipowsky, R. Generic interactions of flexible membranes. In**:** *Structure and Dynamics of Membranes,* Sackmann, Lipowsky, E., R. ; Sackmann E. (Eds.), Elsevier, **1995,** 1–64.

[97] Kusumi, A.; Tsuda, M.; Akino, T.; Ohnishi, S; Terayama, Y. Protein-Phospholipid-Cholesterol Interaction in the Photolysis of Invertebrate Rhodopsin. *Biochemistry*, **1983**, *22*, 1165-1170;

[98] Subczynski, W.K.; Wisniewska, A.; Yin, J.J.; Hyde, J.S.; Kusumi, A. Hydrophobic barriers of lipid bilayer membranes formed by reduction of water penetration by alkyl chain unsaturation and cholesterol. *Biochemistry*, **1994**, *33*(4), 7670-7681.

[99] Bloom, M.; Evans, E; Mouritsen, O.G. Physical properties of the fluid lipid-bilayer component of cell membranes: a perspective. *Q. Rev. Biophys*, **1991**, *24*, 293-397.

[100] (a) Leathes, J.B. Role of fats in vital phenomena. *Lancet*, **1925**, *208*, 853-856; (b) Demel, R.A.; van Deenen, L.L.M.; Pethica, B.A. Monolayer interactions of phospholipids and cholesterol. *Biochim Biophys Acta*, **1967**, *135*, 11-19; (c) Stockton, B.W.; Smith, I.C.P. A deuterium NMR study of the condensing effect of cholesterol on egg phosphatidylcholine bilayer membranes. *Chem. Phys. Lipids*, **1976**, *17*, 251-263; (b) Vist, M.; Davis, J.H. Phase equilibria of cholesterol/dipalmitoylphosphatidylcholine mixtures: deuterium nuclear magnetic resonance and differential scanning calorimetry. *Biochemistry*, **1990**, *29* (2), 451-464.

[101] McMullen, T.P.; McElhaney, R.N. Physical studies of cholesterol-phospholipid interactions. *Curr. Opin. Colloid Interface Sci*., **1996**, *1*, 83-90.

[102] Oldfield, E.; Meadows, M.; Rice, D.; Jacobs, R. Spectroscopic studies of specifically deuterium labeled membrane systems. Nuclear magnetic resonance investigation of the effects of cholesterol in model systems. *Biochemistry*, **1978**, *17*(14), 2727-2740.

[103] Marsh, D.; Smith, I.C.P. Interacting spin labels as probes of molecular separation within phospholipid bilayers. *Biochim. Biophys. Res. Commun*., **1972**, *493*(4), 916-922.

[104] Trouard, T.P.; Nevzorov, A.A.; Alam, T.M.; Job, C.; Zajicek, J. Brown, M.F.; Influence of cholesterol on dynamics of dimyristoylphosphatidylcholine bilayers as studied by deuterium NMR relaxation. *J. Chem. Phys*., **1999**, *110*, 8802-8818.

[105] Smaby, J.M.; Momsen, M.; Brockman, H.L.; Brown, R.E. Phosphatidylcholine acyl unsaturation modulates the decrease in interfacial elasticity induced by cholesterol. *Biophys. J.*, **1997**, *73* (3), 1492-1505.

[106] Davis, J.H. The description of membrane lipid conformation, order and dynamics by 2H-NMR. *Biochim. Biophys. Acta*, **1983**, *737* (1), 117-171.

[107] Pasenkiewicz-Gierula, M.; Ro¨ g, T.; Kitamura, K.; Kusumi, A. Cholesterol effects on the phosphatidylcholine bilayer polar region: a molecular simulation study. *Biophys. J*., **2000**, *78* (3), 1376-1389.

[108] Rög, T.; Pasenkiewicz-Gierula, M. Cholesterol effects on the phospholipid condensation and packing in the bilayer: a molecular simulation study. *FEBS Letters*, **2001**, *502*(1-2), 68-71.

[109] Huang, J; Feigenson, G.W.A Microscopic interaction model of maximum solubility of cholesterol in lipid bilayers. *Biophys. J*., **1999**, *76* (4), 2142-2157.

[110] Sugahara, M.; Uragami, M.; Yan, X.; Regen, S.L. The structural role of cholesterol in biological membranes *J. Am. Chem. Soc*, **2001**, *123* (32), 7939-7940.

[111] Daly, T. A.; Wang, M.; Regen, S. L. The Origin of Cholesterol's Condensing Effect *Langmuir,* **2011,** *27,* 2159-2161.

[112] (a) Pitman, M. C.; Suits, F.; MacKerell, A. D.; Feller S. E. Molecular-level organization of saturated and polyunsaturated fatty acids in a phosphatidylcholine bilayer containing cholesterol *Biochemistry*, **2004**, *43*(49), 15318-15328; (b) Carrillo-Tripp, M.; Feller S. E. Evidence for a mechanism by which omega-3 polyunsaturated lipids may affect membrane protein function *Biochemistry*, **2005**, *44* (30), 10164-9.

[113] Rabinovich, A. L.; KornilovV. V.; Balabaev, N. K.; Leermakers, F. A. M.; Filippov, A. V. Type Iγ phosphatidylinositol phosphate kinase modulates adherens junction and E-cadherin trafficking via a direct interaction with μ1B adaptin. *Biochemistry [Moscow] , Ser. A: Membr. and Cell Biol*., **2007**, *176* (3), 343-353.

[114] (a)Hofsäß, C.; Lindahl, E.; Edholm, O. Molecular Dynamics Simulations of Phospholipid Bilayers with Cholesterol. *Biophys. J*., **2003**, *84* (4), 2192-2206; (b) Jedlovszky, P.; Mezei, M. Effect of Cholesterol on the Properties of Phospholipid Membranes. 1. Structural Features. *J. Phys. Chem. B,* **2003**, *107* (22), 5311-5321.

[115] (a) Falck, E.; Patra, M.; Karttunen, M.; Hyvönen, M. T.; Vattulainen, I. Lessons of slicing membranes: interplay of packing, free area, and lateral diffusion in phospholipid/cholesterol bilayers. *Biophys. J* ., **2004**, *87* (2), 1076-1091; (b) Khelashvili, G. A.; Scott, H. L. Combined Monte Carlo and molecular dynamics simulation of hydrated 18:0 sphingomyelin-cholesterol lipid bilayers. *J. Chem. Phys*, **2004**, *120* (20), 9841-9847; (c) Pandit, S. A.; Bostick, D.; Berkowitz, M. L. Complexation of Phosphatidylcholine Lipids with Cholesterol. *Biophys. J.,* **2004**, *86* (3), 1345-1356.

[116] Czub, J.; Baginski, M. Comparative Molecular Dynamics Study of Lipid Membranes Containing Cholesterol and Ergosterol. *Biophys. J.,* **2006**, *90* (7), 2368-2382.

[117] Martinez-Seara, H.; R´og, T.; Pasenkiewicz-Gierula, M.; Vattulainen, I.; Karttunen, M.; Reigada, R. Interplay of unsaturated phospholipids and cholesterol in membranes: Effect of the double bond position. *Biophys. J,* **2008**, *95*, 3295-3305.

[118] Mihailescu, M.; Vaswani, R.G.; Jardon-Valadez, E.; Castro-Roman, F.; Freites, J.A.; Worcester, D.L.; Chamberlin, A.R.; Tobias, D.J.; White, S.H. Acyl-Chain Methyl Distributions of Liquid-Ordered and -Disordered Membranes. *Biophys. J*, **2011**, *100* (6), 1455-1462.

[119] Olsen, B.N.; Schlesinger, P.H.; Baker, N.A.L. Perturbations of membrane structure by cholesterol and cholesterol derivatives are determined by sterol orientation. *J. Am. Chem. Soc.,* **2009**, *131* (13), 4854-4865.

[120] Olsen, B.N.; Schlesinger, P.H.; Ory, D.S.; Baker, N.A. 25-Hydroxycholesterol increases the availability of cholesterol in phospholipid membranes. *Biophys. J.*, **2011**, *100* (4), 948-956.

[121] Hall, A.; Rog, T.; Karttunen, M.; Vattulainen, I. Role of glycolipids in lipid rafts: a view through atomistic molecular dynamics simulations with galactosylceramide. *J. Phys. Chem. B*, **2010**, *114* (23), 7797-807.

[122] Hall, A.; Rog, T.; Vattulainen, I. Effect of galactosylceramide on the dynamics of cholesterol-rich lipid membranes. *J. Phys. Chem. B*, **2011,** *115* (49), 14424-14434.

[123] Kapla, J.; Stevensson, B.; Dahlberg, M.; Maliniak, A. Molecular dynamics simulations of membranes composed of glycolipids and phospholipids. *J. Phys. Chem.B*, **2011**, *116* (1), 244-252.

[124] Lupyan, D.; Mezei, M.; Logothetis, D.E.; Osman, R.A Molecular dynamics investigation of lipid bilayer perturbation by PIP2. *Biophys. J.,* **2010**, *98* (2), 240-247.

[125] Zidar, J.; Merzel, F.; Hodoscek, M.; Rebolj, K.; Sepcic, K.; Macek, P.; Janezic, D. Liquid-ordered phase formation in cholesterol/sphingomyelin bilayers: all-atom molecular dynamics simulations. *J. Phys. Chem. B*, **2009**, *113* (48), 15795-802.

[126] (a) Gonzalez-Rubio, P.; Gautier, R.; Etchebest, C.; Fuchs, P.F.J. Amphipathic-Lipid-Packing-Sensor interactions with lipids assessed by atomistic molecular dynamics. *Biophys. Acta Biomembr*, **2011**, *1808* (9), 2119-2127; (b) Alwarawrah, M.; Dai, J.; Huang, J. Modification of lipid bilayer structure by diacylglycerol: a comparative study of diacylglycerol and cholesterol. *J. Chem. Theory Comput*., **2012**, *8* (2), 749-758.

[127] (a) Pandit, S.A.; Chiu, S.W., Jakobsson, E.; Grama, A.L; Scott, H.L Cholesterol packing around lipids with saturated and unsaturated chains: a simulation study. *Langmuir*, **2008**, *24* (13), 6858-6865; (b) Cournia, Z.; Ullmann, G.M.; Smith, J.C. Differential effects of cholesterol, ergosterol and lanosterol on a dipalmitoyl phosphatidylcholine membrane: a molecular dynamics simulation study. *J. Phys. Chem. B*, **2007**, *111* (7), 1786-1801.

[128] Rog, T.; Vattulainen, I.; Jansen, M.; Ikonen, E.; Karttunen, M. Comparison of cholesterol and its direct precursors along the biosynthetic pathway: effects of cholesterol, desmosterol and 7-dehydrocholesterol on saturated and unsaturated lipid bilayers. *J. Chem. Phys*, **2008**, *129* (15), 154508.

[129] Jo, S.; Lim, J.B.; Klauda, J.B.; Im, W. CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. *Biophys. J.*, **2009,** *97* (1), 50-58.

[130] Piggot, T.J.; Holdbrook, D.A.; Khalid, S. Electroporation of the E. coli and S. Aureus membranes: molecular dynamics simulations of complex bacterial membranes. *J. Phys. Chem. B*, **2011**, *115*(4), 13381-13388.

[131] Polley, A.; Vemparala, S.; Rao, M. Atomistic simulations of a multicomponent asymmetric lipid bilayers. *J. Phys. Chem. B*, **2012**, *116* (45), 13403-13410.

[132] Robinson, D.; Besley, N. A.; O'Shea, P.; Hirst, J. D. Water order profiles on phospholipid/cholesterol membrane bilayer surfaces. *J. Comput. Chem*, **2011**, *32* (12), 2613-8.

[133] Mihailescu, M.; Vaswani, R. G.; Jardón-Valadez, E.; Castro- Román, F.; Freites, J. A.; ZWorcester, D. L.; Chamberlin, A. R.; Tobias, D. J.; White, S. H. Acyl-chain methyl distributions of liquid-ordered and -disordered membranes. *Biophys. J*, **2011**, *100* (6), 1455-62.

[134] O'Connor, J. W.; Klauda, J. B. Lipid membranes with a majority of cholesterol: applications to the ocular lens and aquaporin 0. *J. Phys. Chem. B*, **2011**, *115* (20), 6455-64.

[135] Zidar, J.; Merzel, F.; Hodoscek, M.; Rebolj, K.; Sepcic, K.; Macek, P.; Janezic, D. Liquid-ordered phase formation in cholesterol/sphingomyelin bilayers: all-atom molecular dynamics simulations. *J. Phys. Chem. B*, **2009**, *113*(48), 15795-812.

[136] Simons, K.; Ikonen, E. ; Functional rafts in cell membranes. *Nature*, **1997**, *387* (6633), 569-72.

[137] Simons, K.; Toomre, D. Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol,* **2000**, *1*(1), 31-39.

[138] Berkowitz, M.L. Detailed molecular dynamics simulations of model biological membranes containing cholesterol. *Bioch. Bioph. Acta*, **2009**, *1788*(1), 86-96.

[139] Robinson, D.A. Polarizable Force-Field for Cholesterol and Sphingomyelin. *J. Chem. Theory Comput*, **2013**, *9*(5), 2498-2503.

[140] van Meer, G., Voelker, D.R.; Feigenson, G.W. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol*. **2008**, *9*(2), 112-124.

[141] Garg, S.; Porcar, L.; Woodka, A.C.; Butler, P.D.; Perez-Salas, U. Noninvasive neutron scattering measurements reveal slower cholesterol transport in model lipid membranes. *Biophys. J.,* **2011**, *101*(2), 370-377.

[142] Bennett, W.F.D.; MacCallum, J.L.; Hinner, M.J.; Marrink, S.J.; Tieleman, D.P. Molecular view of cholesterol flip-flop and chemical potential in different membrane environments. *J. Am. Chem. Soc*., **2009**, *131* (35), 12714-20.

[143] Bennett, W.F.D.; Tieleman, D.P. Molecular simulation of rapid translocation of cholesterol, diacylglycerol, and ceramide in model raft and nonraft membranes. *J. Lipid Res*., **2012**, *53* (3), 421-429.

[144] Bennett, W.F.D.; Tieleman, D.P. Computer simulations of lipid membrane domains. *Bioch. et Bioph.Acta*, **2013**, *1828*(8), 1765-76.

[145] McMullen, T.P.W.; Lewis, R.N.A.H.; McElhaney, R.N. Cholesterol–phospholipid interactions, the liquid-ordered phase and lipid rafts in model and biological membranes. *Curr. Opin. Colloid Interface Sci*, **2004**, *8*, 459-468.

[146] De Joannis, J.; Coppock, P.S.; Yin, F.; Mori, M.; Zamorano, A.; Kindt, J.T. Atomistic simulation of cholesterol effects on miscibility of saturated and unsaturated phospholipids: implications for liquid-ordered/liquid-disordered phase coexistence. *J. Am. Chem. Soc*., **2011**, *133*(10), 3625-34.

[147] (a) Zhang, Z.; Lu, L.; Berkowitz, M.L. Energetics of cholesterol transfer between lipid bilayers. *J. Phys. Chem. B*, **2008**, 112(12), 3807-11; (b) Bennett, W.F.D.; MacCallum, J.L.; Tieleman, D.P. Thermodynamic analysis of the effect of cholesterol on dipalmitoylphosphatidylcholine lipid membranes. *J. Am. Chem. Soc*., **2009**, *131*(5), 1972-8.

[148] Ostrom, R.S.; Insel, P.A. The evolving role of lipid rafts and caveolae in G protein-coupled receptor signaling: implications for molecular pharmacology. *Br. J. Pharmacol*., **2004**, *143* (2), 235-245.

[149] Maguy, A.; Hebert, T.E.; Nattel, S. Involvement of lipid rafts and caveolae in cardiac ion channel function. *Cardiovasc. Res*., **2006**, *69*(4), 798-807.

[150] Wüstner, D. Fluorescent sterols as tools in membrane biophysics and cell biology. *Chem. Phys. Lipids*, **2007**, *146* (1), 1-25.

[151] (a) Engelman, D. M. Membranes are more mosaic than fluid. *Nature*, **2005**, *438* (7068), 578-80. (b) Simons, K.; Ikonen, E. Functional rafts in cell membranes. *Nature*, **1997**, *387* (6633), 569-72.

[152] Sasaki, H.; White, S.H. A novel fluorescent probe that senses the physical state of lipid bilayers. *Biophysical Journal*, **2009**, *96* (11), 4631-41.

[153] Robalo, J. R.; Ramalho, J. P. P.; Loura, L. M. S. NBD-labeled cholesterol analogues in phospholipid bilayers: insights from molecular dynamics . *J. Phys. Chem. B*, **2013**, *117* (44), 13731-42.

[154] Hölttä-Vuori, M.; Uronen, R.; Repakova, J.; Salonen, E.; Vattulainen, I.; Panula, P.; Li, Z.; Bittman, R.; Ikonen, E. BODIPY-cholesterol: a new tool to visualize sterol trafficking in living cells and organisms. *Traffic*, **2008**, *9* (11), 1839-1841.

[155] Kadengodlu, P. A.; Hebishima, T.; Takeshima, S.-N.; Ito, M.; Liu, M.; Abe, H.; Aida, Y.; Aigaki, T.; Ito, Y. Positively charged cholesterol-recombinant human gelatins foster the cellular uptake of proteins and murine immune reactions. *Int. J.Nanomedicine,* **2012**, *7*, 5437-5450.

[156] Watson, D. S.; Endsley, A. N.; Huang, L. Design considerations for liposomal vaccines: influence of formulation parameters on antibody and cell-mediated immune responses to liposome associated antigens. *Vaccine*, **2012,** *30* (13), 2256-2272.

[157] Kaur, R.; enriksen-Lacey, M.; Wilkhu, J.; Devitt, A.; Christensen, D.; Perrie, Y. Effect of incorporating cholesterol into DDA:TDB liposomal adjuvants on bilayer properties, biodistribution, and immune responses. *Mol. Pharmaceutics*, **2014**, *11* (1), 197-207.

[158] Drolle, E.; Kučerka, N.; Hoopes, M.I.; Choi, Y.; Katsaras, J.; Karttunen, M.; Leonenko, Z. Effect of melatonin and cholesterol on the structure of DOPC and DPPC membranes. *Biochimica et Biophysica Acta*, **2013,** *1828* (9), 2247-2254.

[159] (a) Di Paolo, G.; Kim, T.-W. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat. Rev. Neurosci*, **2011,** *12* (5), 284-96; (b) Friedman, R.; Pellarin, R.; Caflisch, A. Amyloid aggregation on lipid bilayers and its impact on membrane permeability. *J. Mol. Biol.*, **2009**, *387* (2), 407-415.

[160] Gellermann, G.P.; Appel, T.R.; Tannert, A.; Radestock, A.; Hortschansky, P.; Schroeckh, V.; Leisner, C.; Lütkepohl, T.; Shtrasburg, S.; Röcken, C.; Pras, M.; Linke, R.P.; Diekmann, S.; Fändrich, M. Raft lipids as common components of human extracellular amyloid fibrils. *Proc. Natl. Acad. Sci.U. S. A*., **2005**, *102* (18), 6297-6302.

[161] Puglielli, L.; Tanzi, R.E.; Kovacs, D.M. Alzheimer's disease: the cholesterol connection. *Nat. Neurosci*, **2003**, *6* (4), 345-351.

