

Integrating Chemical Space Analysis and QSAR Modeling with Virtual Screening Towards the Identification of Novel 5-LOX Inhibitors

*Thesis submitted to the
University of Calicut
in partial fulfillment of the requirements
for the award of the degree of*

*Doctor of Philosophy
in
Chemistry*

by

SHAMEERA AHAMED T.K.



**DEPARTMENT OF CHEMISTRY
UNIVERSITY OF CALICUT
KERALA-673635
JULY 2020**

CERTIFICATE

Certified that the research work embodied in the thesis entitled **“Integrating Chemical Space Analysis and QSAR Modeling with Virtual Screening Towards the Identification of Novel 5-LOX Inhibitors”** has been carried out by **Shameera Ahamed T.K.** under my supervision at the Department of Chemistry, University of Calicut, Kerala and the same has not been submitted elsewhere previously for the award of any other degree or diploma.

University of Calicut

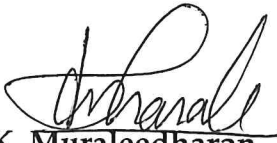
Dr. K. Muraleedharan
Professor
Department of Chemistry
University of Calicut
Kerala

CERTIFICATE

Certified that the research work embodied in the thesis entitled “**Integrating Chemical Space Analysis and QSAR Modeling with Virtual Screening Towards the Identification of Novel 5-LOX Inhibitors**” has been carried out by **Shameera Ahamed T.K.** under my supervision at the Department of Chemistry, University of Calicut, Kerala and the same has not been submitted elsewhere previously for the award of any other degree or diploma.

I also hereby certify that the corrections/suggestions from the adjudicators have been incorporated in the revised thesis. Content of the CD submitted and the hardcopy of the thesis is one and the same.

University of Calicut


Dr. K. Muraleedharan
Professor
Dept. of Chemistry
University of Calicut
Kerala-673635
Professor
Department of Chemistry
University of Calicut
Kerala

DECLARATION

I hereby declare that the research work embodied in the thesis entitled “**Integrating Chemical Space Analysis and QSAR Modeling with Virtual Screening Towards the Identification of Novel 5-LOX Inhibitors**”, submitted to the University of Calicut is a bonafide record of research work done by me during the period 2015-2020 under the supervision and guidance of Dr. K. Muraleedharan, Professor, Department of Chemistry, University of Calicut. The same has not been submitted elsewhere for any degree or diploma. I also declare that I am solely responsible for the genuineness of the findings/observations pertaining to these studies to complete this thesis.

University of Calicut

Shameera Ahamed T.K.

Acknowledgments

In pursuit of this academic endeavor, I feel especially grateful for the motivation, encouragement, and support that came to my way in abundance, and it is an honor for me to acknowledge the same in words.

Firstly, I would like to express my sincere gratitude to my advisor Prof. K. Muraleedharan, Department of Chemistry, University of Calicut, for his patience, motivation, and constructive criticism throughout my research work. I consider myself fortunate in that it would have been impossible to achieve this goal without his support and care.

With a deep sense of gratitude, I copiously thank Dr. A. I. Yahya, Head of the Department, and former Head, Prof. P. Raveendran, for providing me adequate facilities to carry out this research work.

I would like to express my special thanks to Prof. V. M. Abdul Mujeeb for his encouragement and ardent interest shown during the progress of work. I would also like to thank the faculties of the Department, Dr. Abraham Joseph, Dr. N. K. Renuka, Dr. P. Pradeepan, Dr. M. T. Ramesan, Dr. Suresh Babu and other Guest faculty for their kind support for fostering a competing science environment in the Department.

I convey my regards to all the technical and ministerial staff of the Department for their help throughout the course.

I gratefully acknowledge the Council of Scientific and Industrial Research (CSIR) for financial assistance. Also, I would like to express my sincere thanks to the University of Calicut and Central Sophisticated Instrumentation Facility (CSIF) for providing the computational facility.

I would also like to express gratefulness to all the past and present research scholars of computational chemistry group Dr. Kavitha P., Dr. Sindhu N.V., Dr. Sarada K., Dr. Nusrath K., Dr. Vijisha K. Rajan, Dr. Ajmala Shireen P., Ms. Jaseela M.A., Ms. Sabira K., Ms. Sumayya P.C., Ms. Jalala V.K., Ms. Vinduja P., Ms. Neenu Krishna P.U., Ms. Swathi Krishna and friends from different research groups Dr. Rajeena P., Ms. Shamsiya A., Dr. Safna Hussan K.P, Ms. Nibila T.A. for their immense support.

I am indebted to my family, especially my umma and uppa, who always dreamt that I reach golden heights, and to my sisters, brothers, niece and nephews and in-laws for bearing me through all the hardships. And most of all, I owe my deepest gratitude to my husband, Mr. Haris A.K., for his affection, encouragement, understanding, and patience. He supported me without any complaint or regret that enabled me to complete my Ph. D thesis.

Apart from the ones mentioned here, I record my warm thanks to all those who have helped me directly or indirectly in making this thesis successful.

Shameera Ahamed T.K.

*In memory of my father
To my mother
With love and eternal appreciation*

PREFACE

The collaboration between chemistry, biology, and medicine has been remarkably productive over the past century. As a result, the emerging discipline like cheminformatics and computer-aided drug design (CADD) are finding increasing applications in the field of rational drug design. There are thousands of receptor or target proteins are getting identified and characterized as a druggable target for varieties of diseases. Their structure-activity information is also available in public databases. Thus, finding a hidden relationship among this information and uncovering the biological activities of compounds against specific proteins and their chemical structures using computational methods is a preliminary interest in rational drug design.

Pharmacological intervention of 5-lipoxygenase (5-LOX) catalyzed leukotriene biosynthesis has been extensively studied as a promising therapeutic strategy for acute inflammatory, allergic, and respiratory diseases. Due to the toxicity effect of marketed 5-LOX inhibitor zileuton, the scientific community is seeking novel 5-LOX inhibitors. As a result, the significant and relevant amount of structure-activity information of 5-LOX inhibitors has been released and stored in public databases. Computational methods are widely used for the rapid and efficient identification and prediction of potent therapeutic agents against 5-LOX protein using the structure and ligand-based

approaches. So, in this study, we have done a thorough cheminformatic characterization, modeling, and screening studies on 5-LOX inhibitors.

The thesis entitled “Integrating Chemical Space Analysis and QSAR Modeling with Virtual Screening Towards the Identification of Novel 5-LOX Inhibitors” is organized into eight chapters. Chapter 1 provides a brief introduction about chemical space as a source for new drugs, various computational methods for rational drug design, and the detailed description of target 5-LOX and its therapeutic relevance. Besides, this Chapter presents the motivation and objectives of the current research problem. The details of the materials and computational methodology used for this study are discussed in Chapter 2. Chapter 3 presents the detailed report on selection, evaluation, and binding site identification of the crystal structure of Human 5-LOX protein and the development of the 3D model of the rat 5-LOX protein. Moreover, the detailed analysis of the docking result of known 5-LOX antagonists with the ligand-binding sites is given in this Chapter.

The comprehensive cheminformatic characterization of the diversity and complexity of the chemical space of 5-LOX and FLAP inhibitors by comparing it with the Approved drug space and LOX library is described in Chapter 4. Besides, SAR of the datasets is also present using activity landscape analysis and chemotype enrichment. In Chapter 5, the description of the development of robust and statistically significant CoMFA QSAR models to predict the 5-LOX inhibitory potency of redox inhibitors is given. In Chapter 6, we have discussed the development of the QSAR classification model by

incorporating all the complex, diverse structural scaffold and related bioactivity data of 5-LOX inhibitors using non-linear machine learning algorithms. Chapter 7 discusses the virtual screening strategies employed for the successful screening of 2.7 million compounds from ZINC15 databases. Conclusion of the major findings of the studies and some suggestion for future research is given in Chapter 8.