[162] Dante, S.; Hauss, T.; Dencher, N.A. Cholesterol inhibits the insertion of the Alzheimer's peptide Abeta(25-35) in lipid bilayers. *Eur. Biophys. J*, **2006**, *35* (6), 523-531.

[163] Drolle, E.; Gaikwad, R.M.; Leonenko, Z. Nanoscale electrostatic domains in cholesterol-laden lipid membranes create a target for amyloid binding. *Biophys. J*., **2012,** *103* (4), L27-9.

[164] Sheikh, K.; Giordani, C.; McManus, J.J.; Hovgaard, M.B.; Jarvis, S.P. Differing modes of interaction between monomeric Aβ(1-40) peptides and model lipid membranes: an AFM study. *Chem. Phys. Lipids*, **2012**, *165* (2), 142-150.

[165] Cecchi, C.; Nichino, D.; Zampagni, M.; Bernacchioni, C.; Evangelisti, E.; Pensalfini, A.; Liguri, G.; Gliozzi, A.; Stefani, M.; Relini, A. A protective role for lipid raft cholesterol against amyloid-induced membrane damage in human neuroblastoma cells. *Biochim. Biophys. Acta*, **2009**, *1788* (10), 2204-2216.

[166] (a) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput*., **2008**, *4* (3), 435-447; (b) Oostenbrink, C.; Soares, T.A.; van der Vegt, N.F.A.; van Gunsteren, W.F. Validation of the 53A6 GROMOS force field. *Eur. Biophys. J*., **2005**, *34* (4), 273-284.

[167] Wongekkabut, J.; Karttunen, M. Assessment of common simulation protocols for simulations of nanopores, membrane proteins, and channels. *J. Chem. Theory Comput*, **2012**, *8* (8), 2905-2911.

[168] Cino, E. A.; Choy, W.-Y.; Karttunen, M. M. Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J. Chem. Theory Comput*, **2012**, *8* (8), 2725-2740.

[169] Bachar, M.; Brunelle, P.; Tieleman, D.P.; Rauk, A. Molecular dynamics simulation of a polyunsaturated lipid bilayer susceptible to lipid peroxidation. *J. Phys. Chem. B*, **2004**, *108* (22), 7170-7179.**OPEN A**

[170] Martins do Canto, A. M. T.; Palace Carvalho, A. J.; Prates Ramalho, J. P.; Loura, L. M. S. Effect of amphipathic HIV fusion inhibitor peptides on POPC and POPC/cholesterol membrane properties: a molecular simulation study. *Int. J. Mol. Sci.*, **2013**, *14* (7), 14724-14743.

[171] Matthews, T.; Salgo, M.; Greenberg, M.; Chung, J.; DeMasi, R.; Bolognesi, D. Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nat. Rev. Drug. Discov*, . **2004**, *3* (3), 215-225.

[172] Eron, J.J.; Gulick, R.M.; Bartlett, J.A.; Merigan, T.; Arduino, R.; Kilby, J.M.; Yangco, B.; Diers, A.; Drobnes, C.; De Masi, R.; Greenberg, M.; Melby, T.; Raskino, C.; Rusnak, P..; Zhang, Y.; Spence, R.; Miralles, G.D.. Short-term safety and antiretroviral activity of T-1249, a second-generation fusion inhibitor of HIV. *Infect. Dis*., **2004**, *189* (6), 1075-1083.

[173] (a) Colman, P.M.; Lawrence, M.C. The structural biology of type I viral membrane fusion. *Nat. Rev. Mol. Cell Biol*., **2003**, *4* (4), 309-319; (b) Lawless, M.K.; Barney, S.; Guthrie, K.I.; Bucy, T.B.; Petteway, S.R., Jr.; Merutka, G. HIV-1 membrane fusion mechanism: structural studies of the interactions between biologically-active peptides from gp41. *Biochemistry*, **1996**, *35* (42), 13697-13708.

[174] Naider, F.; Anglister, J. Peptides in the treatment of AIDS. *Curr. Opin. Struct. Biol*., **2009**, *19* (4), 473-482.

[175] Chan, D.C.; Fass, D.; Berger, J.M.; Kim, P.S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell,* **1997**, *89* (2), 263-272.

[176] Wild, C.T.; Shugars, D.C.; Greenwell, T.K.; McDanal, C.B.; Matthews, T.J. Peptides corresponding to a predictive alpha-helical domain of human immunodeficiency virus type 1 gp41 are potent inhibitors of virus infection. *Proc. Natl. Acad. Sci. USA*, **1994**, *91* (21), 9770-9774.

[177] Liu, S.; Jing, W.; Cheung, B.; Lu, H.; Sun, J.; Yan, X.; Niu, J.; Farmar, J.; Wu, S.; Jiang, S. HIV gp41 C-terminal heptad repeat contains multifunctional domains. Relation to mechanisms of action of anti-HIV peptides. *J. Biol. Chem*., **2007**, *282* (13), 9612-20.

[178] Kilby, J.M.; Eron, J.J. Novel therapies based on mechanisms of HIV-1 cell entry. *N. Engl. J. Med.*, **2003**, *348* (22), 2228-38.

[179] (a) Baldwin, C.E.; Sanders, R.W.; Berkhout, B. Inhibiting HIV-1 entry with fusion inhibitors. *Curr. Med. Chem*., **2003**, *10* (17), 1633-42; (b) Dwyer, J.J.; Wilson, K.L.; Davison, D.K.; Freel, S.A.; Seedorff, J.E.; Wring, S.A.; Tvermoes, N.A.; Matthews, T.J.; Greenberg, M.L.; Delmedico, M.K. Design of helical, oligomeric HIV-1 fusion inhibitor peptides with potent activity against enfuvirtide-resistant virus. *Proc. Natl. Acad. Sci. U.S.A*., **2007**(31), *104*, 12772-7;

[180] Veiga, S.; Henriques, S.; Santos, N.C.; Castanho, M. Putative role of membranes in the HIV fusion inhibitor enfuvirtide mode of action at the molecular level. *Biochem. J*., **2004**, *377*, 107-10.

[181] Veiga, A.S.; Santos, N.C.; Loura, L.M.; Fedorov, A.; Castanho, M.A. HIV fusion inhibitor peptide T-1249 is able to insert or adsorb to lipidic bilayers. Putative correlation with improved efficiency. *J. Am. Chem. Soc*., **2004**, *126* (45), 14758-63.

[182] (a) Carvalho, A.J.P.; Ramalho, J.P.P.; Loura, L.M.S. Structure and conformation of HIV fusion inhibitor peptide T-1249 in presence of model membranes: A molecular dynamics study. *J. Mol. Struct. THEOCHEM*, **2010**, *946*, 119-124; (b) Martins do Canto, A.M.T.; Carvalho, A.J.; Ramalho, J.P.; Loura, L.M. Molecular dynamics simulations of T-20 HIV fusion inhibitor interacting with model membranes. *Biophys. Chem*, **2011**, *159* (2-3), 275-286. (c) Martins do Canto, A.M.T.; Palace Carvalho, A.J.; Prates Ramalho, J.P.; Loura, L.M. Molecular dynamics simulation of HIV fusion inhibitor T-1249: insights on peptide-lipid interaction. *Int. J. Mol. Sci*, **2012**, 151854.

[183] Franquelim, H. G.; Gaspar, D.; Veiga, A.S.; Santos, N.C.; Castanho, M.A. Decoding distinct membrane interactions of HIV-1 fusion inhibitors using a combined atomic force and fluorescence microscopy approach. *Biochim. Biophys. Acta*, **2013**, *1828* (8), 1777-85.

[184] Pan, C.; Liu, S.; Jiang, S. HIV-1 gp41 fusion intermediate: a target for HIV therapeutics *J. Formos. Med. Assoc*, **2010**, *109* (2), 94-105.

[185] (a) Alyautdin, R.; Khalin, I.; Nafeeza, M. I.; Haron, M. H.; Kuznetsov, D. Nanoscale drug delivery systems and the blood-brain barrier. *Int. J.Nanomedicine*, **2014**, *9*, 795-811; (b)Ozpolat, B.; Sood, A. K.; Lopez-Berestein, G. Liposomal siRNA nanocarriers for cancer therapy. *Adv. Drug. Deliv. Rev*, **2014**, *66C*, 110-116.

[186] (a) Manjila SB, Baby JN, Bijin EN, Constantine I, Pramod K, Valsalakumari, J. Novel gene delivery systems. *Int. J. Pharm.Investig*., **2013**, *3* (1), 1-7; (b) Onoue S, Yamada S, Chan HK. Nanodrugs: pharmacokinetics and safety. *Int, J. Nanomedicine*, **2014**, *9*, 1025-1037.

[187] Zhao, W.; Gurtovenko, A. A.; Vattulainen, I.; Karttunen, M. Cationic dimyristoylphosphatidylcholine and dioleoyloxytrimethylammonium propane lipid bilayers: atomistic insight for structure and dynamics. *J. Phys. Chem. B*, **2012**, *116* (1), 269-276.

[188] Lensink, M. F.; Lonez, C.; Ruysschaert, J.-M.; Vandenbranden, M. Characterization of the cationic DiC(14)-amidine bilayer by mixed DMPC/DiC(14)-amidine molecular dynamics simulations shows an interdigitated nonlamellar bilayer phase. *Langmuir*, **2009**, *25* (9), 5230-8.

[189] (a) Sun, C.; Tang, T.; Uludağ, H., Cuervo J.E. Molecular dynamics simulations of DNA/PEI complexes: effect of PEI branching and protonation state. *Biophys J*, .**2011**, *100* (11), 2754-63; (b) Sun, C.; Tang, T.; Uludağ, H. Molecular dynamics simulations of PEI mediated DNA aggregation. *Biomacromolecules*, **2011**, *12* (10), 3698-707.

[190] (a) Sun, C.; Tang, T.; Uludağ, H. Probing the effects of lipid substitution on polycation mediated DNA aggregation: a molecular dynamics simulations study. *Biomacromolecules* **2012**, *13* (9), 2982-8; (b) Sun, C.; Tang, T.; Uludag H. A. Molecular dynamics simulation study on the effect of lipid substitution on polyethylenimine mediated siRNA complexation. *Biomaterials*, **2013**, *34* (11), 2822-33.

[191] (a) Kopeć, W.; Telenius, J.; Khandelia, H. Molecular dynamics simulations of the interactions of medicinal plant extracts and drugs with lipid bilayer membranes. *FEBS J*., **2013**, *280* (12), 2785-805; (b) Witzke, S.; Duelund, L.; Kongsted, J.; Petersen, M.; Mouritsen, O. G.; Khandelia, H. Inclusion of terpenoid plant extracts in lipid bilayers investigated by molecular dynamics simulations . *J. Phys. Chem. B*, **2010**, *114* (48), 15825-31.

[192] Lindahl, E.; Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol*., **2008**, *18* (4), 425-431.

[193] (a) Harder, E.; MacKerell, A. D.; Roux, B.; Many-body polarization effects and the membrane dipole potential. *J. Am. Chem. Soc*., **2009**, *131* (8), 2760-1; (b) Khalili-Araghi, F.; Gumbart, J.; Wen, P. C.; Sotomayor, M.; Tajkhorshid, E.; Schulten, K. Molecular dynamics simulations of membrane channels and transporters. *Curr.Opin.Struct. Biol*, **2009**, *19* (2), 128-37. (c) Gumbart, J.; Wang, Y.; Aksimentiev, A.; Tajkhorshid, E.; Schulten, K. Molecular dynamics simulations of proteins in lipid bilayers. *Curr.Opin.Struct. Biol*., **2005**, *15* (4), 423-31.

[194] (a) Cordero-Morales, J. F.; Jogini, V.; Lewis, A.; Vasquez, V.; Cortes, D. M.; Roux, B.; Perozo, E. Molecular driving forces determining potassium channel slow inactivation. *Nat. Struct. Mol. Biol*., **2007**, *14* (9), 1062-9; (b) Domene, C.; Klein, M. L.; Branduardi, D.; Gervasio, F. K.; Parrinello, M. Conformational changes and gating at the selectivity filter of potassium channels. *J. Am. Chem. Soc*., **2008**, *130* (29), 9474-80; (c) Treptow, W.; Tarek, M. Environment of the gating charges in the Kv1.2 Shaker potassium channel. *Biophys. J*, .**2006**, *90* (9), L64-6.

[195] (a) Sands, Z. A.; Sansom, M. S. How does a voltage sensor interact with a lipid bilayer? Simulations of a potassium channel domain. *Structure*, **2007**, *15* (2), 235-44; (b) Jogini, V.; Roux, B. Dynamics of the Kv1.2 voltage-gated K+ channel in a membrane environment. *Biophys. J*, **2007**, *93* (9), 3070-3082.

[196] (a) Nishizawa, M.; Nishizawa, K. Molecular dynamics simulation of Kv channel voltage sensor helix in a lipid membrane with applied electric field. *Biophys.J*, **2008**, *95* (4), 1729-44; (b) Bond, P. J.; Sansom, M. Bilayer deformation by the Kv channel voltage sensor domain revealed by self-assembly simulations. *Proc. Natl. Acad. Sci. U.S.A*., **2007**, *104*, 2631-2636.

[197] Jensen, M.Ø; Jogini, V.; Borhani, D.W.; Leffler, A.E.; Dror, R.O.; Shaw, D.E. Mechanism of voltage gating in potassium channels. *Science*, **2012,** *336* (6078), 229-233.

[198] Ngo, V.; Stefanovski, D. ; Haas, S.; Farley, R.A. Non-equilibrium dynamics contribute to ion selectivity in the KcsA channel. *PLOs One*, **2014**, *9* (1), e86079.

[199] Weingarth, M.; van der Cruijsen, E. A. W. ; Ostmeyer, J. ; Lievestro, S.; Roux, B.; Baldus, M. Quantitative analysis of the water occupancy around the selectivity filter of a K+ channel in different gating modes. *J. Am. Chem. Soc*, **2014**, *136* (5), 2000-7.

[200] Rowley, C.N. ; Roux, B.A. Computational study of barium blockades in the KcsA potassium channel based on multi-ion potential of mean force calculations and free energy perturbation. *J Gen Phys*, **2013**, *42* (4), 451-463.

[201] Kasimova, M. A.; Tarek, M.; Shaytan, A. K. ; Shaitan, K. V. ; Delemotte, L. Voltage-gated ion channel modulation by lipids: Insights from molecular dynamics simulations. *Biochimica et Biophysica Acta*, **2014**, *1838*, 1322-31.

[202] Nygaard, R.; Zou, Y. ; Dror, R. O. ; Mildorf, T. J. ; Arlow, D. H. ; Manglik, A.; Pan, A. C. ; Liu, C. W.; Jose Fung, J. ; Bokoch, M. P. ; Sun Thian, F. ; Kobilka, T.S. ; Shaw, D. E. ; Mueller, L. ; Scott Prosser, R. ; Kobilka, B. K. The dynamic process of β(2)-adrenergic receptor activation. *Cell*, **2013**, *152*, 532-42.

[203] Schmidtke, P. ; Ciantar, M. ; Theret, I.; Ducrot, P. Dynamics of HERG closure allow novel insights into HERG blocking by small molecules. *J. Chem. Inf. Model*., **2014**, *54*, 2320-33.

[204] Gumbart, J.; Schulten, K. The Roles of Pore Ring and Plug in the SecY Protein-conducting Channel. *J. Gen. Physiol*, **2008**, *132*, 709-719.

[205] (a) Arkhipov, A.; Shan, Y. ; Das, R.; Endres, ; Eastwood, M. P. ; Wemmer, D.E. ; Kuriyan, J. ; Shaw, D. E. Architecture and membrane interactions of the EGF receptor. *Cell*, **2013**, *152*, 557-69; (b) Endres, N. F.; Das, R. ; Smith, A. W.; Arkhipov, A. ; Kovacs, E.; Huang, Y. ; Pelton, J. G. ; Shan, Y. ; Shaw, D. E. ; Wemmer, D. E. ; Groves, J. T. ; Kuriyan, J. Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell*, **2013**, *152*, 543-556.

[206] Arkhipov, A. ; Shan, Y. ; Kim, E. T. ; Shaw, D. E. Membrane interaction of bound ligands contributes to the negative binding cooperativity of the egf receptor. *PLoS Computational Biology*, **2014**, *10*, e1003742.

[207] Pan, A. C. ; Weinreich, T. M. ; Shan, Y. Scarpazza, D. P. ; Shaw, D. E. Assessing the accuracy of two enhanced sampling methods using egfr kinase transition pathways: The influence of collective variable choice. *Journal of Chemical Theory and Computation*, **2014**, *10*, 2860-5.

[208] Shan, Y.; Arkhipov, A.; Kim, E.T; Pan, AC; Shaw, D.E. Transitions to catalytically inactive conformations in EGFR kinase. *Proc Natl Acad Sci U S A*., **2013**, *110*, 7270-5.

[209] Ivetac, A.; Campbell, J. D.; Sansom, M. S.; Dynamics and Function in a Bacterial ABC Transporter: Simulation Studies of the BtuCDF System and its Components . *Biochemistry*, **2007**, *46*, 2767-78.

[210] Sonne, J.; Kandt, C.; Peters, G.H.; Hansen, F. Y.; Jansen, M. O.; Tieleman, D. P. Simulated nucleotide binding and release shows coupling between nucleotide binding domains and transmembrane domains in the ABC transporter BtuCD. *Biophys. J*., **2007**, *92*, 2727-34.

[211] Treptow, W.; Tarek, M. Molecular Restraints in the Permeation Pathway of Ion Channels. *Biophys. J*., **2006**, *90*, L64.

[212] Freites, J. A.; Tobias, D. J.; White, S. H. A voltage-sensor water pore. *Biophys. J*., **2006**, *91*, L90.

[213] Piasta, K.N.; Theobald, D.L.; Miller, C. Potassium-selective block of barium permeation through single KcsA channels. *J. Gen. Physiol*., **2011**, *138*, 421-36.

[214] Akitake, B.; Anishkin, A.; Liu, N.; Sukharev, S. Straightening and sequential buckling of the pore-lining helices define the gating cycle of MscS. *Nature Struct. Mol. Biol*, **2007**, *14*, 1141-9.

[215] Jeon, J.; Voth, G. A. Gating of the mechanosensitive channel protein MscL: the interplay of membrane and protein. *Biophys. J*., **2008**, *94*, 3497-511.

[216] Sawada, Y.; Murase, M.; Sokabe, M. The gating mechanism of the bacterial mechanosensitive channel MscL revealed by molecular dynamics simulations: from tension sensing to channel opening. *Channels (Austin)*, **2012,** *6*, 317-331.

[217] Cheng, X.; Ivanov, I.; Wang, H.; Sine, S. M.; McCammon, J. A. Nanosecond-timescale conformational dynamics of the human alpha7 nicotinic acetylcholine receptor., *Biophys. J.*, **2007**, *93*, 2622-34.

[218] Holyoake, J.; Sansom, M. S. P. Conformational change in an MFS protein: MD simulations of LacY. *Structure*, **2007**, *15*, 873-84.

[219] Nilsson, J.; Persson, B.; von Heijne, G. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins: Struct. Funct. Bioinf*, **2005**, *60*, 606-16.

[220] Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **2005**, *433*, 377-81.

[221] Johansson, A. C. V.; Lindahl, E. Amino-acid solvation structure in transmembrane helices from molecular dynamics simulations. *Biophys. J*, .**2006**, *91*, 4450-4463.

[222] MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Partitioning of amino acid side chains into lipid bilayers: results from computer simulation and comparison to experiment. *J. Gen. Physiol*, **2007**, *129*, 371-377.

[223] MacCallum, J. L.; Bennett, W. F. D.; Tieleman, D. P. Distribution of amino acids in a lipid bilayer from computer simulations. *Biophys. J.*, **2008**, *94*, 3393-3404.

[224] Yuzlenko, O.; Lazaridis, T. Interactions between Ionizable Amino Acid Side Chains at a Lipid Bilayer-Water Interface. *J Phys Chem B*, **2011,** *115*, 13674-84.

[225] Deol, S. S.; Domene, C.; Bond, P. J.; Sansom, M. S. P. Anionic phospholipid interactions with the potassium channel KcsA: simulation studies. *Biophys. J.*, **2006**, *90*, 822-830.

[226] Grossfield, A.; Feller, S. E.; Pitman, M. C. A role for direct interactions in the modulation of rhodopsin by omega-3 polyunsaturated lipids. *Proc. Natl. Acad. Sci. U.S.A.*, **2006**, *103*, 4888-4893.

[227] Grossfield, A.; Feller, S. E.; Pitman, M. C. Convergence of molecular dynamics simulations of membrane proteins. *Proteins: Struct.Funct.Bioinf.*, **2007**, *67*, 31-40.

[228] Hubbard, S.R.; Till, J.H. Protein tyrosine kinase structure and function. *Annu. Rev. Biochem.*, **2000**, *69*, 373-398.

[229] Robinson, D.R.; Wu, Y.M.; Lin, S.F. The protein tyrosine kinase family of the human genome. *Oncogene*, **2000**, *19*, 5548-57.

[230] Zwick, E.; Bange, J.; Ullrich A. Receptor tyrosine kinase signalling as a target for cancer intervention strategies. *Endocr. Relat. Cancer*, **2001**, *8* (3): 161–173.

[231] Riese, D.J. 2nd; Gallo, R.M.; Settleman, J. Mutational activation of ErbB family receptor tyrosine kinases: insights into mechanisms of signal transduction and tumorigenesis. *Bioessays*, **2007,** *29*, 558–565.

[232] Shaw, D.E.; Dror, R.O.; Salmon, J.K; Grossman, J.P.; Mackenzie, K.M.; Bank, J.A.; Young, C.; Deneroff, M.M.; Batson, B.; Bowers, K.J.; Chow, E.; Eastwood, M.P.; Ierardi, D.J.; Klepeis, J.L.; Kuskin, J.S.; Larson, R.H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M.A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, Portland, Oregon), **2009,** pp. 1–11.

[233] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror R.O.; Shaw D.E. Improved side-chain torsion potentials for the AMBER ff99SB protein force field, Proteins, **2010,** *78* (8) 1950–1958.

[234] Best, R.B.; Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B*, **2009**, *113* (26), 9004–9015.

[235] Nymeyer, H.; Woolf, T. B.; Garcia, A. E. Folding is not required for bilayer insertion: Replica exchange simulations of an alpha-helical peptide with an explicit lipid bilayers. *Proteins: Struct. Funct. Bioinf*., **2005**, *59*, 783-90.

[236] Vogel, A.; Roark, M.; Feller, S.E. A reinterpretation of neutron scattering experiments on a lipidated Ras peptide using replica exchange molecular dynamics. *Biochim Biophys Acta*, **2012,** *1818*(2), 219-24.

[237] Volkov, V.V.; Chelli, R.; Muniz-Miranda, F.; Righini, R. Structural properties of a membrane associated anchor dipeptide. *J Phys Chem B*, **2011**, *115*(18), 5294-5303.

[238] (a) Ikeda, K.; Kameda, T.; Harada, E.; Akutsu, H.; Fujiwara, T. Combined Use of Replica-Exchange Molecular Dynamics and Magic-Angle-Spinning Solid-State NMR Spectral Simulations for Determining the Structure and Orientation of Membrane-Bound Peptide. *J. Phys. Chem. B*, **2011,** 115, 9327–9336; (b) Yeh I. C., Olson, M.A.; Lee M.S.; Wallqvist A. Membrane insertion profiles of peptides probed by molecular dynamics simulations. *DoD HPCMP Users Group Conferenc*e, **2008**.

[239] Jaskierny, A.J.; Panahi A.; Feig M. Effect of flanking residues on the conformational sampling of the internal fusion peptide from Ebola virus. *Proteins*, **2011**, *79*(4), 1109-17.

[240] Vassall, K. A.; Bessonov, K.; De Avila, M.; Polverini, E.; Harauz, G. The Effects of Threonine Phosphorylation on the Stability and Dynamics of the Central Molecular Switch Region of 18.5-kDa Myelin Basic Protein. *PLoS ONE***, 2013**, *8*, e68175.

[241] (a) Harauz, G.; Musse, A. A. A tale of two citrullines-structural and functional aspects of myelin basic protein deimination in health and disease. *Neurochem. Res*., **2007**, *32*, 137-58; (b) Bessonov, K.; Harauz, G. Molecular dynamics investigation of myelin basic protein stability on lipid membranes. *SURG*, **2010**, 4, 179-186; (c) Kattnig, D.R.; Bund, T.; Boggs, J.M.; Harauz, G.; Hinderberger, D. Lateral self-assembly of 18.5-kDa myelin basic protein (MBP) charge component-C1 on membranes. *Biochimica et Biophysica Acta*, **2012,** *1818*, 2636-47.

[242] Kalli, A. C.; Devaney, I.; Sansom, M. S. P. Interactions of Phosphatase and Tensin Homologue (PTEN) Proteins with Phosphatidylinositol Phosphates: Insights from Molecular Dynamics Simulations of PTEN and Voltage Sensitive Phosphatase. *Biochemistry,* **2014**, *53*, 1724−32.

[243] Wang, Y.; Schlamadinger, D.E.; Kim, J. E.; McCammon, J. A. Comparative molecular dynamics simulations of the antimicrobial peptide CM15 in model lipid bilayers. *Biochimica et Biophysica Acta*, **2012**, *1818*, 1402–1409.

[244] Kandt, C.; Ash, W.L.; Tieleman, D. P. Setting up and running molecular dynamics simulations of membrane proteins. *Methods*, **2007**, *41*, 475-88.

[245] MacKerell, A.D. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem*., **2004**, *25*, 1584-604.

[246] Jorgensen, W.L.; Tirado-Rives, J. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc*., **1988**, *110*, 1657-1666.

[247] Ponder J.W.; Case D.A. Force fields for protein simulations. *Adv. Protein. Chem*., **2003**, *66*, 27-85.

[248] Oostenbrink, C.; Villa, A.; Mark, A.E.; Van Gunsteren, W.F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comp. Chem*., **2004**, *25*, 1656-76.

[249] Noskov, S.Y.; Lamoureux, G.; Roux, B. Molecular Dynamics Study of Hydration in Ethanol-Water Mixtures Using a Polarizable Force Field . *J. Phys. Chem. B*, **2005**, *109*, 6705-13.

[250] Chandrasekhar, I.; Kastenholz, M.; Lins, R.D.; Oostenbrink, C.; Schuler, L.D.; Tieleman, D.P.; van Gunsteren, W.F. A consistent potential energy parameter set for lipids: dipalmitoylphosphatidylcholine as a benchmark of the GROMOS96 45A3 force field. *European Biophysics Journal*, **2003**, *32*, 67-77.

[251] Tieleman, D.P.; MacCallum, J.L.; Ash, W.L.; Kandt, C.; Xu, Z.; Monticelli, L. Membrane protein simulations with a united-atom lipid and all-atom protein model: lipid-protein interactions, side chain transfer free energies and model proteins. *J. Phys.: Condens. Matter*, **2006**, *18*, S1221-34.

[252] Villa, A.; Mark, A.E. Calculation of the free energy of solvation for neutral analogs of amino acid side chains. *J. Comput. Chem*., **2002**, *23*, 548-553.

[253] Shirts, M.R.; Pitera, J.W.; Swope, W.C.; Pande, V.S. Extremely Precise Free Energy Calculations of Amino Acid Side Chain Analogs: Comparison of Common Molecular Mechanics Force Fields for Proteins. *J. Chem. Phys*., **2003**, *119*, 5740-5761.

[254] Maccallum, J. L.; Tieleman, D. P. Calculation of the water-cyclohexane transfer free energies of neutral amino acid side-chain analogs using the OPLS all-atom force field. *J. Comput. Chem*., **2003**, *24*, 1930-1935.

[255] Aliste, M.P.; Tieleman, D.P. Computer simulation of partitioning of ten pentapeptides Ace-WLXLL at the cyclohexane/water and phospholipid/water interfaces. *BMC Biochemistry*, **2005**, *6*, 30-31.

[256] Anezo, C.; de Vries, A.H.; Holtje, H.D.; Tieleman, D.P.; Marrink, S.J. Methodological issues in lipid bilayer simulations. *J. Phys. Chem. B*, **2003**, *107*, 9424-33.

[257] Faraldo-Gomez, J. D.; Smith, G. R.; Sansom, M. S.P. Setting up and optimization of membrane protein simulations. *Eur. Biophys. J.*, **2002**, *31*, 217-227.

[258] Biggins, P. C.; Bond, P. Molecular dynamics simulations of membrane proteins. *J.Methods Mol. Biol*., **2008**, *443*, 147-60.

[259] Henin, J.; Pohorille, A.; Chipot, C. Insights into the recognition and association of transmembrane alpha-helices. The free energy of alpha-helix dimerization in glycophorin A. *J. Am. Chem. Soc.*, **2005**, *127*, 8478-84.

[260] Tang, P.; Xu, Y. Large-scale molecular dynamics simulations of general anesthetic effects on the ion channel in the fully hydrated membrane: the implication of molecular mechanisms of general anesthesia. *Proc. Natl. Acad. Sci. U. S. A*., **2002**, *99*, 16035-40.

[261] Koradi, R.; Billeter, M.; Wuthrich, K. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **1996**, *14*, 51-55.

[262] (a) Hsin, J.; Arkhipov, A.; Yin, Y.; Stone, J.E.; Schulten, K. Using VMD: an introductory tutorial. *Curr Protoc Bioinformatics*, **2008**, Chapter 5:Unit 5.7; (b) Guixà-González, R.; Rodriguez-Espigares, I.; Ramírez-Anguita, J.M.; Carrió-Gaspar, P.; Martinez-Seara, H.; Giorgino, T.; Selent, J. MEMBPLUGIN: studying membrane complexity in VMD. *Bioinformatics,* **2014**, *30*(10), 1478-80.

[263] (a) Rees, D.C.; DeAntonio, L.; Eisenberg, D. Hydrophobic organization of membrane proteins. *Science*, **1989**, *245*, 510-3; (b)Wallin, E.; Tsukihara, T.; Yoshikawa, S.; von

Heijne, G.; Elofsson, A. Architecture of helix bundle membrane proteins: An analysis of cytochrome c oxidase from bovine mitochondria. *Protein Science*, **1997**, *6*, 808-15.

[264] Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev*, *42*, **2013**, 6801-6822.

[265] (a) Shih, A.Y.; Arkhipov, A, .; Freddolino, P.L.; Schulten, K. Coarse grained protein−lipid model with application to lipoprotein particles. *J Phys Chem B*, **2006**, *110*, 3674-84; (b) Shih, A.Y.; Arkhipov, A, .; Freddolino, P.L.; Schulten, K. Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. *J Struct Biol*, **2007**, *157*, 579-92.

[266] Arkhipov, A.; Yin, Y.; Schulten, K. Four-scale description of membrane sculpting by BAR domains. *Biophysical Journal*, **2008,** *95*, 2806-21.

[267] (a) Marrink, S.J.; de Vries, A.H.; Mark, A.E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B*, 2004, *108*, 750-60; (b) Marrink, S. J.; JelgerRisselada, H.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, **2007**, *111*, 7812-24.

[268] Martinetz, T.; Schulten, K. Topology representing networks. *Neural Networks*, **1994**, *7*, 507-22.

[269] Arkhipov, A., Freddolino, P.L.; Schulten, K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, **2006,** *14*, 1767-1777.

[270] Maftouni, N.; Amininasab, M.; Vali, M.; Ejtehadi, M.; Kowsari, F. A molecular dynamics simulation study of nanomechanical properties of asymmetric lipid bilayers. *J. Membrane Biol,* **2013**, *246*, 67-73.

[271] Marrink, S. J. ; Mark, A. E. Molecular dynamics simulation of the formation, structure, and dynamics of small phospholipid vesicles . *J. Am. Chem. Soc*., **2003**, *125*, 15233-42.

[272] Raudino, A.; Pannuzzo, M.; Karttunen, M. Combined depletion and electrostatic forces in polymer-induced membrane adhesion: a theoretical model. *J. Chem. Phys*., **2012**, *136*, 055101.

[273] (a) Baoukina, S. ; Tieleman, D. P. Direct simulation of protein-mediated vesicle fusion: lung surfactant protein B. *Biophys. J*., **2010**, *99*, 2134-42; (b) Risselada, H. J. ; Kutzner, C. ; Grubmüller, H. Caught in the act: visualization of SNARE-mediated fusion events in molecular detail. *ChemBioChem*, **2011**, *12*, 1049-55.

[274] Catte, A. ; Patterson, J. C. ; Bashtovyy, D.; Jones, M. K. ; Gu, F. ; Li, L.; Rampioni, A.; Sengupta, D.; Vuorela, T. ; Niemela, P. ; Karttunen, M. ; Marrink, S. J. ; Vattulainen I. ; Segrest, J. P. Structure of spheroidal hdl particles revealed by combined atomistic and coarse-grained simulations. *Biophys. J*., **2008**, *94*, 2306-19.

[275] Vuorela, T. A.; Catte, A. ; Niemela, P. S. ; Hall, A. ; Hyvonen, M. T.; Marrink, S. J.; Karttunen, M. ; Vattulainen, I. Role of lipids in spheroidal high density lipoproteins. *PLoS Comput. Biol*., **2010**, *6*, e1000964.

[276] Murtola, T. ; Vuorela, T. A. ; Hyvonen, M. T. ; Marrink, S. J. ; Karttunen, M.; Vattulainen, I. Low density lipoprotein: structure, dynamics, and interactions of apoB-100 with lipids. *Soft Matter*, **2011**, *7*, 8135-41.

[277] Baoukina, S. ; Monticelli, L. ; Marrink, S. J. ; Tieleman, D. P. Pressure-area isotherm of a lipid monolayer from molecular dynamics simulations. *Langmuir*, **2007**, *23*, 12617-23.

[278] Baoukina, S. ; Monticelli, L. ; Risselada, H. J. ; Marrink, S. J. ; Tieleman, D. P. The molecular mechanism of lipid monolayer collapse. *Proc. Natl. Acad. Sci. U. S. A*., **2008**, *105*, 10803-8.

[279] Hu, M.; de Jong, D.H. ; Marrink, S.J. ; Deserno, M. Gaussian curvature elasticity determined from global shape transformations and local stress distributions: A comparative study using the MARTINI model. *Farad. Discuss*., **2013**, *161*, 365-82; (b) Yesylevskyy, S. O. ; Demchenko, A. P. ; Kraszewski, S. ; Ramseyer, C. Cholesterol Induces Uneven Curvature of Asymmetric Lipid Bilayers. *The Scientific World Journal*, **2013**, ID 965230, 10; (c) Fuhrmans, M.; Marrink, S.J. Molecular view of the role of fusion peptides in promoting positive membrane curvature. *J. Am. Chem. Soc*., **2012**, *134*, 1543-52.

[280] Wassenaar, T. A.; Ingólfsson, H. I.; Prieß, M.; Marrink, S. J.; Schäfer, L. V. Mixing MARTINI: Electrostatic Coupling in Hybrid Atomistic–Coarse-Grained Biomolecular Simulation. *J. Phys. Chem. B*, **2013**, *117*, 3516-30.

[281] Shi, Q.; Izvekov, S.; Voth, G. A. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane-bound ion channel. *J. Phys. Chem. B*, **2006**, *110*, 15045-8.

[282] Michel, J.; Orsi, M.; Essex, J. W. Prediction of partition coefficients by multiscale hybrid atomic-level/coarse-grain simulations. *J. Phys. Chem. B,* **2008**, *112*, 657-60; (b) Orsi, M.; Sanderson, W. E.; Essex, J. W. Permeability of Small Molecules through a Lipid Bilayer: A Multiscale Simulation Study. *J. Phys. Chem. B*, **2009**, *113*, 12019-29.

CHAPTER 9

Data Quality Assurance and Statistical Analysis of High Throughput Screenings for Drug Discovery

Yang Zhong, Zuojun Guo and Jianwei Che*

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA

Abstract: High throughput screening (HTS) is an important tool in modern drug discovery processes. Many recent, successful drugs can be traced back to HTS [1]. This platform has proliferated from pharmaceutical industry to national labs (*e.g.* NIH Molecular Libraries Screening Centers Network), and to academic institutions. Besides throughput improvements from thousand molecules in early times to multimillion molecules now, it has been adapted to increasingly sophisticated biological assays such as high content imaging. The vast amount of biological data from these screens presents a significant challenge for identifying interesting molecules in various biological processes. Due to the intrinsic noise of HTS and complex biological processes in most assays, HTS results need careful analysis to identify reliable hit molecules. Various data normalization and analysis algorithms have been developed by different groups over the years. In this chapter, we briefly describe some common issues encountered in HTS and related analysis.

Keywords: Bayesian model, dose-response analysis, High-throughput Screening, hit analysis, hit identification, normalization, ontology-based pattern identification (OPI) method, quality control, strictly standardized mean difference (SSMD) metric, *t*-test.

1. INTRODUCTION

High throughput screening (HTS) is an automated experimental platform for rapidly identifying small number of molecular entities or conditions with unique biological properties from a large number of tests. Due to its high efficiency, it has been widely used in early stage drug discovery to identify candidate molecules for optimization. In a large scale HTS campaign in drug discovery,

***Corresponding author Jianwei Che:** Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, USA; E-mail: jche@gnf.org

Zaheer Ul-Haq and Jeffry D. Madura (Eds)

several millions of small molecules or tens of thousands antibodies are tested *in vitro* for desirable biological activities [2, 3]. Similarly, genome-wide screenings of genes and proteins are also conducted to determine their biological functions and interactions [4]. More recently, screenings of antibody libraries have been performed to select antibodies with desired physical properties such as thermostability, colloidal stability and solubility in monoclonal antibody (mAb) formulation [5, 6], or specific biological functions [7, 8].

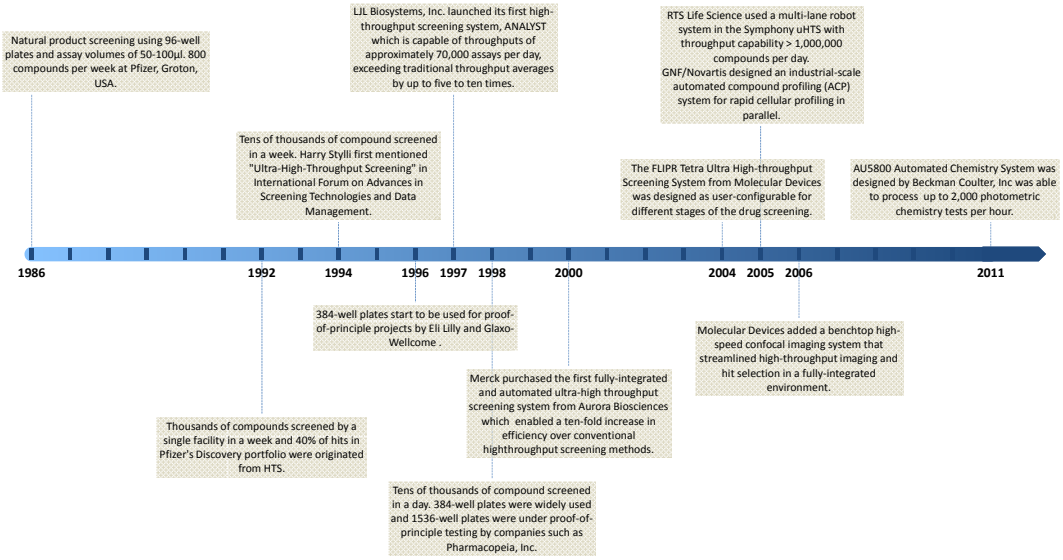


Figure 1: Major milestones in high-throughput screening development.