CONTENTS

CHAPTER 1	INTRODUCTION	1-47
1.1	Chemical space and Libraries	1
1.1.1	<i>Biologically Relevant Chemical Space (BRCS)</i>	3
1.1.2	<i>Exploring BRCS for Drug Discovery</i>	3
1.1.3	<i>Target-Focused Chemical Libraries</i>	5
1.1.4	<i>Learning, Mining, Modelling and Screening the Chemical Libraries</i>	6
1.2	Computational Approaches in Chemical Modeling and Informatics	7
1.2.1	<i>Computational Chemistry</i>	7
1.2.2	<i>Cheminformatics</i>	8
1.2.3	<i>Computer-Aided Drug Design</i>	10
1.2.4	<i>Data Mining</i>	13
1.2.5	<i>Quantitative Structure-Activity Relationship (QSAR)</i>	16
1.2.6	<i>Machine Learning Techniques (MLT)</i>	20
1.2.7	<i>Virtual Screening (VS)</i>	22
1.3	Arachidonate 5-Lipoxygenase (5-LOX) and Its Importance	25
1.3.1	<i>Leukotrienes and 5-LOX</i>	25
1.3.2	<i>Biochemistry</i>	27
1.3.3	<i>5-LOX Catalytic Mechanism</i>	28
1.3.4	<i>Classification of 5-LOX Inhibitors</i>	30
1.3.5	<i>Computational Methods in the Search for Novel 5-LOX Inhibitors</i>	34
1.4	Scope of the Present Study	36
1.5	Objectives of the Present Study	36
CHAPTER 2	THEORETICAL AND METHODOLOGICAL OVERVIEW	49-88
2.1	Molecular Representation	49
2.2	Molecular Modeling	52
2.3	Molecular Descriptors and Fingerprints	53
2.4	Feature Selection Methods	58
2.5	Molecular Similarity	60

2.5.1	<i>Similarity Metrics</i>	60
2.6	Diversity Analysis	62
2.7	Chemical Space Analysis	62
2.7.1	<i>PCA</i>	63
2.7.2	<i>SOM</i>	63
2.8	3D-QSAR	64
2.8.1	<i>CoMFA</i>	65
2.8.2	<i>Molecular Alignment</i>	65
2.8.3	<i>CoMFA Procedure</i>	66
2.8.4	<i>Validation of CoMFA PLSR Model</i>	68
2.8.5	<i>Model Validation through Progressive Scrambling</i>	70
2.9	Machine Learning Modeling	70
2.9.1	<i>Machine Learning Algorithms</i>	71
2.9.2	<i>Model Performance Evaluation</i>	73
2.10	Homology Modeling	74
2.11	Molecular Docking	75
2.11.1	<i>Theory of Docking</i>	76
2.11.2	<i>Search / Sampling Algorithms</i>	76
2.11.3	<i>Scoring Functions</i>	77
2.12	Software	78
2.12.1	<i>Molecular Modeling Software</i>	78
2.12.2	<i>Molecular Editors and Visualization Tools</i>	79
2.12.3	<i>Datamining, Visualisation, QSAR Modeling software</i>	82
CHAPTER 3	PROTEIN STRUCTURE EVALUATION AND MODELING	89-116
3.1	Introduction	89
3.2	Evaluation of Human 5-LOX Protein Crystal Structure	90
3.3	Protein Preparation	92
3.4	Ligand Binding Site Identification	92
3.5	Druggability Assessment of 5-LOX Protein	96
3.6	Molecular Docking	100
3.7	Homology Modeling Studies	103
3.7.1	<i>Analysis of the Homology Models</i>	105
3.7.2	<i>Binding Site Investigation</i>	110
3.8	Conclusion	112
CHAPTER 4	CHEMICAL SPACE CHARACTERIZATION AND SAR ANALYSIS	117-176

4.1	Introduction	117
4.2	Compound Databases and Bioactivity Representations	119
4.3	Physicochemical Properties of 5-LOX Chemical Space	123
4.4	Visual Representation of the Property Space	129
4.5	Diversity Analysis	134
4.5.1	<i>Diversity based on PCP</i>	135
4.5.2	<i>Fingerprint Diversity</i>	136
4.5.3	<i>Molecular Scaffolds and Scaffold Diversity</i>	141
4.5.4	<i>Scaled Shannon Entropy (SSE)</i>	148
4.5.5	<i>Consensus Diversity Plot</i>	150
4.6	Structure-Activity Relationship	152
4.6.1	<i>Structure-Activity Similarity (SAS) Maps</i>	153
4.6.2	<i>Activity Cliff Generators and SAR Interpretation</i>	158
4.6.3	<i>Chemotype Enrichment</i>	166
4.7	Conclusion	169
CHAPTER 5 MODELING CoMFA BASED 3D-QSAR		177-222
5.1	Introduction	177
5.2	CoMFA on 3', 4'-dihydroxyflavones as Rat 5-LOX Inhibitors	178
5.2.1	<i>Dataset of Flavone Derivatives</i>	179
5.2.2	<i>Molecular Modeling and Alignment of Flavones</i>	182
5.2.3	<i>Statistical Analysis of CoMFA Models of Flavones</i>	183
5.2.4	<i>Graphical Interpretation of the CoMFA Contour Maps of Flavones</i>	188
5.2.5	<i>Docking Analyses of Flavones</i>	191
5.3	CoMFA on 3, 4-dihydrochalcones as Rat 5-LOX Inhibitors	194
5.3.1	<i>Dataset of Chalcone Derivatives</i>	195
5.3.2	<i>Molecular Modeling and Alignment of Chalcones</i>	197
5.3.3	<i>Statistical Analysis of CoMFA Models of Chalcone</i>	198
5.3.4	<i>Graphical Interpretation of the CoMFA Contour Maps of chalcones</i>	203
5.3.5	<i>Docking Analyses of Chalcones</i>	206
5.4	CoMFA of Benzoquinone Derivatives as Human 5-LOX Inhibitors	207

5.4.1	<i>Dataset of Benzoquinone Derivatives</i>	208
5.4.2	<i>Molecular Modeling and Alignment of Benzoquinones</i>	210
5.4.3	<i>Statistical Analysis of CoMFA Models of Benzoquinones</i>	211
5.4.4	<i>Graphical Interpretation of the CoMFA Contour Maps of Benzoquinones</i>	215
5.4.5	<i>Docking Analyses of Benzoquinones</i>	217
5.5	Conclusion	219
CHAPTER 6 MODELING MACHINE LEARNING BASED QSAR		223-267
6.1	Introduction	223
6.2	Dataset	224
6.3	Molecular Modeling and Descriptor Calculation	226
6.4	Chemical Space Characterization	229
6.5	Structure-Activity Landscape Analysis	234
6.6	Predictive Modeling Using Machine Learning Techniques	238
6.6.1	<i>Feature Selection</i>	239
6.6.2	<i>Performance Evaluation of Classifiers</i>	247
6.6.3	<i>Model Validation Through Y-Scrambling</i>	253
6.7	Virtual Screening of e-Drug3D Database	254
6.7.1	<i>Virtual Screening (VS)</i>	254
6.7.2	<i>Molecular Docking Analysis</i>	258
6.8	Conclusion	263
CHAPTER 7 DOCKING-BASED VIRTUAL SCREENING OF SMALL MOLECULE INHIBITORS		269-311
7.1	Introduction	269
7.2	Screening Database	271
7.3	Protein Selection and Preparation	273
7.4	Molecular Docking Programs	274
7.5	Performance Evaluation of Docking Programs	279
7.5.1	<i>Actives and Decoys Selection</i>	279
7.5.2	<i>Evaluation Metrics</i>	280
7.6	Application of Virtual Machines in Virtual Screening	288
7.6.1	<i>Application of VMs in Research</i>	289
7.6.2	<i>Engineering Infrastructure</i>	290
7.6.3	<i>Benchmarking Computational Power</i>	290
7.7	Virtual Screening of ZINC 15 Database	292

7.8	Binding Free Energy Calculation	294
7.9	Interaction of Virtual Hits at the Active Site	298
7.10	ADME Property and Toxicity Analysis	301
7.11	Conclusion	305
CHAPTER-8	CONCLUSION AND FUTURE OUTLOOK	313-317

1

INTRODUCTION

1.1. Chemical Space and Libraries

Synthetic chemistry is all about the production of varieties of compounds by making covalent bonds between atoms. Hundreds of molecules are getting synthesized and reported day by day. Recently, combinatorial chemistry and High-Throughput Screening (HTS) methods allowed the simultaneous synthesis of thousands to millions of compounds. Besides, advances in the speed, storage capacity, and processing power of computers have permitted scientists to develop trillions of molecules in virtual chemical libraries. The constant increase in the number of molecules that are either real or virtual raises the question of how many molecules exist in the world [1]. In order to conceptualize the total number of molecules, either real or virtual, the concept of ‘chemical space’ is used.