Technological innovations and scientific advances have played essential roles in the development of HTS. In early 1980's, typical screens were conducted manually on several hundred samples per week. For example, Pfizer screened 800 natural products per week from fermentation tubes in 1986 [9]. Since then, the increasing capacity of HTS platform showed its potential to rapidly sift through enormous number of compounds, antibodies and genes, therefore, promised efficient starting points of discovering new drugs. This has led to a paradigm shift in HTS technology (Fig. **1**). From the late 1980s to early 1990s, drug discovery companies have embraced automation and miniaturization techniques for

developing new assays on high-density plates. These technologies significantly advanced HTS, lowered the screening costs, and improved reproducibility [10]. During the 2000s, HTS further evolved to "ultra high-throughput screening" (uHTS), with throughput over 100,000 samples per day [11, 12] and ten-fold reduction in screening cost when comparing to the mid-1990s [13]. Fig. **2** summarizes the evolution of miniaturization during this period along with the cost of HTS assays. Currently, the screening throughput has reached multimillion compounds per week in large pharmaceutical and biotech companies. Many types of biological targets have been investigated over the years under HTS platform. The emergence of high content screening (HCS) in cell based assays, another HTS advance, introduces simultaneous readout of multi-parameters per sample. It yields much richer information on various phenotypes of cells than traditional single measurement screens [14, 15] such as fluorescent intensity.

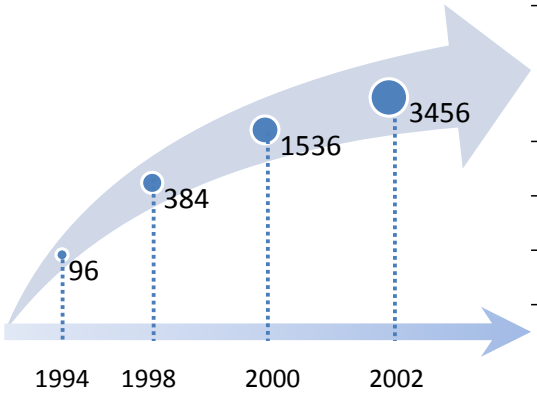


Plate Type	Assay Volume (µL)	Assay Cost per Well ($)
96-well	100-200	1.0-2.0
384-well	20-50	0.20-0.5-
1536-well	2-10	0.05-0.10
3456-well	0.001-0.002	<0.01

Figure 2: Progress on miniaturization in HTS. Figure on the left shows the evolution of plate types used in HTS experiments. Plate types are labeled by the number of wells per plate. Table on the right compares the assay volume and cost per well for each type (From Refs. [13, 16-19]).

HTS technology continuously enhances its position in biomedical research over time. According to the analysis by Swinney *et al.* [20], majority of approved drugs (new molecular entities - NMEs) between 1999 to 2008 originated from screening approaches (Fig. **3**). 28 out of 75 (37%) first-in-class drugs originated from phenotypic screening – the most productive technology for early stage drug discovery, and 17 NMEs (23%) from target-based screens. For follow-on drugs,

the most productive strategy is target-based screens with more than half of the drug approvals (83 out of 164 NMEs), and 30 NMEs (18%) originated from phenotypic screens. Macarron *et al.* [1] studied HTS contributions to different stages of drug discovery projects in 2009 from four pharmaceutical companies. They found 48% to 84% of chemical optimization efforts were initiated from in-house HTS. The authors also listed 12 drugs approved from 2003 to 2009 that originated from HTS hits.

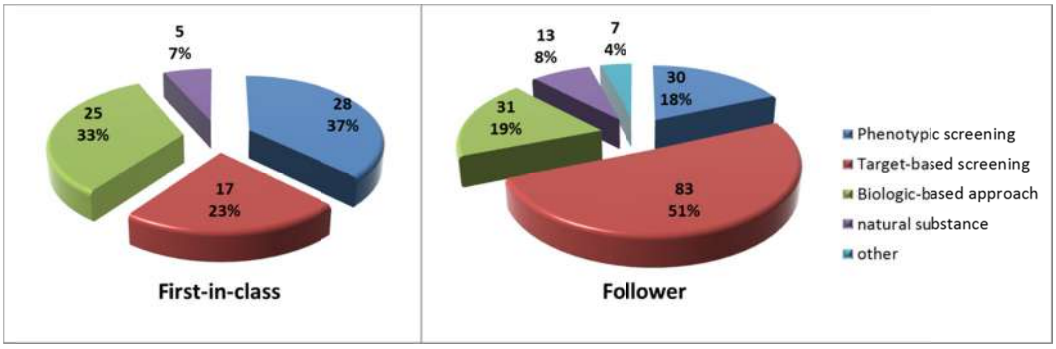


Figure 3: The origin of small molecule NMEs approved during 1999–2008 (Data based on reference [20]).

At the early stage of modern drug discovery projects at large pharmaceutical companies, HTS is routinely performed to profile millions compounds in a single screening campaign [21]. Besides small molecule screens, genome wide siRNAs were also reported [4]. As illustrated in Fig. **4**, HTS serves as an efficient tool to identify promising structures from a large number of compounds for further medicinal chemistry optimization. The primary screening of several million compounds usually leads to thousands hits [22, 23]. These hit molecules are typically tested in secondary screens to confirm their activities. The confirmatory screens normally are done in dose response fashion and include both functional and specificity assays. This process eliminates toxic and false positive compounds, and often reduces the hit list to a few hundred molecules [24, 25].

Proper HTS data analysis is essential for identifying high quality structures for the downstream drug discovery processes, as subsequent chemical optimization and biological evaluation are costly and time consuming. In most cases, simply choosing molecules with highest positive signals from HTS is insufficient and

sometime even misleading [26], because of intrinsic properties of chemical library and assay technology. For example, compounds with certain physicochemical properties can interfere with assay detection mechanism such as fluorescence [27, 28]. In addition, HTS signal is typically lower than conventional assays due to miniaturization. Moreover, confounding factors such as edge evaporation and batch effects can lead to artifacts that render the straightforward interpretation of raw signal meaningless [29].

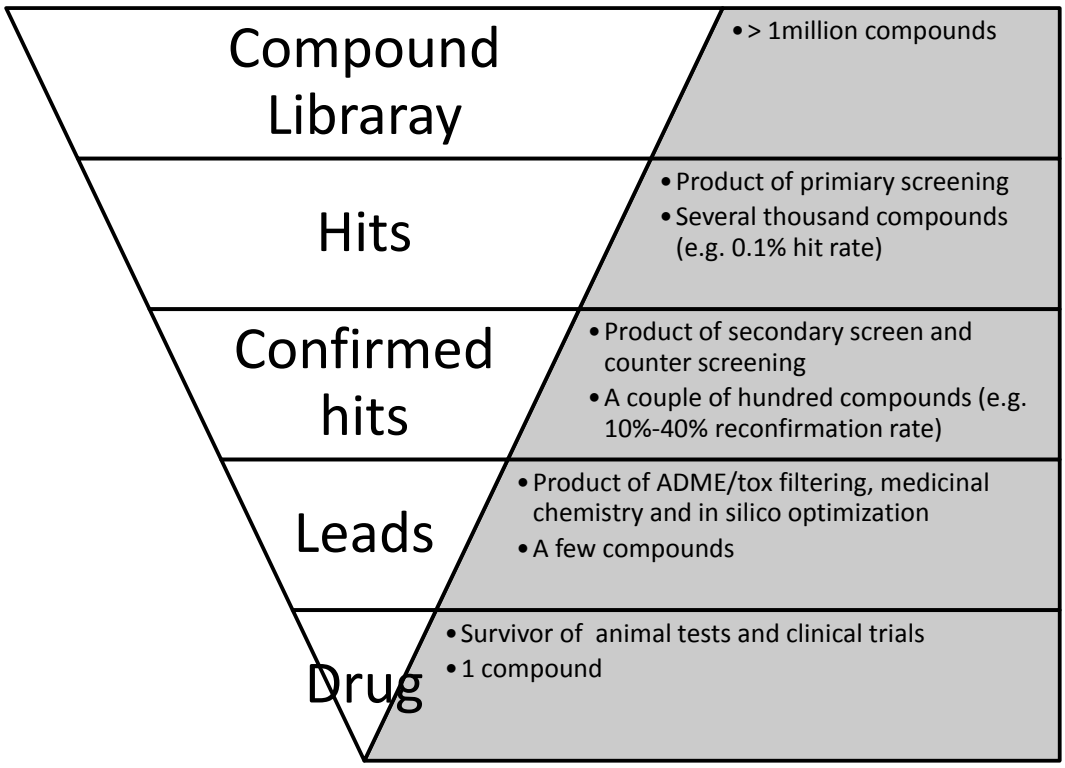


Figure 4: General early drug discovery process.

While challenged by large scale data and assay miniaturization of industrial screening operations, current HTS data management benefits from the advancements of computational technology. According to a recent report, the computational costs have declined over 30% annually during the past two decades [30]. The advances in computer hardware not only accelerated the evolution of HTS automation, but also helped to address the challenges from data acquisition,

storage and analysis of large data sets [31]. As of September 2013, more than 200 million bioactivity summaries and 1.2 trillion data points have been deposited into NCBI PubChem BioAssay database for 2.8 million compounds, 1.9 million chemical structures and 108 thousand RNAi reagents [32]. Such databases provide valuable resources for mining and analyzing hits from wide range of assays [26, 27, 33, 34]. Many analytical tools and software have been developed to analyze high volume screening data [35, 36] including commercial software such as IDBS Activitybase, Oxford RS3 and MDL Assay Explorer, and tools like Spotfire [37] for visual inspection and exploration.

From data analysis perspective, a future direction is to integrate primary screening data with results from subsequent confirmation and orthogonal assays to expand identified hits (see Section 4 for more details). In addition, the integration of early HTS data, even "inactive" data, with data from later stages and *in silico* analysis might also benefit the hit characterization and prioritization. Another trend of HTS analysis lies on the data management for high content screening (HCS). Although a few generic data analysis packages such as CellProfiler [38] and PerkinElmer Acapella are commonly used, it is still far from satisfactory in terms of automatic HCS quality control and hit identification due to the complexity of imaging data.

In this chapter we attempt to provide an overview of data processing and hit identification methods for HTS in drug discovery process. We first summarize the typical HTS process and introduce the general workflow of HTS analysis. Subsequently, we review both general and specific approaches on data normalization, quality control, hit picking, and hit analysis. We also compare the advantages and limitations of different hit identification strategies reported in literature. Lastly, we discuss the follow-up analysis on the biological properties of hit molecules, particularly from recent data mining efforts on HTS databases.

2. HIGH THROUGHPUT SCREENING PROCESS

Regardless of assay types and targets, a typical HTS protocol can be divided into three major steps (Fig. **5**).

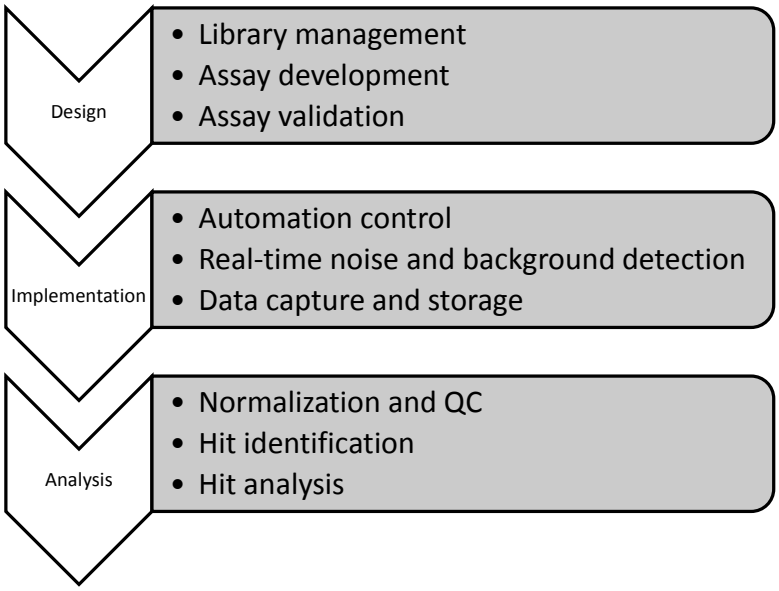


Figure 5: A typical HTS process.

A good HTS design consists of a biologically relevant and robust assay, high-quality chemical library, and sufficient throughput. Advances in human genomics introduce more and more potential targets and the number is expected to increase in the future [1]. To identify leads for a selected target, particularly novel targets without much prior knowledge, the size and diversity of compound collections are crucial for the success of HTS. In the early days of HTS, compound libraries were mainly derived from natural products [39] or internally assembled [1]. Modern compound libraries utilize the advances in organic and combinatorial chemistry to increase their size and diversity. As a result, millions of compounds were routinely screened in the early 2000s [39]. Meanwhile, the advent of cheminformatics and data mining methodologies enable characterizations of individual compounds with great details. Diversity of a compound library is determined in terms of pharmacophoric motifs, chemical groups, target-based discriminations, *etc.* [40-42]. A number of software (*e.g.*, ChemAxon, Accelrys, Tripos, Schrodinger, OpenEye, and Oxford Molecular) has been developed for multi-dimensional characterization of libraries based on *in silico* descriptions [43]. Assay development, on the other hand, aims at maximizing the dynamic range between active and inactive compounds in miniature format. Both biochemical and cellular assays are routinely employed in drug discovery projects.

Biochemical assays have the advantage of clear determination of target engagement, while cellular assays have better representation of biological context. In recent years, cell-based assays are taking a more prominent role [44] due to its tighter biological relevance and the ability for phenotypic/target agnostic readouts. Conditions of HTS assays are often optimized to achieve both sensitivity and stability in the biological system of interest, followed by statistical evaluations [45].

Laboratory automation is a major driving force behind HTS implementation. Together with microplate design, liquid handling, and detection technology, it greatly reduces screening time and improves data quality and consistency. Moreover, modern uHTS systems are able to monitor data quality on the fly to ensure early detection and prevention of problems through preventive maintenance, equipment self-check and validation [46]. From plate loading to reagent dispensing to error detection and data storage, fully integrated automation is able to perform assays and deliver data free of any human intervention.

While data analysis is also involved in the step of library management and the implementation of statistical process control, we focus our discussions here on methods of analyzing data from screening experiments.

2.1. HTS Analysis Overview and Workflow

Usually, raw data from screening systems go through multiple steps before it can be reliably interpreted biologically. The main analysis process can be divided into four steps, as illustrated in Fig. **6**.

In an industry setting, HTS data usually is integrated with a centralized compound management database in order to work efficiently with millions of compounds. HTS is intrinsically an error-prone process due to factors such as miniaturization, stochastic error, batch effects, and edge effects, the raw signals have to be properly investigated and corrected before extracting the meaningful information. Therefore, quality control (QC) is a critical step prior to hit selection processes. In Section 2, we describe various stages of QC and corresponding statistical analysis. The goal of HTS is to identify a few “active” hits from large number of biologically inactive samples. Based on different statistical principles, many approaches have been applied to select and/or prioritize compounds or genes.

They will be discussed in detail in Section 3. The follow-up analysis is described in Section 4. It characterizes selected hits, summarizes their chemical and biological properties that can be integrated into a comprehensive report for decision making.

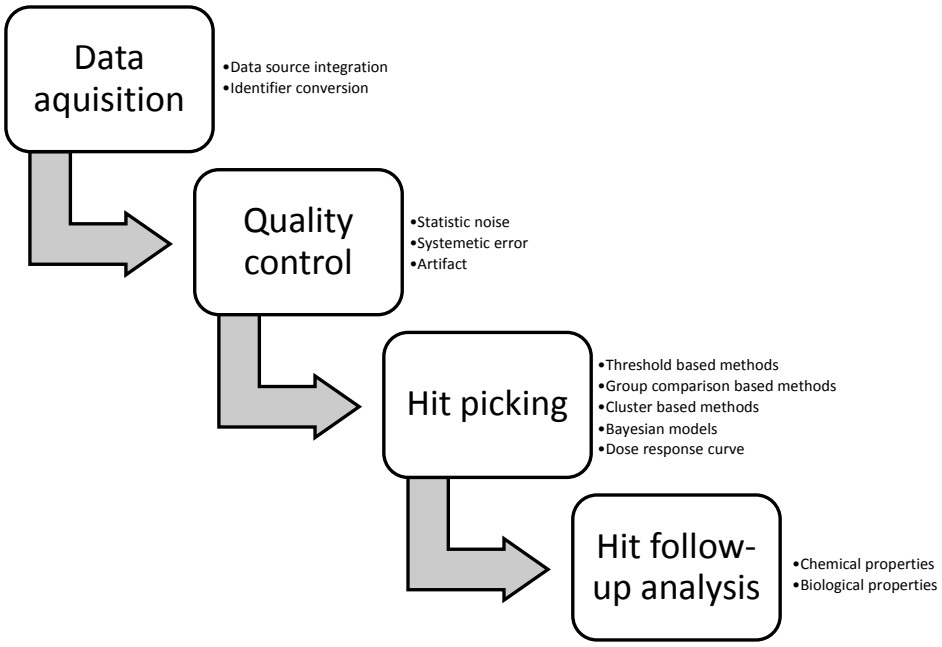


Figure 6: HTS data analysis workflow.

2.2. Quality Control (QC) in HTS Results

Like in any experiments, QC is an extremely important initial step. It is particularly true for HTS. The state-of-the-art "industrial" HTS data production has surpassed 20-fold of the initial throughput in the 1990s, and over a million data points can be obtained in a day in a uHTS campaign [47]. The quantity and complexity of modern HTS results present a challenge for data processing. Traditional manual QC based on human visualization is no longer practical under this circumstance. Thus, large efforts have been devoted to automated QC process for error detection and correction from pharmaceutical companies [37, 48-51], software companies [52-55] and academic labs [56-58]. Here, we review the fundamental statistical principles used in HTS QC at different stages.

2.2.1. Normalization

Normalization is a critical step to remove inter and intra plate bias and systematic errors in raw signal. The raw data is usually normalized based on plate-to-plate and/or batch-to-batch variations.