Lipinski and Hopkins compared the chemical space with the cosmic space and state that “chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars”[2]. However, by generalized definition, chemical space is defined as “the set of all possible molecular structures.” In cheminformatics, “A chemical space

is a multi-dimensional descriptor space, *i.e.*, it is spanned by a particular choice of (molecular) descriptors and the limits placed on them"[3]. It is widely accepted that the chemical space is vast and that there are only a small fraction of molecules that are known. There are no precise techniques available for calculating this space's exact size. The chemical space has been developed from enumerating acyclic hydrocarbons in the 1800s to the recent assembly of the chemical universe database GDB [4]. Several open-access databases of compounds that are currently accessible to the public in which the structures are written as SMILES [5–7], or related formats such as InChI [8] and some are given in Table 1.

Table 1.1 Various open-access databases of compounds with compounds size and their web address

Database	Description	Size	Web address
PubChem [9]	Known molecules from various public sources. world's largest collection of freely accessible chemical information	252 M	http://pubchem.ncbi.nlm.nih.gov
Chemspider [10]	An online resource from the Royal Society of Chemistry	67 M	http://www.chemspider.com/
ZINC [11]	Database of commercially-available compounds for virtual screening	230 M	http://zinc.docking.org
NCI Open [12]	Anticancer and AIDS compounds with screening data	0.25 M	http://cactus.nci.nih.gov/ncidb2.1
ChemDB [13]	Commercially available small molecules from 150 chemical vendors	5 M	http://cdb.ics.uci.edu
BindingDB [14]	Public web-accessible bioactive molecules with binding affinity data	0.78 M	http://www.bindingdb.org
ChemBank [15]	small molecules annotated with screening data	1.6 M	http://chembank.broadinstitute.org/
ChEMBL [16]	Small molecules with experimental data	1.8 M	https://www.ebi.ac.uk/chembl/db
DrugBank [17]	Experimental and approved small molecule drugs	0.0065 M	http://www.drugbank.ca

1.1.1. Biologically Relevant Chemical Space (BRCS)

Biologically relevant chemical space is defining as the areas of chemical space that possess biologically active compounds for a specific target or target class. They can, therefore, modulate a given biological system and then influence the development and progression of the disease. This space has a statistically definable Physico-chemical property limit, occupy discrete 'pockets' within chemical spaces. So, it is evident that, at least in terms of the number of compounds, 'biologically relevant chemical space' is only a small fraction of the total chemical space. Some estimates of this number are more than 10^{60} with a molecular weight of less than 500, which is the result of a thought experiment constructing molecules with 30 atoms by Bohacek *et al.* [18]. On the other hand, the "known drug space," *i.e.*, molecules that can be an active substance with a desirable effect against a specific target, is estimated to be only $1.1-2.0 \times 10^6$ molecules [19]. Databases of enumerated drug-like chemicals such as the Generated Chemical universe Database (GDB) [20] and the small molecule universe (SMU) [21] are also developed to explore more diverse drug-like chemical space.

1.1.2. Exploring BRCS for Drug Discovery

The discovery of novel therapeutic agents requires looking into a diverse and expanded chemical space. The random quest for bioactive compounds in the entire chemical space without prior information is equivalent to the hunt for "a needle in a haystack" scenario. Any viable method that can reduce the cost or time required

or that can improve the drug candidate success rate deserves attention from all parties involved, including the pharmaceutical industry and the regulatory government bodies.

For research in medicinal chemistry, it is crucial to identify particular molecules that fall within the biologically relevant category of large chemical spaces. By comparing to the total space, only a small proportion of the vast chemical space is determined to be biologically active and is, therefore, essential for drug development. However, even if we restrict ourselves to the chemical space of small molecules, it is evident from its massive size that it is not an easy task to explore the BRCS. The common strategy to avoid this problem is to set certain constraints in the search space so that the search space can be restricted. One of these ideal constraints is the physicochemical properties of molecules as they are single numerical quantities that can be experimentally calculated for real molecules and computed for virtual ones effectively. Also, these properties act as measures of bioavailability. There were also made several attempts to define a global coordinate system for molecules within the BRCS, in order to explore the chemical space more effectively. The first one is Molecular Quantum Numbers (MQN) [22] uses integer properties such as counting of atoms of different types, bonds of different types, hydrogen bond donors and acceptors, charges, and small topological features. The second example is Chemical Global Positioning System (ChemGPS) [23] uses molecule descriptors based on size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity. The third is a new and intuitive visualization tool based on

ligand–receptor interactions (LiRIf), which introduced for guiding medicinal chemists in analyzing the R-groups from a congeneric series [24]. In this way, using a particular choice of descriptors, many more chemical spaces can be defined.

1.1.3. Target-Focused Chemical Libraries

Concentrating chemical libraries on a specific target group is an attractive alternative for improving the process of hit identification in drug discovery. The growing number of structural data relating to targets and their small molecule inhibitors has contributed to the development of focused structural and activity databases. These libraries are generally known as target-focused chemical libraries. Universal chemical libraries are typically large, while the focused libraries are fairly small, usually with thousands of compounds. The development of target-focused chemical libraries needs 3D structural knowledge of targets and a set of known active small molecules. Focused chemical libraries are typically built using cheminformatics tools based on ligand knowledge or target information, or maybe even from both, and are produced by screening compounds in a wider and diverse set of compounds. Various approaches to visualizing and comparing large chemical databases have been developed, and some are described in Chapter 2.

1.1.4. Learning, Mining, Modelling and Screening the Chemical Libraries

The developments in HTS have contributed to a wide range of systemic activity datasets that have been integrated into commercial, open, and public repositories such as ChEMBL and PubChem over the last decade. The HTS campaigns in pharmaceutical companies have accumulated a great deal of information of hundreds of millions of compounds over a couple of hundred assays and saved as target-focused chemical libraries. The emergence of academic HTS screening centers and more research articles published as well as the growing push towards academia for early-stage drug development indicate that computational informatics tools and methods are needed to gather and learn from such information. Although there have been few reports on a systematic approach for data mining that can efficiently extract relevant knowledge of the interest of chemists and biologists, it is common knowledge that comprehensive data is concealed within the vast amounts of data. Wide ranges of tools were developed by Collaborative Drug Discovery (CDD) [25] for storing, mining, secure and selective sharing and learning from these HTS data. Computational approaches, especially cheminformatics, have played an increasingly important role in the learning, mining, modeling, and screening the chemical libraries. Consistent developments in cheminformatics and bioinformatics tools are needed to explore and visualize the biologically relevant chemical space or target focused libraries and to identify interesting molecules that could have therapeutic potential.

1.2. Computational Approaches in Chemical Modeling and Informatics

The way we do science has changed by computers. Like any other field, its influence has made it easy to handle most of the chemical problems. With tremendous computational power and the immense amount of data, the technique and approach to solving a chemical problem are transforming into a degree of collaboration between research, theory, and data analysis. For decades, computer methods have played a role in theoretical and physical chemistry. The prediction of molecular properties and the theory testing was regularly based on computational models with the help of on theoretical principles of classical and quantum mechanical physics, and the name of the discipline used for this purpose is called computational chemistry. It provides a better understanding of the behavior of matter at its most fundamental level and helps in the calculation of molecular geometry, molecular energy, and transition states, chemical reactivity, rate, spectra, substrate-enzyme interaction, *etc.*

1.2.1. Computational Chemistry

Computational chemistry methods can generally be classified into two such as classical mechanical methods that include Molecular Mechanics (MM) and Molecular Dynamics (MD), and quantum mechanical methods that include *Ab initio*, Semi-Empirical (SE), and Density Functional Theory methods (DFT). MM is based on the classical model in which a molecule as a collection of balls (atoms) held together by springs (bonds), while MD is based on Newton's laws

of motion. A set of functions and constants, termed as a force field, is used in MM. *Ab initio* calculations are based solely on finding approximate solutions to Schrödinger wave equations, while SE methods included some empiric parameterization to solve Schrödinger wave equations. The most popular method, DFT, replaces the N-electron wave function with the simpler concept of ‘electron density.’ Choosing the most appropriate method for the task in question by evaluating the strengths and weaknesses of numerous computational tools is a serious challenge for computational chemists. Despite this, the work of computational chemists affects the way the world functions, helps manufacturers design more efficient and productive processes, aid to characterize new compounds, pharmaceuticals, and materials and assists researchers in gathering useful knowledge from mountains of data.