2.2.1.1. Control Versus Non-Control

Although improved by automation, noise and variation is still inevitable in measurements and can be affected by factors such as well location, liquid dispensing and signal intensity. For example, well along the plate edge has higher rate of evaporation. To help assess background signal level, control compounds are usually included in assays. In typical biochemical screens for antagonists, negative control is usually assay vehicle such as dimethyl sulphoxide (DMSO) solution without any compound. It usually corresponds to the high activity *H*. In addition, positive controls such as potent reference compounds are also included to give benchmark activity level, *i.e.* the low activity *L*. On the other hand, the low activity level *L* in agonist assays is the basal level signal and the high activity *H* is determined by positive controls, *i.e.* reference agonist molecules. In control-based normalization, the normalized activities are calculated as the ratio with respect to inhibited or activated and basal level signals (Equation 1).

Equation 1

$$p_i = \frac{x_i - L}{H - L}$$

where x_i and p_i are the signals pre and post normalization for well i, respectively. However, outliers in the control wells can introduce biases in controls-based methods [59, 60].

Sample-based normalization assumes most samples from a diverse library are inactive. Therefore, the background level can be estimated by the average signal of all wells on a plate (Equation 2). Obviously, it may not be true if a plate contains many active samples such as a focused library of a specific target or target class [48, 49].

Equation 2

$$p_i = \frac{x_i - \bar{x}}{H - \bar{x}}$$

where $\bar{x}$ can be either the mean(x), median(x) or mode(x) function (usually the value with highest probability density). The median is often used because the mean is sensitive to outliers and the mode is lack of consistent definition [61].

In HTS, non-control based normalization commonly uses *Z*-score (Equation 3) and B-score (Equation 4) [62]. Z score computes the deviation of individual well signal from the mean in the unit of the standard deviation. B score is designed to correct systematic biases along rows and columns. Both methods are widely used in HTS hit identification process [59, 63].

Equation 3

$$z_i = \frac{x_i - \bar{x}}{s}$$

where z_i is the *Z*-score of well *i*, and *s* is the standard deviation or median absolute deviation (MAD) of $\{x_i\}$ in the same plate.

Equation 4

$$B_{ij} = \frac{r_{ij}}{MAD}$$

where B_{ij} is the B-score of well at row *i* and column j. $r_{ij} = x_{ij} - (\bar{x} + \bar{R}_\iota + \bar{C}_\jmath)$ is the residual of raw signal x_{ij} and a fitted value based on the mean of plate signals $\bar{x}$, systematic signal offsets of row *i* $\bar{R}_\iota$ and column *j* $\bar{C}_\jmath$ [64]. MAD is the median absolute deviation of r_{ij} within the plate.

A Previous study on normalization methods suggested that no single method was versatile for all HTS data sets due to the complexity of different screening campaigns [65, 66]. Shun *et al.* [66] employed five normalization methods on six HTS campaigns and proposed a 3-step statistical decision methodology to select the optimal QC method based on statistical analysis of assay signals. In some

cases, the choice of normalization method showed little effect on the final results. For instance, seven normalization approaches were applied in a large-scale Drosophila RNAi screen, and no significant performance difference was found based on confirmation rate from a secondary screening [67].

2.2.1.2. By Plate Versus by Experiment

Plate-wise normalization is commonly done to correct plate-specific systematic bias, in particular, when hit rates are low. However, screens of non-random samples, such as focused compound libraries or a transcription factor siRNA library, may have large number of true positives on a single plate. Normalizing within plate under these scenarios can underestimate the signals from true hits. It is not unusual for similar compounds to be on a same plate due to library generation, which can contribute to the appearance of clustered hits. Without strong evidence of intra-plate systematic bias, normalization across all plates in an assay is widely used [68].

2.2.1.3. Statistical Noise Versus Systematic Error

Statistical noises from intrinsic fluctuations in raw data can occur at any given time during screening. Conceptually, random noises can be minimized by increasing sample size [69]. Systematic errors, on the other hand, are time dependent and can be attributed to certain known cause(s), such as failures of robotic systems, malfunctions of pipettes, variation in incubation time, temperature, lighting, air flow, decay of cells, and evaporation of reagents [70]. Systematic errors usually result in large scale signal patterns. For example, certain area of a plate consistently gives high signal across an experiment or within consecutive plates (Fig. **7**). It is a common error and usually challenging to correct automatically. False positives and false negatives are often the results of these errors [70, 71].

One way to correct systematic error is to combine normalization with pattern recognition. For instance, plates sharing a similar signal pattern and/or from a consecutive time interval are grouped, and normalization is applied first within the groups then on individual plates. However, such process usually relies on

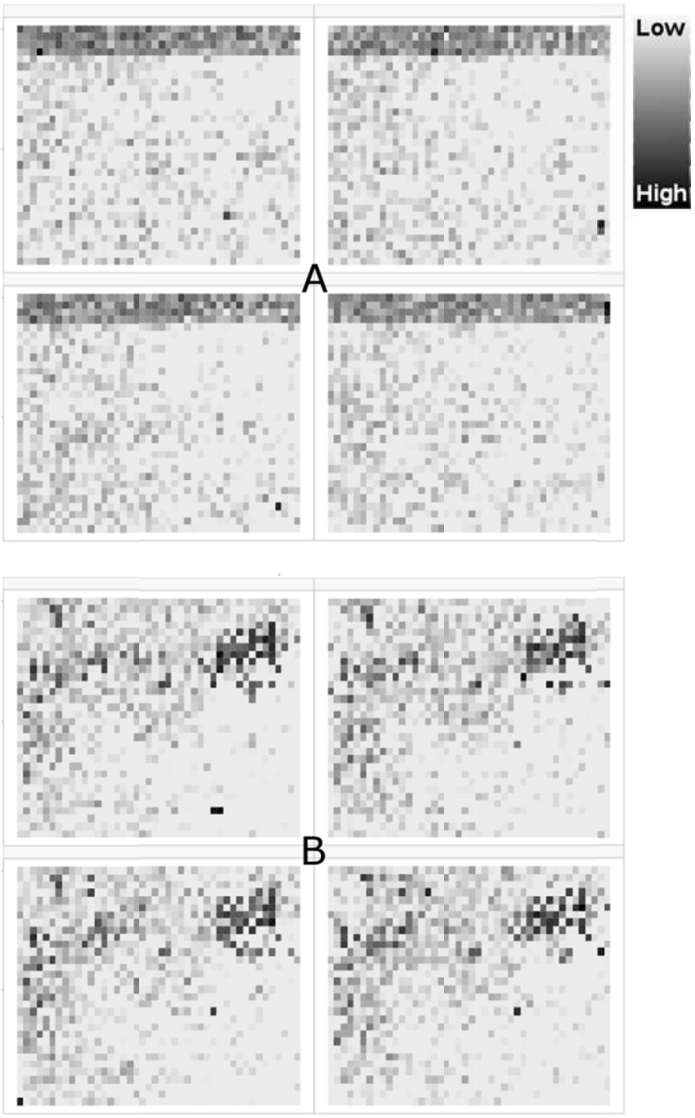


Figure 7: Two examples of systematic errors: edge effects on the top rows for four plates in panel A and hit clusters on the upper right side of four plates in panel B.

intensive manual examination of spatial or time-dependent patterns by visualization. This process can lead to unreliable and inconsistent error correction. To improve the process, Gubler *et al.* [61] grouped plates from a time window and changed point detection according to their time course: plates screened consecutively should share similar conditions. Similarity-based method [57]

employed unsupervised *k*-means clustering method to group plates based on their signal similarities. One limitation of these methods is the uncertainty of time frame size or *k* value. Scientists may still need visual examination to properly choose these parameters. The "well correction" (WC) method [72] corrects each well signal by centering the distribution of the same well of all plates at zero using the linear least-squares fit. Despite its robustness and overall good performance, WC method cannot correct errors affecting only a subset of plates. Scoring methods, such as B-score [59, 69], estimate background signal using a two-way median polish method [64]. It corrects signal shifts in both rows and columns (Fig. **7A**). However, B-score method cannot correct irregular error patterns, such as the hit clusters in Fig. **7B**. It can lead to over-correction of signals [73].

2.2.2. Removal of Random Artifacts

For HTS screens on diverse libraries, concentrated active wells are statistically abnormal. As shown in Fig. **8A**, there is only 1% probability of finding a cluster of hits with more than five wells in a 1536-well plate based on 2000 permutation tests with a 10% hit rate. Therefore, continuous (or nearly continuous) region of actives beyond certain size is more likely to be artifacts than true actives in a normal HTS, unless they are screens with known high hit rate (for example, focused library). Fig. **7B** displays four plates subject to artifacts with unique and random shapes. Such artificial blocks are usually caused by dispensing, reagents or sample preparations [48], and no longer retain the information of true signals. They need to be removed from HTS hit analysis prior to normalization [71, 72].

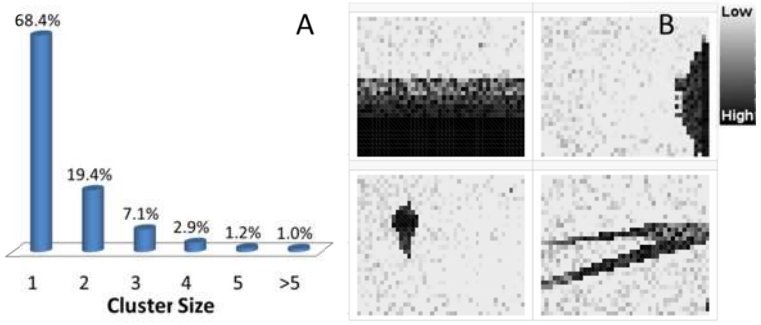


Figure 8: Random artifacts in HTS results. (A) Percentages of blotches of different sizes randomly formed by simulating 2000 hypothetical 1536-well plates with 10% hit rate. (B) Examples of artifacts shown as unique patterns in four plates.

2.2.3. Quality Evaluation

Before hit selection, the last step of QC assesses if the quality of corrected data meets the minimum requirements for hit selections. The assay quality is usually characterized by statistical parameters based on dynamic range and variance of the measurements. Such parameters are also widely used for assay-to-assay comparison in assay development and optimization.

2.2.3.1. Z or Z' Factor

Two widely-used quality metrics proposed by Zhang *et al.* are Z and Z' factor [74].

Equation 5

$$Z = 1 - \frac{3\sigma_s + 3\sigma_c}{|\mu_s - \mu_c|}$$

$$Z' = 1 - \frac{3\sigma_{Hc} + 3\sigma_{Lc}}{|\mu_{Hc} - \mu_{Lc}|}$$

where μ and σ indicate mean and standard deviation, s and c represent sample and control, and *Hc* and *Lc* are the upper and lower bounds of control activities, respectively. In general, assays with $Z < 0$ are considered to have insufficient dynamic range because sample and control signal are overlapping. With $Z > 0.5$, a good separation of signal from noise can be achieved. Z factor measures assay quality based on signal levels from sample and control, and is used in QC process. Z' factor, on the other hand, represents the resolution of controls in an assay and is optimized during assay development.

2.2.3.2. Strictly Standardized Mean Difference (SSMD)

For screens using moderate controls (for example, siRNA screens), SSMD parameter is considered to be statistically more appropriate than Z factor [75].

Equation 6

$$SSMD = \frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}$$

SSMD measures the difference between two groups, A and B, according to their means μ and standard deviations σ, where A and B can be samples or controls. It is analogous to Z factor for actual sample quality assessment, and Z' factor for assay optimization with controls. SSMD is found to be more accurate and less conservative than Z and Z' methods in siRNA screenings [75, 76].

3. HIT IDENTIFICATION STRATEGIES

Following data QC, the data is analyzed to identify hit compounds or genes with good activities and properties for confirmation and optimization in medicinal chemistry and/or *in silico*. For a typical HTS campaign, the primary hit picking can be conceptually simplified to a statistical question: how to separate positives from negatives with practical constraints such as time and cost. In a primary screen, sample activities are usually measured at a single concentration for high throughput. Despite the intrinsic limitations of single-dose HTS [77], several statistical tools were developed to optimize sensitivity and specificity for hit finding. Usually, hits are chosen from the top ranked samples, where actual number of hits is limited by the capacity of follow-up experiments or a threshold defined by relevant biological rationales [78]. However, the threshold is largely ad hoc in many cases [23, 25, 79, 80]. Table **1** summarizes the commonly used methods for HTS hit identification and the details of individual methods are described in Section 3.1 to 3.5.

Table 1: Summary of HTS hit identification methods

Category	Application	Statistical Model	Notes
Threshold-based	Single-dose screens without replicate	Percent of control or normalized percent of control (Equation 1)	Lacking robustness.
		Z-scores (Equation 3)	Based on constant error assumption. Standard *Z*-score is not robust; *Z*-score based on median and MAD preferred.
		B-score (Equation 4)	Robust and more resistant to outliers than *Z*-score methods
Group-comparison-based	Single-dose screens with replicates	Paired *t*-test or modified *t*-test (randomized variance model) (Equation 7)	Low sensitivity if sample size is limited.
		SSMD-based methods (Equation 8)	Robust for small sample size. Comparable across experiments.

Table 1: contd…

Cluster-based	Small-molecule screens with built-in chemical redundancy or redundant RNAi screens or high content screens	Hypergeometric model (Equation 9)	Robust, high confirmation rate than threshold methods.
Bayesian model	Screens without replicates, noisy screens (*e.g.* compound mixture screens)	Bayesian models	Robust, noise tolerant, better performance than Median +/- *k* MAD. Able to combine plate-wise and experimental-wise information.
Dose response	Secondary confirmatory screens for small molecules	Sigmoidal curve, Hill's Equation (Equation 10)	Requires curve fitting process.

3.1. Threshold Based Methods

Threshold based methods rank compounds or genes directly by their normalized activities. It is very simple conceptually and computationally. The thresholds are usually chosen according to the breakdown point [81] theory in statistics. The breakdown point is considered as the threshold to determine outliers (hits in HTS analysis) without drastically changing the original distribution of the sample set. For example, the breakdown point of a normal distribution can be estimated based on normality plots. Studies have shown that the approach is preferred for data sets with clear outliers, in particular, when robust statistics still applies with median absolute deviation (MAD) even when half of the data set are outliers [82].

Table 2: Comparison of different threshold based methods

Strategy	Threshold	Application	Notes
Top X	Activity of the X^{th} sample	Any screen	No statistical basis.
Mean +/- *k* SD*	*k* standard deviations (SD) away from the average of normalized activity (z score).	Screen data with low outlier rate or after outlier removal.	Based on random error assumption. Sensitive to outliers.
Median +/- *k* MAD*	*k* median absolute deviation (MAD) away from the median of normalized activity (z score or B score)	Data contains outliers.	Robust and insensitive to outliers. Potentially lower false positive rate [83] than Mean +/- *k* SD methods. Log transformation may apply for certain assays such as gene activation assays, cell growth assays *etc.*

Table 2: contd…

Quartile	Quartile or interquartile	Non-symmetrical data.	Independent upper and lower thresholds and insensitive to outliers. Some studies showed higher true positive rate [84].

* *k* is typically 3 according to 3-sigma rule [84, 85]

3.2. Group-Comparison-Based Methods

In order to minimize both false positive and false negative rates, it is preferred to screen replicates. Particularly, it is now a routine practice for siRNA confirmatory screens [86]. In fact, group comparison methods were originally developed specifically for such screens with replicates to prioritize siRNA hits.

3.2.1. Multiple t-Tests

T-test method is usually applied to HTS screens with replicates. It is used to evaluate if a sample activity is significantly different from negative references with the consideration of variance from replicates. Paired *t*-test eliminates plate-to-plate variability by calculating *t* statistic based on intra-plate comparison, thus is preferred to unpaired *t*-test [87] (Equation 7).

Equation 7

$$t_i = \frac{\bar{d}_i}{s_i/\sqrt{n}}$$

where t_i is the t score of sample *i* with *n* replicates, $\bar{d}_i$ and s_i are the mean and standard deviation of intra-plate difference between sample and reference across all plates, respectively. Instead of sample activity, the *p* value of *t*-test is used to rank samples. One condition of the *t*-test method is the assumption that t_i follows a normal distribution instead of t distribution. This is not always accurate especially for the small number of replicates in a standard HTS screen [87]. Wright and Simon introduced the randomized variance model (RVM) *t*-test method [88] that uses a corrected variance from an F-distribution based on the variance of replicates. RVM *t*-test method is found to improve the sensitivity and lower false-negative rate [62].

3.2.2. Strictly Standardized Mean Difference (SSMD) Metric

Alternatively, Zhang *et al.* introduced SSMD-based "hit" selection methods to HTS [86, 87, 89-92]. Similar to *t*-test, SSMD measures the statistical significance between samples and negative references. It is more robust with small sample size and can be used for inter-experiment comparisons [93]. Without a replicate, SSMD applies the constant error assumption to estimate the sample variance from variance of all measurements and is essentially equivalent to *Z*-score methods. With replicates, paired SSMD is appropriate because it considers plate-to-plate variations. SSMD of sample *i* with n replicates by uniformly minimal variance unbiased estimate (UMVUE) [89] is shown in Equation 8.

Equation 8

$$SSMD_i = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\sqrt{\frac{2}{n-1}}\frac{\overline{d}_i}{s_i}$$

where $\overline{d}_i$ and s_i are the mean and standard deviation of all paired difference between sample and negative reference in the screen, and they can be substituted by median and MAD for the robust version. These parameters have been used for hit identification in several siRNA screens with good controls on false negative and false positive rates [86, 87, 89-92]. Moreover, SSMD-based cutoffs were also developed to quantify the significance of siRNA hits [78].

3.3. Cluster-Based Method

While many hit identification methods depend solely on assay activities, recent advances also incorporate cheminformatics techniques. The underlying hypothesis is that a group of hits with similar chemical structure reinforce the confidence of positive signal based on the presumption that similar compounds tend to have similar biological activities. Klekota *et al.* [94] clustered a compound library using k-mode algorithm based on Daylight fingerprints [95], and the resulting compound clusters were scored using the p-value from the hypergeometric distribution (Fisher's exact test):

Equation 9

$$P(N,n,m,r) = \sum_{i=r}^{\min(m,n)} \frac{\binom{n}{i}\binom{N-n}{m-i}}{\binom{N}{m}}$$

where *m* is the number of compounds in a cluster, *r* is number of potential active compounds, and *n* is the total number of potential hits in the compound library of *N* compounds. The method was able to identify 87% of known active compounds from a set of published test compounds, suggesting better performance than ones based on individual sample activity.

Instead of defining one activity threshold, Yan *et al.* [96] adopted ontology-based pattern identification (OPI) method from microarray analysis [97, 98] to calculate minimum p-value for each cluster. The *p* value was based on the accumulative hypergeometric distribution in Equation 9, and computed for r = 1, 2, …min(*m,n*). The minimum *p* value was used to score the cluster. Retrospective studies on a set of HTS data showed a significant improvement of confirmation rates from 55% to 85% over threshold based methods.

The OPI-based method was also adopted for RNAi screens, called redundant siRNA activity (RSA) analysis method [99]. In this context, siRNAs were clustered by gene, and *p*-value (Equation 9) was calculated for each gene. The authors assessed the performance of RSA method using two genome-wide RNAi screens and found it significantly better than conventional activity-based methods.