1.2.2. Cheminformatics

Our best interest, however, is in the manipulation of information on chemical structures and biological activity data rather than the calculation of the fundamental properties of molecules. In recent years, the term ‘Cheminformatics’ or ‘Chemoinformatics’ has been recognized as a distinct discipline in computational molecular sciences with a primary focus on analyzing/simulating/modeling/manipulating chemical information that can represent either in 2D structure or 3D structure and related metadata such as endpoints of biological activity and physicochemical properties [26,27]. Unlike conventional computational chemistry, cheminformatics emphasis on

practical issues. The term chemoinformatics has been introduced and defined by F.K. Brown only in 1998: "Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization" [28]. Nevertheless, for more than 30 years, the core principles behind cheminformatics, such as Quantitative Structure-Activity Relationships (QSARs) and compound property prediction, have been around.

Generally, cheminformatics methods follow an inductive learning strategy where knowledge is extracted from chemical information gained from lots of data on structures, activities, *etc.* Until recently, cheminformatics, which had a relatively small presence in academia or industry, was a fairly obscure discipline [29]. Computer-assisted synthesis, design, structure representation, and chemometrics are the first modules of cheminformatics [29–31]. With the emergence of high throughput drug screening and the need for millions of combined chemical libraries, cheminformatics is now playing a significant role in many areas of drug discovery and drug development. These techniques are also behind the terms "computer-aided molecular design" and "computer-aided drug design."

A vital goal of the cheminformatics is to understand the interaction of small molecules with their biological targets, thus leading to the identification of structural characteristics that determine the biological activities of the small molecules. Previously, we looked

at the expanding nature of the BRCS due to the vast amount of compound bioactivity data produced by HTS techniques. Navigating sub-spaces that represent compounds with desired activity in such a vast chemical space are a rather daunting task. Data mining and visualization approaches can navigate through chemical space and unveil the underlying patterns in spaces of chemical and pharmacological properties decisive for the discovery and development of drugs. So, cheminformatics currently encompasses a global range of computational methodologies, including compound mining, library design and optimization, molecular similarities and diversity analysis, chemical space analysis, chemical structure, and property prediction, QSAR, QSPR, and the like.

1.2.3. Computer-Aided Drug Design

The difficult issue of determining which compounds is to be druggable for a specific target has always been a concern for medicinal chemists. The process of marketing a new medicine is very costly, and that takes much time. The current estimates of costs range from \$500 million to \$2 billion. It can take up to 6-10 years for the approval process and has a low success rate due to lower-than-expected efficacies or higher-than-expected toxicities. At every stage of the drug design process, the chemist must choose from tens of millions of possible molecules. The development of computer algorithms and methods in the field of drug design reduces this difficulty. More and more computational methods have been introduced in the drug development process over the last decades, enabling researchers to

analyze the situation much more quickly and develop drugs within a limited time and cost. Software simulation, such as MD, Machine Learning (ML) *etc.*, models that predict drug candidate activity or toxicity is an evident approach to lowering the overall costs.

The term Computer-Aided Drug Discovery (CADD) arises and revolutionizes the area of drug design by identifying compounds with desirable characteristics, speed up the hit-to-lead process and improve the chances of getting your compound past the hurdles of preclinical testing and is rapidly gaining in popularity, implementation, and appreciation. Using certain computer-aided tools such as molecular modeling, simulation, and virtual screening now we can able to identify promising candidates before synthesis. Such methods are generally classified as either structure-based or ligand-based methods. Structure-Based Drug Designing (SBDD) [32,33] is the approach where the structural features of the drug target are used for the development of its inhibitor. In contrast, the Ligand-Based Drug Designing (LBDD) approach takes into account the knowledge of known inhibitors that bind to the biological target of interest and use this information to derive a pharmacophore model that describes the minimum necessary structural characteristics that a molecule must hold in order to bind to the target.

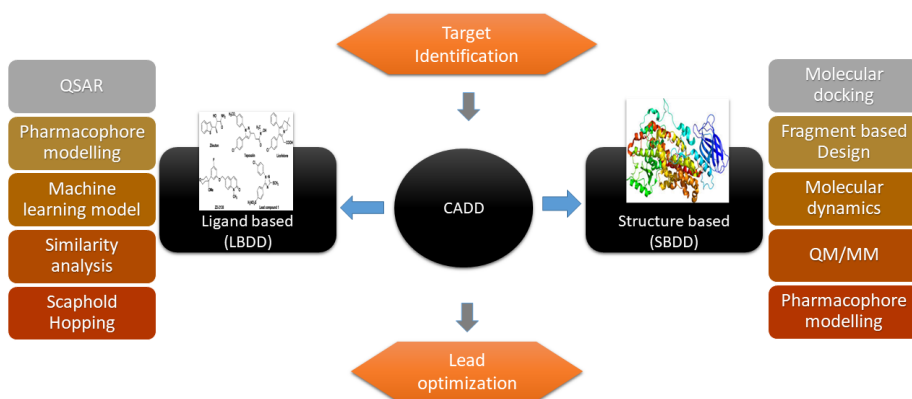


Fig. 1.1 Schematic representation of the CADD techniques.

Protein structures obtained from various techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), cryo-electron microscopy (EM), homology modeling, and molecular dynamic (MD) simulations are the primary requisite for SBDD [34–36]. Exploiting information from these 3D receptors to find small fragments that match well with the binding site is called De novo drug design. These fragments should be connected to ensure synthetic accessibility based on connection rules that provide a structurally novel scaffold that can be synthesized for further screening [37]. Where us, exploiting 3D receptor structural information to screen compounds with specific bioactivity from available large small-molecule libraries are called Virtual Screening (VS) [38,39]. Several SBDD methodologies to identify a drug molecule against a particular target have been developed over the past few years. Molecular docking, Fragment-based Design, Molecular dynamics, QM/MM, and Pharmacophore modeling are some of that have been used as SBDD strategy.

Molecular docking is one of the most well-known SBDD methods that predict possible binding modes of a compound in a particular target binding site and estimates affinity based on its conformation and complementarity with the features found in the binding pocket.

If there is a lack of a 3D protein structure, pharmacophore information from a group of ligands active against the specific target (receptor or enzyme) may be used to classify compounds either active or inactive. Classification can be done based on physicochemical and structural properties (molecular descriptors) that are responsible for observed biological activity. Here, it is assumed that structurally similar compounds exhibit similar biological response and difference in biological activity is due to structural differences [40]. Common LBDD techniques are QSAR, Pharmacophore modeling, ML model, Similarity analysis, and Scaffold Hopping and the rest.

1.2.4. Data Mining

We have already addressed the fact that the biologically relevant chemical space is expanding due to the unprecedented increase in the number of compounds and associated activity data stored in the public domain. Advances in Information technology and increasing automation also have a contribution to vast amounts of data being generated and collected. Compound activity data provide an important knowledge base for the discovery of drugs if data can efficiently be mined. Researchers now routinely screen millions of compounds with a technology known as virtual screening in the search for some that are biologically active. Structure-Activity Relations

(SARs) can also be systematically extracted for compounds that are active against current targets and used for compound design and optimization. Nevertheless, in addition to massively increasing amounts of compound data, diversity and complexity of chemical data are also increasing rapidly. Even though it is thought that advances in computational speed and capacity make analyzing extensive data much more feasible, it is impossible to explore such vast amounts of data and uncover useful patterns through traditional statistical methods and human analysis alone. Therefore, an innovative field is developed that uses a range of analytical and modeling techniques to find patterns and relationships in data and is known as data mining.