Molecular clustering can also be used to design a better HTS experiments. For example, molecular clusters have been used to reduce the total number of molecules in the primary screen by about two thirds [100]. Two common strategies used to cluster small molecules are similarity-based and scaffold-based clustering. Similarity-based clustering groups structurally similar molecules together based on whole molecule fingerprint. In this method, two molecules with similar fingerprint are believed to have overall similar chemical fragments, but it allows small changes even at its core as long as the overall similarity is maintained. On the other hand, scaffold-based clustering groups molecules based on a well-defined common core substructure. Scaffold based clustering often

produces scaffolds similar to medicinal chemists' intuition. Scaffolds are often the starting points from which lead optimization proceeds. One commonly used approach to compute scaffolds is the molecular framework algorithm described by Bemis and Murcko [101, 102]. Wilkens *et al.* [103] also developed an exhaustive ring-based algorithm. It recursively identifies all possible ring-delimited substructures within a set of compounds. Molecules are grouped by shared ring substructures (scaffolds) so that common scaffolds obtain higher membership. Once all of the scaffolds for a set of compounds are identified, the hierarchical structural relationships between the scaffold structures are established. The complex network of hierarchical relationships is then utilized to navigate compounds in a structurally directed fashion.

3.4. Bayesian Models

Bayesian statistics were introduced to HTS hit identification in late 1990s and early 2000s [104-112]. Bayesian statistics uses probabilities with degrees of confidence to assess a hypothesis based on given evidences, and updates the probabilities with new evidences according to the Bayes' theorem [113]. Besides assay activity, Bayesian approaches incorporate compound structure properties. As a robust and noise tolerant method, Bayesian inference method utilizes molecular descriptors, such as fingerprints [106, 114, 115] as representations of compound structures and the correlation between molecular features and biological activities were examined to build a probability distribution of active and inactive compounds [106]. Naïve Bayesian models such as Laplacian-modified Naïve Bayesian (LMNB) [106, 116] and multiple-category Bayesian models (MCBM) [117, 118] simplify the joint probabilities by assuming the independence and equal importance of molecular descriptors.

It is found that robust performance could be obtained even with noisy data. Zhang *et al.* integrated plate-wise and experiment-wise information for hit selection based on Bayesian hypothesis for siRNA screens [112]. The authors developed two models to compute the posterior probability for siRNA activities. One was based on negative control alone, and the other was based on both positive and negative controls. They found that the Bayesian methods outperformed the Median +/- *k* MAD method in both cases.

3.5. Dose-Response Analysis

Single dose primary screens measure compound activity at a single concentration. As a result, they often fail to identify complex pharmacological actions such as weak partial agonists or antagonists. They are known to have high false positive and false negative rates [65]. In subsequent confirmation screens, assay activities for individual compound are usually measured at a series of concentrations to investigate the true compound activities. Due to more reliable signals from multiple dose experiments, assays with concentration-dependent activity measurements have also been adopted in some primary screens [119].

Dose-response curves are normally sigmoidal curves in log-based compound concentration. It is defined by four parameters: top asymptote (maximal response) *T*, bottom asymptote (baseline response) *B*, slope (Hill slope or Hill coefficient) *h*, and the EC_{50} value (the concentration at half-maximal effect). The Hill equation [120] for assay readout signal and compound concentration *x* is:

Equation 10

$$signal = B + \frac{T - B}{1 + \left(\frac{EC_{50}}{x}\right)^{h}}$$

Three values derived from Equation 10 are usually used to describe compound potency in hit identification purposes, *i.e.* EC_{50}, fold change $FC = \frac{T-B}{T}$, and the area under the dose-response curve (AUC) [121]. *T*, *B,* $EC_{50}$and AUC are labeled for a dose-response curve (blue) of an antagonist assay in Fig. **9**. In general, EC_{50} represents compound potency; FC measures efficacy and AUC measures total effects of a compound. In inhibitory assays, IC_{50} is commonly used in place of EC_{50}. Partial agonists/antagonists are usually defined as compounds with reduced efficacy, *i.e.* $FC = \frac{T-B}{T} < 1$ and $\frac{FC}{FC_{ref}} < 1$ (FC_{ref}= 1 when there's no reference compound with full efficacy) [122]. In a simple second-order competitive inhibition (Fig. **9**, black curve), $T = 1$, $B = 0$, and $h = 1$ [121]. AUC, on the other hand, measures both potency and efficacy simultaneously in a given dose range [121].

The pattern of dose responsive can also be attributed to different biological processes. Hill slope *h* implies the ratio of stoichiometry for enzyme-inhibitor or

protein-ligand complexes, thus can be used to derive binding mechanism such as cooperativity, multiple binding sites, and to determine compound behavior such as protein denaturant or micelle formation [123]. Moreover, the sign of Hill slope *h* indicates agonism or antagonism behavior of a compound. As shown in Fig. **9**, an agonist triggers a boost of activity with increasing compound concentrations (Fig. **9** orange curve), while an antagonist shows a reversed effect (Fig. **9** blue curve). Bell curves (Fig. **9** green curve) are also common in agonist assays due to compound toxicity at higher concentration or the initiation of other cellular responses.

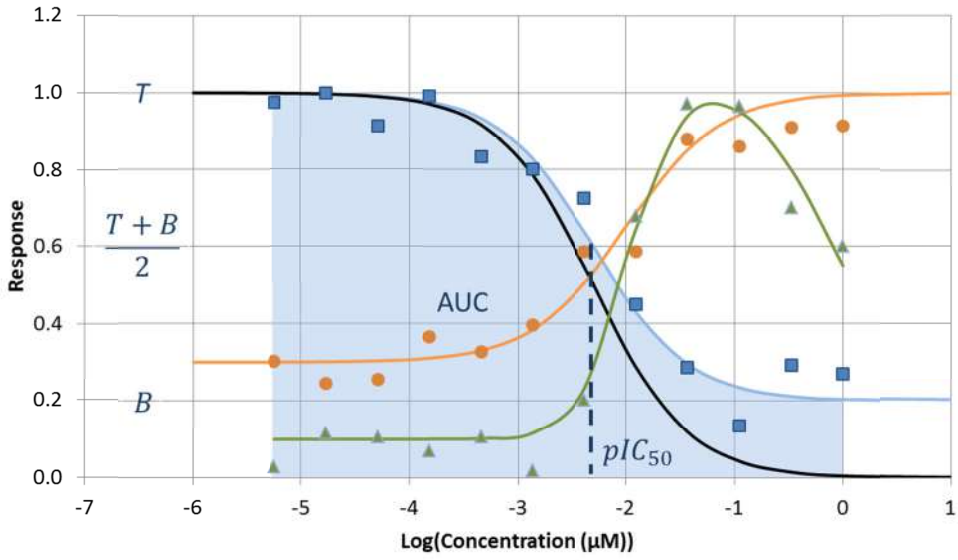


Figure 9: Examples of dose-response curves. Black: A simple case of second-order competitive inhibition. Blue: a partial antagonist curve. Orange: a partial agonist curve. Green: a bell curve for an agonist assay. *T*, *B,* and $\boldsymbol{pIC_{50}}$ are labled for the inhibitory curve in blue and the shaded area represents AUC.

4. HTS HITS ANALYSIS

4.1. Analysis of Small Molecule Screen Results

Hit identification process usually is followed by investigations of chemical and biological properties of selected compounds. Hit characterization provides valuable information for hit selection, such as scaffold information and SAR profiles in biological systems. At the same time, the accumulation of HTS data

across different biological systems presents a unique opportunity for understanding the interactions between chemical and biological systems. To this end, HTS data are being used repeatedly for comprehensive mining of chemical biology information across assays (bio-profiling).

4.1.1.Chemical Properties

Various cheminformatics methods are developed for HTS data analysis including clustering analysis, recursive partitioning, similarity search and pharmacophore modeling *etc.* [77] These methods were originally designed for predictive (quantitative) structure-activity relationship [(Q)SAR] models to select compounds for sequential screening or smart screening [124-126]. The predictive (Q)SAR models are continuously refined by new experimental results and ultimately used to aid the hit-to-lead process. Recently, a growing number of studies focused on compound activities in multiple physiological targets, *i.e.* drug polypharmacology [127, 128]. Che *et al.* [129] explored chemical properties of three million structures screened across 277 assays and identified 6% hits as "frequent hits" or "promiscuous hits", *i.e.* molecules that are active across many independent assays. Novartis and others also used the scaffolds from large-scale HTS campaigns to generate novel antimalarial pharmacophores [130]. Known promiscuous chemical structural features can be used to filter out hits in the hit-to-lead process.

4.1.2. Biological Properties

With large amount of HTS data, people have realized some molecules appear to be active across large number of assays, *i.e.* "frequent hits" or "promiscuous hits". The mechanisms of these promiscuous hits have been extensively studied [131-136]. One reason is compound aggregation [131-135]. Aggregated molecules interrupt enzymatic activity or cell viability in a nonspecific way. By mining an HTS database covering large chemical space and diverse biological functions, Che *et al.* identified many novel chemical features and several biological processes that are associated with a significant portion of frequent hits [129]. It is also noted that several marketed drugs also contained characteristics that are commonly associated with frequent hits. This observation suggested that simple filtering strategies for triaging compounds may result in discarding compounds

with good properties. Therefore, a novel strategy that overlaid chemical scaffolds, biological processes, along with empirical hit frequency data was developed. The risk of removing biologically-relevant frequent hits was reduced compared to the simple empirical hit frequency-based filtering strategy.

In the work of Che *et al.* [127], Mode of action (MOA) and target analysis of frequent hits from 55 biochemical assays and 222 cellular assays suggest that frequent hits generated in biomedical screens are due to either signal artifacts associated with their physicochemical proprieties or non-specific binding to a large number of proteins, while frequent hits in cellular assays are usually related to disruption of critical biological pathways. Therefore, it is recommended that biochemical frequent hits should be removed from hit selection while cellular frequent hits should be retained for further evaluation. Similar concept has also been applied on screen sub-library design and compound-target diversity of HTS library can improve both hit rate and hit scaffold diversity [137].

4.2. Analysis of RNAi Screen Results

Results of RNAi screens require additional follow-up studies based on gene functional annotations for hit prioritization to minimize false discovery rate. Biological implications of hits from gene function in known pathways and networks are often used during the downstream analysis. The analysis usually involves gene annotation, gene ontology, pathway analysis, protein-protein interactions, *etc.* Gene annotations are continuously revised by projects like the ENCyclopedia Of DNA Elements (ENCODE) [138-140] and modENCODE [141], which improves RNAi reagent library design [142] to reduce off-target effects [143]. Enrichment analysis on gene ontology (GO) is the most frequently used method for biological functional annotations for hits [144, 145]. In addition, GO and pathway analysis provides meaningful comparisons on hits from different screens on functional levels. For example, substantially more overlaps were identified from screening hits at the level of GO or pathways than at the gene level in the studies of human immunodeficiency virus (HIV) infection in mammalian cells [146] and the JAK/STAT signaling pathway using two different *Drosophila* cell types [147]. Tu *et al*. developed a network-based approach to link RNAi screening hits with PPI related to insulin signaling pathway, and it significantly enriched siRNA hits associated with insulin

resistance [148]. Furthermore, combinatorial studies on multiple screens are often useful for a comprehensive understanding of specific gene functions as well as their environmental dependence. For example, Schlabach *et al*. implemented pooled shRNA screens across four different cancer cell lines to identify shRNAs related to cell proliferation and viability in cancer cells and normal cells [149]. Bakal *et al*. designed 12 combinatory screens on 1395 dsRNAs with a total of 17,724 combinatorial tests to study *Drosophila* JUN NH2-terminal kinase (JNK) phosphorylation network [150].

SUMMARY

HTS is a powerful tool to identify active molecules from a large library. Recent scientific and technological advances improved HTS technique and dramatically enhanced its throughput and quality. In this chapter, we discussed the key components of HTS data analysis. For quality control of HTS results, we summarize key statistical principles and their applications on the removal of artifacts. For hit identification, we review the state of the art methodologies commonly applied in primary and secondary screens, and summarize their pros and cons. At the end, we discuss efforts for characterizing HTS hits. In summary, HTS data analysis provides rich opportunities from identifying novel molecular entities for modulating biological processes to understanding mechanism of action of molecules to exploring new applications of known molecules. With advancements in assay technologies and informatics methods, we anticipate that HTS will be increasingly popular and productive.

ACKNOWLEDGEMENTS

The authors thank Dr. Yingyao Zhou and Dr. Edward Ainscow for their comments and suggestions in the preparation of this manuscript. We also thank Dr. Yufeng Zhai for providing HTS image examples. This work is supported by the National Institutes of Health through the grants R01GM100785 (Y. Zhong and J. Che) and R01GM096188 (Z. Guo and J. Che).

CONFLICT OF INTEREST

The authors confirm that this chapter content have no conflict of interest.

REFERENCES

[1] Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S., Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011,** *10* (3), 188-195.

[2] Delvin, J. P., *High Throughput Screening: The Discovery of Bioactive Substances*. Marcel Dekker: New York, 1997.

[3] Janzen, W. P., *High Throughput Screening: Methods and Protocols*. Humana Press: Totowa, New Jersey, 2002.

[4] Boutros, M.; Kiger, A. A.; Armknecht, S.; Kerr, K.; Hild, M.; Koch, B.; Haas, S. A.; Paro, R.; Perrimon, N.; Consortium, H. F. A., Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science* **2004,** *303* (5659), 832-835.

[5] He, F.; Woods, C. E.; Becker, G. W.; Narhi, L. O.; Razinkov, V. I., High-throughput assessment of thermal and colloidal stability parameters for monoclonal antibody formulations. *J Pharm Sci* **2011,** *100* (12), 5126-41.

[6] Gibson, T. J.; Mccarty, K.; Mcfadyen, I. J.; Cash, E.; Dalmonte, P.; Hinds, K. D.; Dinerman, A. A.; Alvarez, J. C.; Volkin, D. B., Application of a High-Throughput Screening Procedure with PEG-Induced Precipitation to Compare Relative Protein Solubility During Formulation Development with IgG1 Monoclonal Antibodies. *J Pharm Sci-Us* **2011,** *100* (3), 1009-1021.

[7] Yoshikawa, M.; Mukai, Y.; Okada, Y.; Tsumori, Y.; Tsunoda, S.; Tsutsumi, Y.; Aird, W. C.; Yoshioka, Y.; Okada, N.; Doi, T.; Nakagawa, S., Robo4 is an effective tumor endothelial marker for antibody-drug conjugates based on the rapid isolation of the anti-Robo4 cell-internalizing antibody. *Blood* **2013,** *121* (14), 2804-13.

[8] Li, M.; Gao, J.; Feng, R.; Wang, Y.; Chen, X.; Sun, J.; Zhang, D.; Zhu, Z.; Ellis, L. M.; Lu, M.; Lee, J. E.; Feng, Z.; Liu, B., Generation of monoclonal antibody MS17-57 targeting secreted alkaline phosphatase ectopically expressed on the surface of gastrointestinal cancer cells. *PloS one* **2013,** *8* (10), e77398.

[9] Pereira, D. A.; Williams, J. A., Origin and evolution of high throughput screening. *British journal of pharmacology* **2007,** *152* (1), 53-61.

[10] Razvi, F. M.; Kritzinger, E. E.; Taylor, K.; Jones, S.; Ryder, R. E. J., Oral fluorescein angiography as an additional tool in diabetic retinopathy screening. *Diabetologia* **1999,** *42*, A313-A313.

[11] Sundberg, S. A., High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. *Curr Opin Biotech* **2000,** *11* (1), 47-53.

[12] Yi, L.; Li, Z. Q.; Yuan, K. H.; Qu, X. X.; Chen, J.; Wang, G. W.; Zhang, H.; Luo, H. P.; Zhu, L. L.; Jiang, P. F.; Chen, L. R.; Shen, Y.; Luo, M.; Zuo, G. Y.; Hu, J. H.; Duan, D. L.; Nie, Y. C.; Shi, X. L.; Wang, W.; Han, Y.; Li, T. S.; Liu, Y. Q.; Ding, M. X.; Deng, H. K.; Xu, X. J., Small molecules blocking the entry of severe acute respiratory syndrome coronavirus into host cells. *J Virol* **2004,** *78* (20), 11334-11339.

[13] Mayr, L. M.; Fuerst, P., The future of high-throughput screening. *J Biomol Screen* **2008,** *13* (6), 443-448.

[14] Ramm, P., Imaging systems in assay screening. *Drug Discov Today* **1999,** *4* (9), 401-410.

[15] Gasparri, F., An overview of cell phenotypes in HCS: limitations and advantages. *Expert Opin Drug Dis* **2009,** *4* (6), 643-657.

[16] Brenan, C. J.; Roberts, D.; Hurley, J., Nanoliter high-throughput PCR for DNA and RNA profiling. *Methods in molecular biology* **2009,** *496*, 161-74.
[17] Horiuchi, K. Y.; Ma, H., Innovative chemical compound microarrays for drug screening. In *DrugsPlus.International*, 2006.
[18] Brooks Automation, I. AURORA MICROPLATES COST SAVINGS FROM ASSAY MINIATURIZATION. http://www.brooks.com/products/life-science-systems/consumables/~/media/Files/BLSS%20Literature/BLSS%200014%20Aurora%20Cost%20Savings.pdf.
[19] Rodems, S. M.; Hamman, B. D.; Lin, C.; Zhao, J.; Shah, S.; Heidary, D.; Makings, L.; Stack, J. H.; Pollok, B. A., A FRET-based assay platform for ultra-high density drug screening of protein kinases and phosphatases. *Assay Drug Dev Techn* **2002,** *1* (1), 9-19.
[20] Swinney, D. C.; Anthony, J., How were new medicines discovered? *Nat Rev Drug Discov* **2011,** *10* (7), 507-519.
[21] Mayr, L. M.; Bojanic, D., Novel trends in high-throughput screening. *Curr Opin Pharmacol* **2009,** *9* (5), 580-588.
[22] Oprea, T. I.; Bologa, C. G.; Edwards, B. S.; Prossnitz, E. R.; Sklar, L. A., Post-high-throughput screening analysis: an empirical compound prioritization scheme. *J Biomol Screen* **2005,** *10* (5), 419-26.
[23] McFadyen, I.; Walker, G.; Alvarez, J., Enhancing hit quality and diversity within assay throughput constrains. In *Chemoinformatics in Drug Discovery*, Wiley-VCH: Weinheim, 2005; pp 143-173.
[24] Walters, W. P.; Namchuk, M., Designing screens: How to make your hits a hit. *Nat Rev Drug Discov* **2003,** *2* (4), 259-266.
[25] Hann, M. M.; Oprea, T. I., Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **2004,** *8* (3), 255-263.
[26] Thorne, N.; Auld, D. S.; Inglese, J., Apparent activity in high-throughput screening: origins of compound-dependent assay interference. *Curr Opin Chem Biol* **2010,** *14* (3), 315-324.
[27] Inglese, J.; Johnson, R. L.; Simeonov, A.; Xia, M. H.; Zheng, W.; Austin, C. P.; Auld, D. S., High-throughput screening assays for the identification of chemical probes. *Nat Chem Biol* **2007,** *3* (8), 466-479.
[28] Shapiro, A. B.; Walkup, G. K.; Keating, T. A., Correction for Interference by Test Samples in High-Throughput Assays. *J Biomol Screen* **2009,** *14* (8), 1008-1016.
[29] Carralot, J. P.; Ogier, A.; Boese, A.; Genovesio, A.; Brodin, P.; Sommer, P.; Dorval, T., A novel specific edge effect correction method for RNA interference screenings. *Bioinformatics* **2012,** *28* (2), 261-268.
[30] Meeker, M. Internet trends 2014 - code conference. www.kpcb.com/InternetTrends.
[31] Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D. P.; Kania, R.; Schaeffer, M.; St Pierre, S.; Twigger, S.; White, O.; Rhee, S. Y., Big data: The future of biocuration. *Nature* **2008,** *455* (7209), 47-50.
[32] Wang, Y. L.; Suzek, T.; Zhang, J.; Wang, J. Y.; He, S. Q.; Cheng, T. J.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H., PubChem BioAssay: 2014 update. *Nucleic Acids Res* **2014,** *42* (D1), D1075-D1082.
[33] Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P., A high-throughput screen for aggregation-based inhibition in a large compound library. *J Med Chem* **2007,** *50* (10), 2385-2390.

[34] Li, Q. L.; Wang, Y. L.; Bryant, S. H., A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **2009,** *25* (24), 3310-3316.

[35] Team, R. d. C., *R: A language and environment for statistical computing.* . R Foundation for Statistical Computing: Vienna, Austria, 2008.

[36] Becker, S.; Tetsch, P.; Karcher, R.; Trippler, S., Statistical Evaluation of 9000 Apicoectomies - Technical Changes in Intraoperative Root Filling. *Int J Oral Max Surg* **1988,** *17* (2), 147-147.

[37] Spotfire, Spotfire User Conference Presentations, Life Sciences.

[38] Li, M.; Gao, J. P.; Feng, R. H.; Wang, Y. L.; Chen, X. H.; Sun, J. Y.; Zhang, D. Q.; Zhu, Z. G.; Ellis, L. M.; Lu, M.; Lee, J. E.; Feng, Z. Q.; Liu, B. Y., Generation of Monoclonal Antibody MS17-57 Targeting Secreted Alkaline Phosphatase Ectopically Expressed on the Surface of Gastrointestinal Cancer Cells. *PloS one* **2013,** *8* (10).

[39] Huser, J.; Lohrmann, E.; Kalthof, B.; Burkhardt, N.; Bruggemeier, U.; Bechem, M., High-throughput Screening for Targeted Lead Discovery. In *High Throughput-Screening in Drug Discovery*, Huser, J., Ed. Wiley-VCH: Weinheim, 2006; Vol. 35, pp 15-36.

[40] Harper, G.; Pickett, S. D.; Green, D. V., Design of a compound screening collection for use in high throughput screening. *Combinatorial chemistry & high throughput screening* **2004,** *7* (1), 63-70.

[41] Wintner, E. A.; Moallemi, C. C., Quantized surface complementarity diversity (QSCD): a model based on small molecule-target complementarity. *J Med Chem* **2000,** *43* (10), 1993-2006.

[42] Roth, H. J., There is no such thing as 'diversity'! *Curr Opin Chem Biol* **2005,** *9* (3), 293-5.

[43] Willett, P.; Barnard, J. M.; Downs, G. M., Chemical similarity searching. *J Chem Inf Comp Sci* **1998,** *38* (6), 983-996.

[44] Shoichet, B. K., Screening in a spirit haunted world. *Drug Discov Today* **2006,** *11* (13-14), 607-15.

[45] Macarron, R.; Hertzberg, R. P., Design and Implementation of High-Throughput Screening Assays. In *High Throughput Screening: Methods and Protocols*, 2nd ed.; Janzen, W. P.; Bernasconi, P., Eds. Humana Press: New York, 2009.

[46] J.Q., W., *Six Sigma Quality Design & Control. Desirable Precision and Requisite QC for Laboratory Measurement Processes.* Westgard QC Inc.: Madison, 2001.

[47] Coma, I.; Herranz, J.; Martin, J., Statistics and Decision Making in High-Throughput Screening. In *High Throughput Screening Methods and Protocols*, 2nd ed.; Janzen, W. P.; Bernasconi, P., Eds. Humana Press: New York, 2009; pp 69-106.

[48] Gunter, B.; Brideau, C.; Pikounis, B.; Liaw, A., Statistical and graphical methods for quality control determination of high-throughput screening data. *J Biomol Screen* **2003,** *8* (6), 624-33.

[49] Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A., Improved statistical methods for hit selection in high-throughput screening. *J Biomol Screen* **2003,** *8* (6), 634-47.

[50] Gribbon, P.; Lyons, R.; Laflin, P.; Bradley, J.; Chambers, C.; Williams, B. S.; Keighley, W.; Sewing, A., Evaluating real-life high-throughput screening data. *J Biomol Screen* **2005,** *10* (2), 99-107.

[51] Gubler, H.; Girod, M.; Dressler, S.; Bouhelal, R.; Ottl, J. In *HTS data analysis in the real world: practical experience with HTS data quality assurance systems and recent integration of the GeneData Screener software*, Lab-Automation, San Jose, CA, Association for Laboratory Automation: San Jose, CA, 2004; p 65.

[52] Reimann, S.; Lindemann, M.; Rinn, B.; Lefevre, O.; Heyse, S., Large scale, comprehensive quality control and analysis of high-throughput screening data. In *European BioPharmaceutical Review, Applied R&D*, Spring: Samedan, London, 2003.

[53] Heyse, S. In *Comprehensive analysis of high-throughput screening data*, SPIE, 2002; pp 535-547.
[54] CyBi *SIENA Software*, CyBio: Jena.
[55] BlueFuse *BlueFuse for Microarrays Software*, Blue-gnome: Cambridge.
[56] Root, D. E.; Kelley, B. P.; Stockwell, B. R., Detecting spatial patterns in biological array experiments. *J Biomol Screen* **2003,** *8* (4), 393-8.
[57] Kevorkov, D.; Makarenkov, V., Statistical analysis of systematic errors in high-throughput screening. *J Biomol Screen* **2005,** *10* (6), 557-67.
[58] Kelley, B. P.; Lunn, M. R.; Root, D. E.; Flaherty, S. P.; Martino, A. M.; Stockwell, B. R., A flexible data analysis tool for chemical genetic screens. *Chemistry & biology* **2004,** *11* (11), 1495-503.
[59] Brideau, C.; Gunter, B.; Pikounis, B.; Liaw, A., Improved statistical methods for hit selection in high-throughput screening. *J Biomol Screen* **2003,** *8* (6), 634-647.
[60] Coma, I.; Clark, L.; Diez, E.; Harper, G.; Herranz, J.; Hofmann, G.; Lennon, M.; Richmond, N.; Valmaseda, M.; Macarron, R., Process Validation and Screen Reproducibility in High-Throughput Screening. *J Biomol Screen* **2009,** *14* (1), 66-76.
[61] Gubler, H., Methods for statistical analysis, quality assurance and management of primary high-throughput screening data. In *High-throughput Screening in Drug Discovery* Wiley-VCH: Weinheim, 2006.
[62] Malo, N.; Hanley, J. A.; Carlile, G.; Liu, J.; Pelletier, J.; Thomas, D.; Nadon, R., Experimental design and statistical methods for improved hit detection in high-throughput screening. *J Biomol Screen* **2010,** *15* (8), 990-1000.
[63] Wu, Z. J.; Liu, D. M.; Sui, Y. X., Quantitative assessment of hit detection and confirmation in single and duplicate high-throughput screenings. *J Biomol Screen* **2008,** *13* (2), 159-167.
[64] Tukey, J. W., *Exploratory Data Analysis*. Addison-Wesley: Cambridge, MA, 1977.
[65] Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R., Statistical practice in high-throughput screening data analysis. *Nat Biotechnol* **2006,** *24* (2), 167-175.
[66] Shun, T. Y.; Lazo, J. S.; Sharlow, E. R.; Johnston, P. A., Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *J Biomol Screen* **2011,** *16* (1), 1-14.
[67] Wiles, A. M.; Ravi, D.; Bhavani, S.; Bishop, A. J. R., An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme. *J Biomol Screen* **2008,** *13* (8), 777-784.
[68] Birmingham, A.; Selfors, L. M.; Forster, T.; Wrobel, D.; Kennedy, C. J.; Shanks, E.; Santoyo-Lopez, J.; Dunican, D. J.; Long, A.; Kelleher, D.; Smith, Q.; Beijersbergen, R. L.; Ghazal, P.; Shamu, C. E., Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* **2009,** *6* (8), 569-575.
[69] Malo, J.; Epifanio, I.; Navarro, R.; Simoncelli, E. P., Nonlinear image representation for efficient perceptual coding. *Ieee T Image Process* **2006,** *15* (1), 68-80.
[70] Makarenkov, V.; Zentilli, P.; Kevorkov, D.; Gagarin, A.; Malo, N.; Nadon, R., An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **2007,** *23* (13), 1648-1657.
[71] Huser, D. J., *High-Throughput Screening in Drug Discovery*. Wiley-VCH: Weinheim, 2006; Vol. 35.
[72] Makarenkov, V.; Zentilli, P.; Kevorkov, D.; Gagarin, A.; Malo, N.; Nadon, R., An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **2007,** *23* (13), 1648-57.

[73] Dragiev, P.; Nadon, R.; Makarenkov, V., Systematic error detection in experimental high-throughput screening. *Bmc Bioinformatics* **2011,** *12*, 25.

[74] Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R., A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen* **1999,** *4* (2), 67-73.

[75] Zhang, X. D., A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* **2007,** *89* (4), 552-61.

[76] Birmingham, A.; Selfors, L. M.; Forster, T.; Wrobel, D.; Kennedy, C. J.; Shanks, E.; Santoyo-Lopez, J.; Dunican, D. J.; Long, A.; Kelleher, D.; Smith, Q.; Beijersbergen, R. L.; Ghazal, P.; Shamu, C. E., Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* **2009,** *6* (8), 569-75.

[77] Parker, C. N.; Schreyer, S. K., Application of chemoinformatics to high-throughput screening: practical considerations. *Methods in molecular biology* **2004,** *275*, 85-110.

[78] Zhang, X. D., A method for effectively comparing gene effects in multiple conditions in RNAi and expression-profiling research. *Pharmacogenomics* **2009,** *10* (3), 345-58.

[79] Alanine, A.; Nettekoven, M.; Roberts, E.; Thomas, A. W., Lead generation--enhancing the success of drug discovery by investing in the hit to lead process. *Combinatorial chemistry & high throughput screening* **2003,** *6* (1), 51-66.

[80] Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I., Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2003,** *2* (5), 369-78.

[81] Hampel, F. R., A general qualitative definition of robustness. *Ann. Math. Statist.* **1971,** *42*, 1887-1896.

[82] Rousseeuw, P.; Van Zomeren, B., Robust distances: Simulations and cutoff values. In *Directions in Robust Statistics and Diagnostics Part 2*, Stahel, W.; Weisberg, S., Eds. Springer-Verlag: New York, 1991; pp 195-203.

[83] Chung, N.; Zhang, X. D.; Kreamer, A.; Locco, L.; Kuan, P. F.; Bartz, S.; Linsley, P. S.; Ferrer, M.; Strulovici, B., Median absolute deviation to improve hit selection for genome-scale RNAi screens. *J Biomol Screen* **2008,** *13* (2), 149-58.

[84] Zhang, X. D.; Yang, X. C.; Chung, N.; Gates, A.; Stec, E.; Kunapuli, P.; Holder, D. J.; Ferrer, M.; Espeseth, A. S., Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics* **2006,** *7* (3), 299-309.

[85] Vysochanskij, D. F.; Petunin, Y. I., Justification of the 3-Sigma Rule for Unimodal Distribution. *Theory Probab. Math. Stat.* **1980,** *21*, 25-36.

[86] Zhang, X. D.; Heyse, J. F., Determination of sample size in genome-scale RNAi screens. *Bioinformatics* **2009,** *25* (7), 841-4.

[87] Zhang, X. D., Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens. *J Biomol Screen* **2011,** *16* (7), 775-85.

[88] Wright, G. W.; Simon, R. M., A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **2003,** *19* (18), 2448-2455.

[89] Zhang, X. D., A new method with flexible and balanced control of false negatives and false positives for hit selection in RNA interference high-throughput screening assays. *J Biomol Screen* **2007,** *12* (5), 645-655.

[90] Zhang, X. D., Genome-wide screens for effective siRNAs through assessing the size of siRNA effects. *BMC Research Notes* **2008,** *1*, 33-39.

[91] Zhang, X. H. D., An Effective Method for Controlling False Discovery and False Nondiscovery Rates in Genome-Scale RNAi Screens. *J Biomol Screen* **2010,** *15* (9), 1116-1122.

[92] Zhang, X. H. D.; Lacson, R.; Yang, R. J.; Marine, S. D.; McCampbell, A.; Toolan, D. M.; Hare, T. R.; Kajdas, J.; Berger, J. P.; Holder, D. J.; Heyse, J. F.; Ferrer, M., The Use of SSMD-Based False Discovery and False Nondiscovery Rates in Genome-Scale RNAi Screens. *J Biomol Screen* **2010,** *15* (9), 1123-1131.

[93] Zhang, X. H. D., Illustration of SSMD, z Score, SSMD*, z* Score, and t Statistic for Hit Selection in RNAi High-Throughput Screens. *J Biomol Screen* **2011,** *16* (7), 775-785.

[94] Klekota, J.; Brauner, E.; Schreiber, S. L., Identifying biologically active compound classes using phenotypic screening data and sampling statistics. *J Chem Inf Model* **2005,** *45* (6), 1824-1836.

[95] Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J Chem Inf Model* **2010,** *50* (5), 742-754.

[96] Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Y., Novel statistical approach for primary high-throughput screening hit selection. *J Chem Inf Model* **2005,** *45* (6), 1784-1790.

[97] Zhou, Y. Y.; Young, J. A.; Santrosyan, A.; Chen, K. S.; Yan, S. F.; Winzeler, E. A., In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* **2005,** *21* (7), 1237-1245.

[98] Young, J. A.; Fivelman, Q. L.; Blair, P. L.; de la Vega, P.; Le Roch, K. G.; Zhou, Y. Y.; Carucci, D. J.; Baker, D. A.; Winzeler, E. A., The Plasmodium falciparum sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Mol Biochem Parasit* **2005,** *143* (1), 67-79.

[99] Konig, R.; Chiang, C. Y.; Tu, B. P.; Yan, S. F.; DeJesus, P. D.; Romero, A.; Bergauer, T.; Orth, A.; Krueger, U.; Zhou, Y.; Chanda, S. K., A probability-based approach for the analysis of large-scale RNAi screens. *Nat Methods* **2007,** *4* (10), 847-849.

[100] Karnachi, P. S.; Brown, F. K., Practical approaches to efficient screening: information-rich screening protocol. *J Biomol Screen* **2004,** *9* (8), 678-86.

[101] Bemis, G. W.; Murcko, M. A., The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996,** *39* (15), 2887-93.

[102] Swamidass, S. J.; Calhoun, B. T.; Bittker, J. A.; Bodycombe, N. E.; Clemons, P. A., Enhancing the rate of scaffold discovery with diversity-oriented prioritization. *Bioinformatics* **2011,** *27* (16), 2271-8.

[103] Wilkens, S. J.; Janes, J.; Su, A. I., HierS: Hierarchical scaffold clustering using topological chemical graphs. *J Med Chem* **2005,** *48* (9), 3182-3193.

[104] Labute, P., Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* **1999**, 444-55.

[105] Labute, P.; Nilar, S.; Williams, C., A probabilistic approach to high throughput drug discovery. *Combinatorial chemistry & high throughput screening* **2002,** *5* (2), 135-45.

[106] Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W., Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J Biomol Screen* **2004,** *9* (1), 32-6.

[107] Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D., Classification of kinase inhibitors using a Bayesian model. *J Med Chem* **2004,** *47* (18), 4463-70.

[108] Rogers, D.; Brown, R. D.; Hahn, M., Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* **2005,** *10* (7), 682-6.

[109] Klon, A. E.; Diller, D. J., Library fingerprints: a novel approach to the screening of virtual libraries. *J Chem Inf Model* **2007,** *47* (4), 1354-65.

[110] Crisman, T. J.; Jenkins, J. L.; Parker, C. N.; Hill, W. A.; Bender, A.; Deng, Z.; Nettles, J. H.; Davies, J. W.; Glick, M., "Plate cherry picking": a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J Biomol Screen* **2007,** *12* (3), 320-7.

[111] Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M., Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J Chem Inf Model* **2007,** *47* (4), 1319-27.

[112] Zhang, X. D.; Kuan, P. F.; Ferrer, M.; Shu, X.; Liu, Y. C.; Gates, A. T.; Kunapuli, P.; Stec, E. M.; Xu, M.; Marine, S. D.; Holder, D. J.; Strulovici, B.; Heyse, J. F.; Espeseth, A. S., Hit selection with false discovery rate control in genome-scale RNAi screens. *Nucleic Acids Res* **2008,** *36* (14), 4667-79.

[113] Bayes, T., An Essay Towards Solving a Problem in the Doctrine of Chances. *Biometrika* **1958,** *45* (3-4), 296-315.

[114] Klon, A. E.; Lowrie, J. F.; Diller, D. J., Improved naive Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. *J Chem Inf Model* **2006,** *46* (5), 1945-56.

[115] Davies, J. W.; Glick, M.; Jenkins, J. L., Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr Opin Chem Biol* **2006,** *10* (4), 343-51.

[116] Xia, X. Y.; Maliski, E. G.; Gallant, P.; Rogers, D., Classification of kinase inhibitors using a Bayesian model. *J Med Chem* **2004,** *47* (18), 4463-4470.

[117] Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Bin Kang, Z.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M., Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J Chem Inf Model* **2007,** *47* (4), 1319-1327.

[118] Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L., Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* **2006,** *46* (3), 1124-1133.

[119] Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P., Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc Natl Acad Sci U S A* **2006,** *103* (31), 11473-8.

[120] Hill, A. V., The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J Physiol (London)* **1910,** *40*, 4-7.

[121] Fallahi-Sichani, M.; Honarnejad, S.; Heiser, L. M.; Gray, J. W.; Sorger, P. K., Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nat Chem Biol* **2013,** *9* (11), 708-+.

[122] Lambert, D. G., Drugs and receptors. *Contin Educ Anaesth Crit Care Pain* **2004,** *4* (6), 181-184.

[123] Copeland, R. A., *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists*. John Wiley & Sons: Hoboken, New Jersey, 2005.

[124] Young, S. S.; Hawkins, D. M., Using recursive partitioning analysis to evaluate compound selection methods. *Methods in molecular biology* **2004,** *275*, 317-34.

[125] Engels, M. F.; Venkatarangan, P., Smart screening: approaches to efficient HTS. *Current opinion in drug discovery & development* **2001,** *4* (3), 275-83.

[126] Young, S. S.; Lam, R. L.; Welch, W. J., Initial compound selection for sequential screening. *Current opinion in drug discovery & development* **2002,** *5* (3), 422-7.

[127] Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L., Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J Chem Inf Model* **2009,** *49* (2), 308-317.

[128] Hu, Y.; Bajorath, J., Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes with Different Activity Profiles and Comparison to Approved Drugs. *J Chem Inf Model* **2010,** *50* (12), 2112-2118.

[129] Che, J. W.; King, F. J.; Zhou, B.; Zhou, Y. Y., Chemical and Biological Properties of Frequent Screening Hits. *J Chem Inf Model* **2012,** *52* (4), 913-926.

[130] Guiguemde, W. A.; Shelat, A. A.; Garcia-Bustos, J. F.; Diagana, T. T.; Gamo, F. J.; Guy, R. K., Global phenotypic screening for antimalarials. *Chem Biol* **2012,** *19* (1), 116-29.

[131] McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K., A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J Med Chem* **2002,** *45* (8), 1712-1722.

[132] Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K., Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding. *J Med Chem* **2009,** *52* (7), 2067-2075.

[133] Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K., High-throughput assays for promiscuous inhibitors. *Nat Chem Biol* **2005,** *1* (3), 146-148.