Data mining is a complex process for identifying novel, potentially useful, valid, and ultimately logical patterns in a dataset [41] and is referred to as a particular step in the process of Knowledge Discovery in Databases (KDD) [42]. Data mining has become a major area of research in cheminformatics, which contributes to the development of artificial intelligence (AI) and machine learning. Data mining techniques will reveal underlying patterns that are critical for drug discovery and production in chemical and pharmacological property spaces [41]. The tasks of data mining include pre-processing, clustering, classification, regression, visualization, and feature selection [43]. According to Feelders *et al.*, Data mining process has six significant steps [44] which are:

1. Problem definition.
2. Knowledge acquisition.
3. Data selection.
4. Data pre-processing.
5. Analysis and interpretation.
6. Reporting and use.

Predictive data mining is the most important common branch of data mining that has the most direct applications in science and business [45]. In predictive modeling, a model is generated to predict an outcome (dependent parameters), which is commonly bioactivity in drug discovery that can be either categorical (classification) or numerical (regression) developed based on one or more variables (independent parameters) like molecular properties. A subset of the dataset, the training set, is used to build predictive models, and a test set is used to validate the built model. The model evaluation, which is an integral part of the model development process, leads to the selection of the best model representing the data. Data mining models can provide a simple parametric equation derived from linear techniques and complex non-linear models derived from non-linear techniques. The primary applications of data-mining approaches in cheminformatics are VS and QSAR. Details of these techniques are provided in the following section.

1.2.5. Quantitative Structure-Activity Relationship (QSAR)

QSAR is one of the powerful ligand-based CADD techniques used to build computational or Mathematical models that attempt to find a statistically significant correlation between various molecular properties of a set of molecules with their experimentally known biological activity using a chemometric technique [46].

The development of QSAR as a practical drug design method started in the early 1960s with the works of Hansch and Free-Wilson's one or two-dimensional linear free-energy relationships. This model involves the correlation of various electronic, hydrophobic, and steric features with biologic activity. The journey of QSAR continued via Crammer's three-dimensional QSAR [47,48] to Hopfinger's fourth [49] and Vedani's fifth [50] and now reached sixth-dimensions [51]. Over the years, several effective QSAR models that cover a wide range of biological and physicochemical properties have been published.

The diverse applications of QSAR extend not only in science but also in industry, academia, and government (regulatory) agencies. QSAR models are also commonly used to determine the potential impacts of chemicals, materials, and nanomaterials on human health and ecological systems [52–55]. Nevertheless, the use of predictive models for lead optimization in drug discovery remains an essential field of active QSAR expansion, where a growing number of specialized tools and databases are being developed and validated [56].

Objectives of QSAR are to quantitatively correlate and recapitulate the relationships between trends in chemical structure alterations and respective changes in the biological endpoint. This then used to comprehend which chemical properties are most likely determinants for their biological activities and, after that, optimize the existing leads to improve their biological activities. QSAR could also be used to predict the biological activities of untested and sometimes yet unavailable compounds [57].

The use of QSAR techniques is based upon two underlying principles: (1) structurally similar compounds behave (activity) comparably under similar environmental conditions, and (2) behavioral (activity) differences among compounds are linked to structural and compositional variations. The predictor variables are usually described as “descriptors” (or features, attributes, independent variables, structural/compositional components, *etc.*), and the resultant response variables (*e.g.*, reactivity, toxicity or bioactivity) are termed as “activities” (or endpoints, dependent variables, *etc.*). QSAR is in the form of a mathematical model:

$$\text{Activity} = f(\text{physicochemical properties and/or structural properties}) + \text{error}$$

The error includes a model error (bias) and observational variability, that is, the variability in observations even on a correct model. QSAR becomes a useful alternative because of the following reasons: Conventional syntheses methods are expensive and time-consuming, biological assays are also too costly, often requiring time,

the sacrifice of animals, also, need compounds in pure forms. Drugs fail due to poor ADMET profiles at a later stage of drug discovery or even after the commercialization stage of drug development. All of these process exceedingly expensive [57].

Key steps of QSAR (Figure 1.2), including (i) Data set selection and molecular modeling (ii) descriptors extraction (ii) feature selection, (iii) model construction and (iv) statistical validation [58]. Details of these steps are provided in Chapter 2.



Fig.1.2 General Steps involved in QSAR modeling

Classification of QSAR methodologies

Based on dimensionality - The QSAR methods are most often categorized according to the structural representation and how the descriptor values are derived, into the following classes [57]:

- 1D-QSAR: correlating biological activity with basic molecular property descriptors like HBD, HBA, log P, *etc.*
- 2D-QSAR: correlating biological activity with 2D descriptors derived from 2D structural patterns like connectivity indices, 2D-pharmacophores, *etc.*

- 3D-QSAR: correlating biological activity with non-covalent interaction fields like electrostatic, steric, and hydrophobic, so on.
- 4D-QSAR: Additionally, it includes an ensemble of ligand configurations in 3D-QSAR.
- 5D-QSAR: explicitly representing different induced-fit models in 4D-QSAR.
- 6D-QSAR: incorporating various solvation models in 5D-QSAR.

Based on the chemometric methods employed – QSAR approaches are also divided into two groups based on the type of correlation methodology used to create a relationship between structural properties and biological activity [57]:

- **Linear Model:** this model is based on the fact that only a linear relationship operates between a set of molecular descriptors and a specific biological activity. It includes Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Partial Least-Squares (PLS).
- **Non-linear model:** this model describes non-linear relationships in molecular descriptors and specific biological activity and consisting of Artificial Neural Networks (ANN), k-Nearest Neighbors (kNN), Bayesian neural nets, and so on.

Based on the nature of response property - QSAR is a statistical approach correlating the response property or activity data

with descriptors encoding chemical information. Such correlation may be derived either in a regression-based approach (in cases where the response property is quantitative and available on a continuous scale) or a classification-based approach (in cases where the response property is graded or semi-quantitative).

1.2.6. *Machine Learning Techniques (MLT)*

Machine learning is, in fact, an application of artificial intelligence that deals with the way machines learn from experience without being explicitly programmed. They are commonly used for grouping observations or instances into classes. Machine learning approaches are commonly divided into three, such as supervised learning, unsupervised learning, and reinforcement learning (Figure 1.3).

In supervised learning, the algorithm learns on a labeled dataset. That is, models are based on known inputs and their desired outputs given to the system. It requires two steps: first, creating a model based on known data and known responses using the correct algorithm, and second, predicting responses to new data based on the created model. Classification and regression are the two crucial supervised learning problems. In classification, the responses are categorical variables, while in regression responses are continuous variables. Unsupervised learning is where only input data and no associated response variables are available, *i.e.*, no output labels are given to the learning algorithm. Here, the grouping of data into different categories based on some measure of inherent similarity. Clustering and association are two forms of unsupervised learning

problems. In clustering, the inherent groupings in the given data are explored, whereas, in the association, the rules that describe large portions of the given data are explored. Reinforcement learning: It involves the interaction of a computer program with a dynamic environment in which it must perform a specific goal.

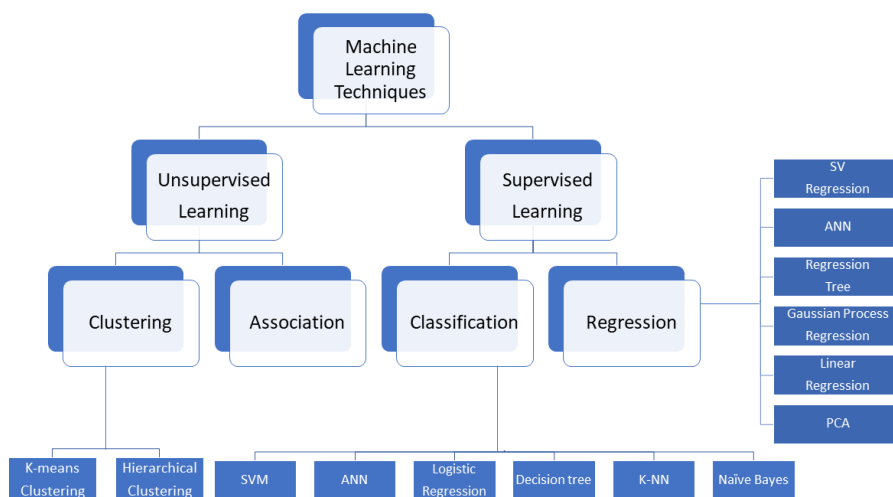


Fig.1.3 Various machine learning techniques

Machine learning techniques show outstanding performance in QSAR modeling, where the relationship between structure and activity is often complex and non-linear [59]. A wide range of machine learning algorithms has been used to build QSAR classification models from an input data set of molecular descriptors and activity labels. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers, *etc.*, are different machine learning techniques to solve a classification problem. Each technique adopts a learning algorithm to identify a

model that best fits the relationship between the descriptor set and the class label of the input data.

1.2.7. Virtual Screening (VS)

We have seen there are thousands of chemical ‘libraries’ and every library can contain, in theory, a considerable number of compounds—possibly billions. We also realized that there is this ‘virtual chemistry space’ which may contain 10^{100} potential molecules. So, there is an important question arises, how should a chemist filter the huge virtual chemistry space to find out the "right one"? Combinatorial Chemistry and HTS are facing challenges because of the need to miniaturize and automate as a means of controlling costs, saving time, minimizing resource-intensive, and reducing the volume of waste materials. So on this occasion, computer chemists have developed some computer programs that can automatically evaluate and screen extensive compound libraries using a specific method and algorithm [60]. This process is generally known as virtual screening. It is an alternative approach to HTS, where screening of large chemical libraries for potential hit compound candidates that would have a positive affinity towards a specific target of the known structure via computational method [61]. While searching throughout the chemical universe can be a theoretically interesting issue, more realistic VS possibilities concentrate on constructing, optimizing, and enriching targeted libraries of accessible compounds from in-house compound databases or vendor products.

With the accuracy of the VS method rising, it has now become an essential part of the drug discovery process; nonetheless, certain

issues remain major disadvantages and barriers to the reliability of the virtual screening programs. Inadequate structural information or incomplete information, ambiguous knowledge of the characteristics of drug-like molecules, failure to translate 3D properties to 2D structures, poor performance, poor scoring function, incorrect estimation of existing SAR data, and poor docking techniques also cause significant barriers in the virtual screening process [62]. Despite all these drawbacks, VS still hardcore *in silico* method that has been extensively used in drug discovery processes because it achieved a high level of precision and sensitivity in identifying known inhibitors of targets.

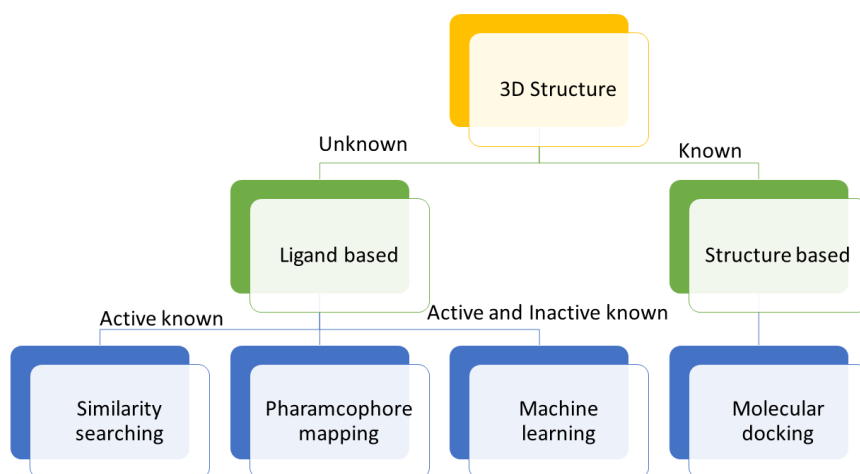


Fig. 1.4 Flow chart of Virtual Screening

Based on the information provided to the system, the VS have been traditionally subdivided to Structural-Based VS (SBVS) methods like those of molecular docking or Ligand-Based VS (LBVS) methods like similarity searching (Figure 1.4). In LBVS, molecular descriptors or fingerprints that are derived from 2D or 3D chemical structures of

known active or sometimes inactive molecules are often used to extract more active compounds with similar structures from a database through the application of similarity-searching techniques. Compounds with similar structures tend to target similar proteins are the underlying assumption of similarity-based screening techniques. Furthermore, searching through the hunt for a particular substructure, pharmacophore, or shape parameter that is the reason for its activity within the active set is also one of the ways of LBVS techniques. Another frequently used method for screening a vast number of small molecule inhibitors is ligand-based approaches using QSAR models. In structure-based or receptor-based virtual screening, compounds from screening library databases are docked into a ligand-binding site of target protein, predicting the most preferred orientation of one small molecule and bound to a target, resulting in a stable complex, then ranked using one or several scoring functions. The scoring function is used to calculate non-bonded interaction terms between the receptor and ligand atoms. The process can then be replicated employing various types of post-processing methods if deemed appropriate. Depending on the number of computer resources available and the type of target, the receptor and ligand flexibilities can be handled by various strategies and at different stages of the process.

The collaboration between chemistry, biology, and medicine has been remarkably productive over the past century. As a result, the emerging discipline like Cheminformatics and CADD and are finding increasing applications in the field of rational drug design. Drug discovery and development begins with the capture of the clinical

spectrum of diseases and the identification and validation of disease-causing target genes. Human Genome Project has uncovered novel functional pathways and therapeutic targets in several human diseases such as cancers and autoimmunity. There are thousands of receptor or target proteins are getting identified and characterized as a druggable target for varieties of diseases. Their structure-activity information is also available in public databases. Thus, finding various hidden relationships or logical patterns in this structure-activity data and based on it, uncovering the biological activities of novel compounds against specific proteins and using computational methods, is vital in rational drug design. Arachidonate 5-lipoxygenase (5-LOX) is an important enzyme that catalyzes the production of leukotrienes, a group of potent inflammatory mediators. It is a validated target for varieties of allergic and inflammatory conditions such as asthma and allergic rhinitis. The 5-LOX protein is, therefore, selectively chosen for our research. More details on these proteins and their importance as a therapeutic target are given in the following section.

1.3. Arachidonate 5-Lipoxygenase (5-LOX) and Its Importance

1.3.1. Leukotrienes and 5-LOX

Leukotrienes (LTs) are lipid mediators produced in leukocytes that are responsible for bodies' innate immunity and inflammatory responses and contain a conjugated triene as part of their structure. LTs use lipid signaling to transmit information either to the cells that produce it or to neighboring cells in order to control immune responses. Due to its function in the pathogenesis of acute and chronic inflammatory responses such as rheumatoid arthritis, gastroesophageal

reflux disease, atherosclerosis, inflammatory bowel disease, and autoimmune, LTs have received negative publicity in the last few years [63]. Besides, LTs are also involved in the development and progression of pulmonary inflammatory diseases such as Asthma and Rhinitis. Also, the role of LTs in cancer, such as leukemia, pancreas, prostate, and colon cancer, has been reported recently [64–67]. Pharmacological interventions to control the formation of LTs are, therefore, essential to reduce LT related diseases.