[134] Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A., Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J Med Chem* **2010,** *53* (1), 37-51.

[135] McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K., A specific mechanism of nonspecific inhibition. *J Med Chem* **2003,** *46* (20), 4265-4272.

[136] Baell, J. B.; Holloway, G. A., New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem* **2010,** *53* (7), 2719-2740.

[137] Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M., Biodiversity of small molecules - a new perspective in screening set selection. *Drug Discov Today* **2013,** *18* (13-14), 674-680.

[138] Feingold, E. A.; Good, P. J.; Guyer, M. S.; Kamholz, S.; Liefer, L.; Wetterstrand, K.; Collins, F. S.; Gingeras, T. R.; Kampa, D.; Sekinger, E. A.; Cheng, J.; Hirsch, H.; Ghosh, S.; Zhu, Z.; Pate, S.; Piccolboni, A.; Yang, A.; Tammana, H.; Bekiranov, S.; Kapranov, P.; Harrison, R.; Church, G.; Struhl, K.; Ren, B.; Kim, T. H.; Barrera, L. O.; Qu, C.; Van Calcar, S.; Luna, R.; Glass, C. K.; Rosenfeld, M. G.; Guigo, R.; Antonarakis, S. E.; Birney, E.; Brent, M.; Pachter, L.; Reymond, A.; Dermitzakis, E. T.; Dewey, C.; Keefe, D.; Denoeud, F.; Lagarde, J.; Ashurst, J.; Hubbard, T.; Wesselink, J. J.; Castelo, R.; Eyras, E.; Myers, R. M.; Sidow, A.; Batzoglou, S.; Trinklein, N. D.; Hartman, S. J.; Aldred, S. F.; Anton, E.; Schroeder, D. I.; Marticke, S. S.; Nguyen, L.; Schmutz, J.; Grimwood, J.; Dickson, M.; Cooper, G. M.; Stone, E. A.; Asimenos, G.; Brudno, M.; Dutta, A.; Karnani, N.; Taylor, C. M.; Kim, H. K.; Robins, G.; Stamatoyannopoulos, G.; Stamatoyannopoulos, J. A.; Dorschner, M.; Sabo, P.; Hawrytycz, M.; Humbert, R.; Wallace, J.; Yu, M.; Navas, P. A.; McArthur, M.; Noble, W. S.; Dunham, I.; Koch, C. M.; Andrews, R. M.; Clelland, G. K.; Wilcox, S.; Fowler, J. C.; James, K. D.; Groth, P.; Dovey, O. M.; Ellis, P. D.; Wraight, V. L.; Mungall, A. J.; Dhami, P.; Fiegler, H.; Langford, C. F.; Carter, N. P.; Vetrie, D.; Snyder, M.; Euskirchen, G.; Urban, A. E.; Nagalakshmi, U.; Rinn, J.; Popescu, G.; Bertone, P.; Hartman, S.; Rozowsky, J.; Emanuelsson,

O.; Royce, T.; Chung, S.; Gerstein, M.; Lian, Z.; Lian, J.; Nakayama, Y.; Weissman, S.; Stoic, V.; Tongprasit, W.; Sethi, H.; Jones, S.; Marra, M.; Shin, H.; Schein, J.; Clamp, M.; Lindblad-Toh, K.; Chang, J.; Jaffe, D. B.; Kamal, E. S.; Lander, E. S.; Mikkelsen, T. S.; Vinson, J.; Zody, M. C.; de Jong, P. J.; Osoegawa, K.; Nefedov, M.; Zhu, B.; Baxevanis, A. D.; Wolfsberg, T. G.; Collins, F. S.; Crawford, G. E.; Holt, E.; Vasicek, T. J.; Zhou, D.; Luo, S.; Green, E. D.; Bouffard, G. G.; Margulies, E. H.; Portnoy, M. E.; Hansen, N. F.; Thomas, P. J.; Mcdowell, J. C.; Maskeri, B.; Young, A. C.; Idol, J. R.; Blakesley, R. W.; Schuler, G.; Miller, W.; Hardison, R.; Elnitski, L.; Shah, P.; Salzberg, S. L.; Pertea, M.; Majoros, W. H.; Haussler, D.; Thomas, D.; Rosenbloom, K. R.; Clawson, H.; Siepe, A.; Kent, W. J.; Weng, Z.; Jin, S.; Halees, A.; Burden, H.; Karaoz, U.; Fu, Y.; Yu, Y.; Ding, C.; Cantor, C. R.; Kingston, R. E.; Dennis, J.; Green, R. D.; Singer, M. A.; Richmond, T. A.; Norton, J. E.; Farnham, P. J.; Oberley, M. J.; Inman, D. R.; McCormick, M. R.; Kim, H.; Middle, C. L.; Pirrung, M. C.; Fu, X. D.; Kwon, Y. S.; Ye, Z.; Dekker, J.; Tabuchi, T. M.; Gheldof, N.; Dostie, J.; Harvey, S. C.; Consortium, E. P., The ENCODE (ENCyclopedia of DNA elements) Project. *Science* **2004,** *306* (5696), 636-640.

[139] Birney, E.; Stamatoyannopoulos, J. A.; Dutta, A.; Guigo, R.; Gingeras, T. R.; Margulies, E. H.; Weng, Z. P.; Snyder, M.; Dermitzakis, E. T.; Stamatoyannopoulos, J. A.; Thurman, R. E.; Kuehn, M. S.; Taylor, C. M.; Neph, S.; Koch, C. M.; Asthana, S.; Malhotra, A.; Adzhubei, I.; Greenbaum, J. A.; Andrews, R. M.; Flicek, P.; Boyle, P. J.; Cao, H.; Carter, N. P.; Clelland, G. K.; Davis, S.; Day, N.; Dhami, P.; Dillon, S. C.; Dorschner, M. O.; Fiegler, H.; Giresi, P. G.; Goldy, J.; Hawrylycz, M.; Haydock, A.; Humbert, R.; James, K. D.; Johnson, B. E.; Johnson, E. M.; Frum, T. T.; Rosenzweig, E. R.; Karnani, N.; Lee, K.; Lefebvre, G. C.; Navas, P. A.; Neri, F.; Parker, S. C. J.; Sabo, P. J.; Sandstrom, R.; Shafer, A.; Vetrie, D.; Weaver, M.; Wilcox, S.; Yu, M.; Collins, F. S.; Dekker, J.; Lieb, J. D.; Tullius, T. D.; Crawford, G. E.; Sunyaev, S.; Noble, W. S.; Dunham, I.; Dutta, A.; Guigo, R.; Denoeud, F.; Reymond, A.; Kapranov, P.; Rozowsky, J.; Zheng, D. Y.; Castelo, R.; Frankish, A.; Harrow, J.; Ghosh, S.; Sandelin, A.; Hofacker, I. L.; Baertsch, R.; Keefe, D.; Flicek, P.; Dike, S.; Cheng, J.; Hirsch, H. A.; Sekinger, E. A.; Lagarde, J.; Abril, J. F.; Shahab, A.; Flamm, C.; Fried, C.; Hackermuller, J.; Hertel, J.; Lindemeyer, M.; Missal, K.; Tanzer, A.; Washietl, S.; Korbel, J.; Emanuelsson, O.; Pedersen, J. S.; Holroyd, N.; Taylor, R.; Swarbreck, D.; Matthews, N.; Dickson, M. C.; Thomas, D. J.; Weirauch, M. T.; Gilbert, J.; Drenkow, J.; Bell, I.; Zhao, X.; Srinivasan, K. G.; Sung, W. K.; Ooi, H. S.; Chiu, K. P.; Foissac, S.; Alioto, T.; Brent, M.; Pachter, L.; Tress, M. L.; Valencia, A.; Choo, S. W.; Choo, C. Y.; Ucla, C.; Manzano, C.; Wyss, C.; Cheung, E.; Clark, T. G.; Brown, J. B.; Ganesh, M.; Patel, S.; Tammana, H.; Chrast, J.; Henrichsen, C. N.; Kai, C.; Kawai, J.; Nagalakshmi, U.; Wu, J. Q.; Lian, Z.; Lian, J.; Newburger, P.; Zhang, X. Q.; Bickel, P.; Mattick, J. S.; Carninci, P.; Hayashizaki, Y.; Weissman, S.; Dermitzakis, E. T.; Margulies, E. H.; Hubbard, T.; Myers, R. M.; Rogers, J.; Stadler, P. F.; Lowe, T. M.; Wei, C. L.; Ruan, Y. J.; Snyder, M.; Birney, E.; Struhl, K.; Gerstein, M.; Antonarakis, S. E.; Gingeras, T. R.; Brown, J. B.; Flicek, P.; Fu, Y. T.; Keefe, D.; Birney, E.; Denoeud, F.; Gerstein, M.; Green, E. D.; Kapranov, P.; Karaoz, U.; Myers, R. M.; Noble, W. S.; Reymond, A.; Rozowsky, J.; Struhl, K.; Siepel, A.; Stamatoyannopoulos, J. A.; Taylor, C. M.; Taylor, J.; Thurman, R. E.; Tullius, T. D.; Washietl, S.; Zheng, D. Y.; Liefer, L. A.; Wetterstrand, K. A.; Good, P. J.; Feingold, E. A.; Guyer, M. S.; Collins, F. S.; Margulies, E. H.; Cooper, G. M.; Asimenos, G.; Thomas, D. J.; Dewey, C. N.; Siepel, A.; Birney, E.; Keefe, D.; Hou, M. M.; Taylor, J.; Nikolaev, S.; Montoya-Burgos, J. I.; Loytynoja, A.; Whelan, S.; Pardi, F.; Massingham, T.; Brown, J. B.; Huang, H. Y.; Zhang, N. R.; Bickel, P.; Holmes, I.; Mullikin, J.

C.; Ureta-Vidal, A.; Paten, B.; Seringhaus, M.; Church, D.; Rosenbloom, K.; Kent, W. J.; Stone, E. A.; Gerstein, M.; Antonarakis, S. E.; Batzoglou, S.; Goldman, N.; Hardison, R. C.; Haussler, D.; Miller, W.; Pachter, L.; Green, E. D.; Sidow, A.; Weng, Z. P.; Trinklein, N. D.; Fu, Y. T.; Zhang, Z. D. D.; Karaoz, U.; Barrera, L.; Stuart, R.; Zheng, D. Y.; Ghosh, S.; Flicek, P.; King, D. C.; Taylor, J.; Ameur, A.; Enroth, S.; Bieda, M. C.; Koch, C. M.; Hirsch, H. A.; Wei, C. L.; Cheng, J.; Kim, J.; Bhinge, A. A.; Giresi, P. G.; Jiang, N.; Liu, J.; Yao, F.; Sung, W. K.; Chiu, K. P.; Vega, V. B.; Lee, C. W. H.; Ng, P.; Shahab, A.; Sekinger, E. A.; Yang, A.; Moqtaderi, Z.; Zhu, Z.; Xu, X. Q.; Squazzo, S.; Oberley, M. J.; Inman, D.; Singer, M. A.; Richmond, T. A.; Munn, K. J.; Rada-Iglesias, A.; Wallerman, O.; Komorowski, J.; Clelland, G. K.; Wilcox, S.; Dillon, S. C.; Andrews, R. M.; Fowler, J. C.; Couttet, P.; James, K. D.; Lefebvre, G. C.; Bruce, A. W.; Dovey, O. M.; Ellis, P. D.; Dhami, P.; Langford, C. F.; Carter, N. P.; Vetrie, D.; Kapranov, P.; Nix, D. A.; Bell, I.; Patel, S.; Rozowsky, J.; Euskirchen, G.; Hartman, S.; Lian, J.; Wu, J. Q.; Urban, A. E.; Kraus, P.; Van Calcar, S.; Heintzman, N.; Kim, T. H.; Wang, K.; Qu, C. X.; Hon, G.; Luna, R.; Glass, C. K.; Rosenfeld, M. G.; Aldred, S. F.; Cooper, S. J.; Halees, A.; Lin, J. M.; Shulha, H. P.; Zhang, X. L.; Xu, M. S.; Haidar, J. N. S.; Yu, Y.; Birney, E.; Weissman, S.; Ruan, Y. J.; Lieb, J. D.; Iyer, V. R.; Green, R. D.; Gingeras, T. R.; Wadelius, C.; Dunham, I.; Struhl, K.; Hardison, R. C.; Gerstein, M.; Farnham, P. J.; Myers, R. M.; Ren, B.; Snyder, M.; Thomas, D. J.; Rosenbloom, K.; Harte, R. A.; Hinrichs, A. S.; Trumbower, H.; Clawson, H.; Hillman-Jackson, J.; Zweig, A. S.; Smith, K.; Thakkapallayil, A.; Barber, G.; Kuhn, R. M.; Karolchik, D.; Haussler, D.; Kent, W. J.; Dermitzakis, E. T.; Armengol, L.; Bird, C. P.; Clark, T. G.; Cooper, G. M.; de Bakker, P. I. W.; Kern, A. D.; Lopez-Bigas, N.; Martin, J. D.; Stranger, B. E.; Thomas, D. J.; Woodroffe, A.; Batzoglou, S.; Davydov, E.; Dimas, A.; Eyras, E.; Hallgrimsdottir, I. B.; Hardison, R. C.; Huppert, J.; Sidow, A.; Taylor, J.; Trumbower, H.; Zody, M. C.; Guigo, R.; Mullikin, J. C.; Abecasis, G. R.; Estivill, X.; Birney, E.; Bouffard, G. G.; Guan, X. B.; Hansen, N. F.; Idol, J. R.; Maduro, V. V. B.; Maskeri, B.; McDowell, J. C.; Park, M.; Thomas, P. J.; Young, A. C.; Blakesley, R. W.; Muzny, D. M.; Sodergren, E.; Wheeler, D. A.; Worley, K. C.; Jiang, H. Y.; Weinstock, G. M.; Gibbs, R. A.; Graves, T.; Fulton, R.; Mardis, E. R.; Wilson, R. K.; Clamp, M.; Cuff, J.; Gnerre, S.; Jaffe, D. B.; Chang, J. L.; Lindblad-Toh, K.; Lander, E. S.; Koriabine, M.; Nefedov, M.; Osoegawa, K.; Yoshinaga, Y.; Zhu, B. L.; de Jong, P. J.; Consortium, E. P., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007,** *447* (7146), 799-816.

[140] Thomas, D. J.; Rosenbloom, K. R.; Clawson, H.; Hinrichs, A. S.; Trumbower, H.; Raney, B. J.; Karolchik, D.; Barber, G. P.; Harte, R. A.; Hillman-Jackson, J.; Kuhn, R. M.; Rhead, B. L.; Smith, K. E.; Thakkapallayil, A.; Zweig, A. S.; Haussler, D.; Kent, W. J., The ENCODE project at UC Santa Cruz. *Nucleic Acids Res* **2007,** *35*, D663-D667.

[141] Celniker, S. E.; Dillon, L. A. L.; Gerstein, M. B.; Gunsalus, K. C.; Henikoff, S.; Karpen, G. H.; Kellis, M.; Lai, E. C.; Lieb, J. D.; MacAlpine, D. M.; Micklem, G.; Piano, F.; Snyder, M.; Stein, L.; White, K. P.; Waterston, R. H.; Consortium, m., Unlocking the secrets of the genome. *Nature* **2009,** *459* (7249), 927-930.

[142] Tilesi, F.; Fradiani, P.; Socci, V.; Willems, D.; Ascenzioni, F., Design and validation of siRNAs and shRNAs. *Curr Opin Mol Ther* **2009,** *11* (2), 156-164.

[143] Mohr, S.; Bakal, C.; Perrimon, N., Genomic Screening with RNAi: Results and Challenges. *Annu Rev Biochem* **2010,** *79*, 37-64.

[144] Boutros, M.; Bras, L. P.; Huber, W., Analysis of cell-based RNAi screens. *Genome Biol* **2006,** *7* (7).

[145] Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009,** *4* (1), 44-57.

[146] Bushman, F. D.; Malani, N.; Fernandes, J.; D'Orso, I.; Cagney, G.; Diamond, T. L.; Zhou, H. L.; Hazuda, D. J.; Espeseth, A. S.; Konig, R.; Bandyopadhyay, S.; Ideker, T.; Goff, S. P.; Krogan, N. J.; Frankel, A. D.; Young, J. A. T.; Chanda, S. K., Host Cell Factors in HIV Replication: Meta-Analysis of Genome-Wide Studies. *Plos Pathog* **2009,** *5* (5).

[147] Beller, M.; Sztalryd, C.; Southall, N.; Bell, M.; Jackle, H.; Auld, D. S.; Oliver, B., COPI Complex Is a Regulator of Lipid Homeostasis. *Plos Biol* **2008,** *6* (11), 2530-2549.

[148] Tu, Z.; Argmann, C.; Wong, K. K.; Mitnaul, L. J.; Edwards, S.; Sach, I. C.; Zhu, J.; Schadt, E. E., Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. *Genome research* **2009,** *19* (6), 1057-67.

[149] Schlabach, M. R., Cancer proliferation gene discovery through functional genomics (vol 319, pg 620, 2008). *Science* **2008,** *320* (5874), 316-316.

[150] Bakal, C.; Linding, R.; Llense, F.; Heffern, E.; Martin-Blanco, E.; Pawson, T.; Perrimon, N., Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science* **2008,** *322* (5900), 453-456.

Subject Index

D

E

F

N

O

P

Q

R

S

T

Zaheer Ul-Haq

Dr. Zaheer Ul-Haq is directing a computational chemistry group at the Dr. Panjwani Center for Molecular Medicine and Drug research (PCMD), University of Karachi. He obtained his PhD under the supervision of Prof. Atta-ur-Rahman and completed his post-doctoral studies on computation chemistry with Prof. Bernd M. Rode in Innsbruck, Austria. He is a recipient of Fulbright and Humboldt Fellowship from USA and Germany, respectively. Dr. Zaheer also received Gold medal in Chemistry from Pakistan Science Foundation. He has published over 80 research articles in top international journals of computational chemistry. His area of interest includes designing new bio-active compounds using in silico tools, generation and screening of large commercially available compounds, and Molecular Dynamics (MD) simulation of bio-molecules. He is currently serving as an editorial board member to the Journal of Molecular Graphics and Modelling.

Jeffry D. Madura

Jeffry is the Lambert F. Minucci Endowed Chair in Engineering and Computational Sciences and Professor in the Department of Chemistry and Biochemistry at Duquesne University. He earned a B.A. from Thiel College, a Ph.D. in Physical Chemistry from Purdue University and was a postdoctoral fellow at the University of Houston. His research interests include the development and application of biomolecular simulation software, the study of neurotransmitter transporters, the electronic structure of solid-state materials, and the thermoresponsive behavior of smart polymers. He has published 100+ peer-reviewed papers in physical chemistry and received over $6M in external research funding. He was a recipient of a Dreyfus Teacher-Scholar Award, was the chair of the ACS COMP Division and is an ACS Fellow. Dr. Madura received the 2014 ACS Pittsburgh Local Section Award. He is a co-author to the textbook titled "General Chemistry: Principles and Modern Applications" as well as a co-author to a physical chemistry solutions manual. He received the Bayer School of Natural and Environmental Sciences and the Duquesne University Presidential Award for Excellence in Scholarship in 2007 and the Bayer School of Natural and Environmental Sciences Award for Excellence in Service in 2004. He is currently co-editor to the Journal of Molecular Graphics and Modelling.