One of the main approaches for suppressing LT formation is the inhibition of protein 5-LOX enzymatic activity [68]. 5-LOX is an iron-containing non-heme dioxygenase enzyme that plays a vital role in LT biosynthesis. Mammalian LOXs are classified as 5-, 8-, 9-, 11-, 12- and 15-LOXs according to their positional specificity to oxygenate Arachidonic Acid (AA), a polyunsaturated fatty acid, from which LTS are getting synthesized by oxidative metabolism [68]. The substrate AA is generated near the cell wall surface by the enzyme phospholipase A2. Among all these LOXs, 5-lipoxygenase (5-LOX) has been established as the key enzyme for biosynthesis of LTs, which are significant mediators of inflammation and allergic responses [66]. 5-LOXs catalyzes the addition of molecular oxygen at position C-5 of 1, 4-cis–cis-pentadiene of AA to convert them into their hydroperoxy derivatives such as 5-hydroperoxy eicosatetraenoic acid (5-HPETE) and the subsequent dehydration of this compound to the short-lived epoxide leukotriene (LT) A₄ (Figure 1.5) [69]. LTA₄ is then metabolized to LTB₄, LTC₄, LTD₄, and LTE₄, which generally exert pro-inflammatory effects. LTD₄ and LTE₄ are referred to as the cysteinyl LTs (CysLTs) because of the presence of amino acid cysteine

and three conjugated double bonds, in contrast to the non-cysteine-containing LTB₄. LTB₄ potent inflammatory mediator, as well as a potent chemotactic agent and, is involved in leukocyte activation. LTC₄ and LTD₄ are powerful bronchial spasmogenic agents causing bronchoconstriction, airway edema, and mucus secretion. Pharmacological interruption of the 5-LOX pathway, serving as means for intervention with LTs, therefore, have therapeutic benefits in a variety of inflammatory and allergic diseases. 5-LOX inhibitors restrict the synthesis of LTs from arachidonic acid (AA). So, protein 5-LOX is an important therapeutic target.

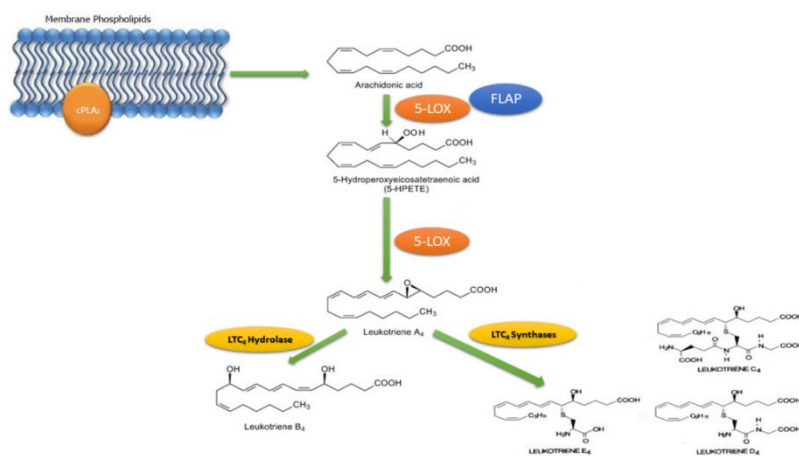


Fig. 1.5 Scheme representing 5-LOX mediated biosynthesis of Leukotrienes

1.3.2. Biochemistry

Human 5-LOX is a soluble, monomeric membrane binding enzyme consisting of 672 or 673 amino acids with a molecular weight of ~78 kDa, which is aided by the proteins FLAP (5-LOX Activation Protein) and CLP (coactosin like protein). It comprises a C-terminal catalytic domain (residues 126-673) and N-terminal C2-like b-barrel

domain that facilitates its binding to substrates, cellular phospholipid membranes, and Ca^{2+} . The C2-like domain has a regulatory function. Calcium (Ca^{2+}) can activate 5-LOX by inducing binding to phosphatidylcholine membranes and CLP. ATP is crucial for ALOX5's metabolic activity, and it binds to 5-LOX and increases enzyme activity, but hydrolysis of ATP is not required. Instead, it appears that ATP stabilizes the enzyme. Several structures of human 5-LOX have been recently released, including a substrate-bound form shown in Figure 1.6. All were obtained from the so-called human Stable 5-LOX form of the enzyme

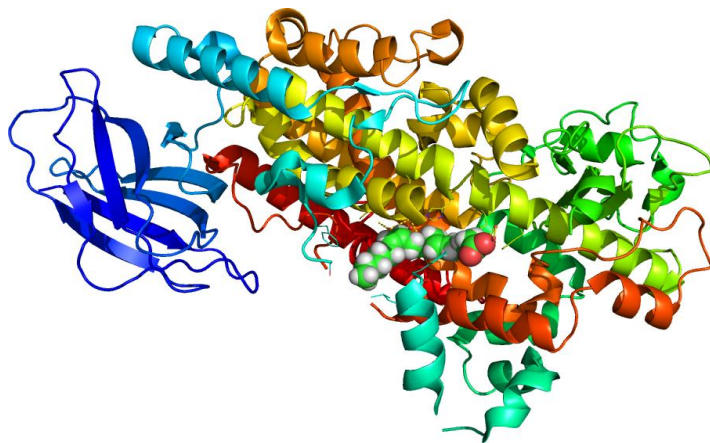


Fig. 1.6 3D Structure of substrate-bound human 5-LOX (PDB ID: 3V99)

1.3.3. 5-LOX Catalytic Mechanism

In the active site of 5-LOX, the catalytic domain contains a non-heme iron that functions as an electron acceptor or donor during catalysis. The catalytic Fe in this enzyme is held into place by coordination with side chains of three conserved His residues such as His 367, His 372 and His 550 and the main chain carboxyl group of the

C-terminal Ile 673 and an Asn 554 residue, which is not close enough to be present in the actual coordination sphere. During enzyme activation, to obtain maximum activity, the iron oxidized from an inactive ferrous (Fe^{2+}) state to the active ferric (Fe^{3+}) state through interaction with an oxidizing agent such as fatty acid hydroperoxide, AOOH (Figure 1.7). In the first step, a hydrogen atom is abstracted from a bisallylic group of AAs (C7 of AA) by the $\text{Fe}^{3+}\text{-OH}^-$ cofactor yielding $\text{Fe}^{2+}\text{-OH}_2$ and a pentadienyl π radical [70] followed by the addition of molecular oxygen to generate 5-HpETE (5-hydroperoxy eicosatetraenoic acid). The second step involves the removal of a hydrogen atom from C-10, resulting in the formation of the allylic epoxide LTA₄, followed by the release of the product from the enzyme and restores the $\text{Fe}^{3+}\text{-OH}^-$ moiety of the 5-LOX active form [70].

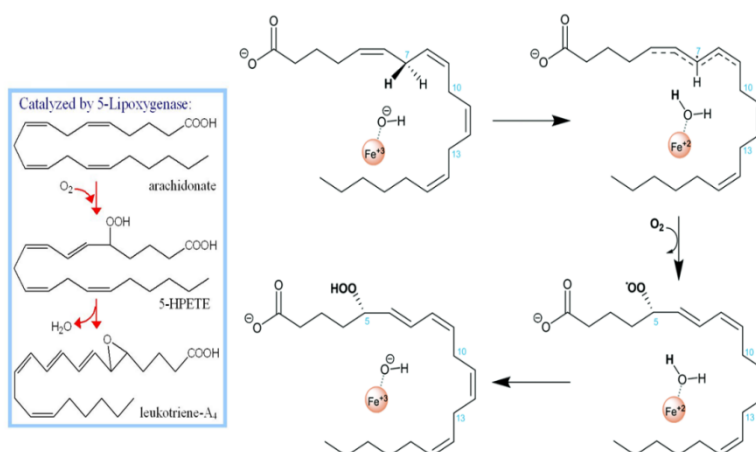


Fig. 1.7 Catalytic reaction mechanism of AA hydroperoxidation by iron at the active site of 5-LOX.

Access to the catalytic site is sealed in stable 5-LOX by the insertion of two aromatic amino acids such as Phe177 (F177) and

Tyr181 (Y181), referred to as the FY cork (17). Moreover, a deep, elongated inner cavity filled with conserved as well as specific amino acids is sufficiently large to accommodate the substrate. 5-LOX specific amino acids such as Tyr181, Ala603, Ala606, His600, and Thr364 and small side chains of Ala603 and Ala606 are needed for the conformation of Tyr181, which, along with Phe177, “corks” the cavity at one end (FY cork) [70]. Another 5-LOX-specific amino acid, Trp599, appears to support one side of the FY cork. The substrate, therefore, gets access to the catalytic iron either through removing the FY cork at one end of the cavity or through moving Trp599 to secure it, or even by moving Trp147 to the opposite end [70]. This observation indicates that AA will reach 5-LOX from the opposite direction as it only requires side-chain rotation. Also, it explains the catalytic process of H abstraction and peroxidation that happens on opposite sides of the pentadiene. The entrance at Trp147 allows AA to reach the methyl end first and place the substrate in order to generate the 5-HPETE's S-isomer.

1.3.4. Classification of 5-LOX Inhibitors

The perception that the 5-LOX activation mechanism is far more complex than the other lipoxygenases provides possibilities for new strategies of inhibition. Thus, a variety of approaches can be considered for the development of inhibitors of 5-LOX. Based on the mechanism of action, previous research has suggested four types of 5-LOX inhibitors:

1. **Redox inhibitors:** many small organic compounds with radical scavenging activity such as phenols, quinones, dihydroquinones, flavonoids, *etc.*, can interact with the redox cycling of 5-LOX [71] and inhibit the redox process such inhibitors are called redox inhibitors. Given the lipophilic nature of the substrate, these inhibitors have been usually small lipophilic molecules, such as mono-and polycyclic aromatics. Phenidone, BW755C, and AA-861 are some of the redox inhibitors. It is relatively difficult to describe Structural Activity Relationships (SAR) of this class of inhibitors. Furthermore, redox inhibitors have poor selectivity for inhibition of 5-LOX relative to inhibition of cyclogenesis (COXs). In other words, although they have a high potency in inhibiting leukotriene biosynthesis, these inhibitors exhibit poor selectivity, non-specificity, and ancillary activity. Such inhibitors also demonstrate significant toxicity due to the methemoglobin formation.
2. **Competitive reversible (non-redox) inhibitors:** The numerous toxicities and challenges faced in producing 5-LOX redox inhibitors have prompted several research groups to try competitive non-redox inhibitors. Non-redox inhibitors do not meddle with the oxidation reactions of 5-LOX. These inhibitors compete for substrate binding sites with arachidonic acid. That is why it used to be called competitive inhibitors. The

methoxyalkylthiazoles and methoxytetrahydropyrans derivatives are deemed to be non-redox inhibitors since redox reactions are unlikely to occur because of the structural features of these series. ZD-2138 and ICI211965 are potent competitive, reversible 5-LOX inhibitors.

3. **Iron chelating inhibitors:** This class of 5-LOX inhibitors is a non-toxic redox inhibitor with iron-chelating properties such as hydroxamic acid or N-hydroxyurea. Among them, Zileuton and Atreleuton are important and well known 5-LOX's iron-chelating inhibitors. Studies have suggested that maybe advancing the iron chelator inhibitors of 5-LOX could be a fascinating idea for further exploration.
4. **FLAP inhibitors:** The 5-lipoxygenase activating protein (FLAP) is a protein that enhances the enzymatic action of 5-LOX. Inhibitors of FLAP reduce the action of 5-LOX, which then controls the formation of leukotrienes. MK-866 is an inhibitor of FLAP that is safe for use and has an effect on the early and late stages of asthma responses to allergens. The other compounds that belong to the FLAP inhibitor class are MK-0591 and Bay-X-1005.

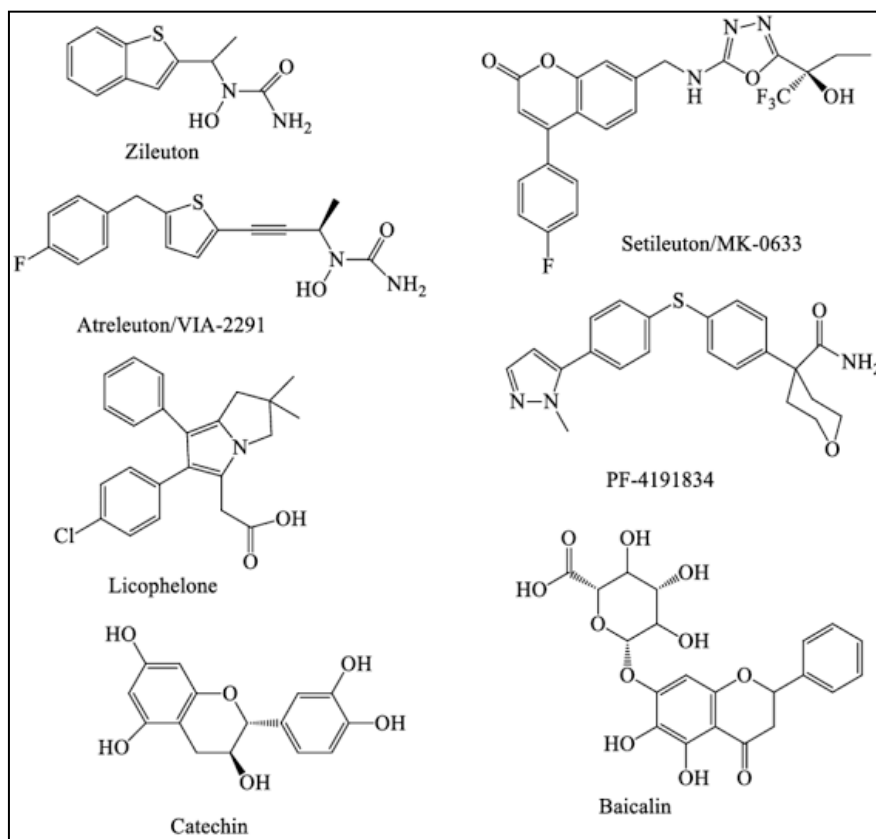


Fig.1.8 5-LOX inhibitors under clinical development

5-LOX inhibitors under clinical development are shown in Figure 1.8. Despite all the intensive efforts in the development of 5-LOX inhibitors, Zileuton (N-hydroxyurea derivative), an iron-chelating inhibitor, is only approved orally active inhibitor of 5-LOX available in the market [72]. All other potential candidates miserably failed due to a lack of efficacy in clinical studies or due to severe side effects. Zileuton itself has many side effects, including liver toxicity probably because of alkylation and irreversible inhibition of glutathione S-transferase M1 [GSTM-1] and unfavorable

pharmacokinetics with a short half-life [73]. Hence, the enhancement of the pharmacokinetic properties of 5-LOX inhibitors in terms of efficacy, selectivity, and safety has been a principal challenge and requires new leads with novel modes of molecular action. Therefore, the design of novel 5-LOX inhibitors triggers great enthusiasm among the scientific community.

1.3.5. Computational Methods in the Search for Novel 5-LOX Inhibitors

Computational methods are expected to play an important role in finding novel, better 5-LOX inhibitors with minimal side effects. Many articles have been published in recent years, documenting computational studies on 5-LOX inhibitors for the discovery of novel leads [74]. At present, computational approaches to predict 5-LOX inhibitors can be divided into two types: 1) Structure-based approach based on homology modeling and 2) Ligand-based approach based on pharmacophore and QSAR models. The comparative homology model of the human 5-LOX has been used to identify novel inhibitors until the newly resolved structure of human 5-LOX is published [75]. The newly resolved structure of human 5-LOX is an apo structure without any bound inhibitor and is a tough target due to its supposed flexibility. However, the reliable structure of 5-LOX has become an excellent source for structure-based lead optimization study. It can be used as a working tool for more precise predictions of binding interactions and affinities of inhibitors.

Wu *et al.*, have developed a comparative model for the active conformation of human 5-LOX using homology modeling based on the

closed conformation of 15-LOX [76]. They conducted docking studies and molecular dynamics simulations of known inhibitors to create the most rational conformation that explains inhibitory activities. This comparative model was then used for docking-based virtual screening to identify novel 5-LOX inhibitors. Aparoy P *et al.*, also used a 5-LOX comparative model to conduct virtual screening studies. Nevertheless, for both the model and the experimental assay, they used potato 5-LOX [77]. Due to the low similarity between human and plant enzymes, we need to ensure that the screened compounds inhibit human 5-LOX as well.

In LBDD, a few QSAR works have been reported with the aim of formulating an excellent predictive model for 5-LOX inhibitors [78]. Some of these studies [79,80] have used structure-based Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) as a tool to model the activity of human 5-LOX inhibitors. Several other studies [81,82] have used the conventional 2D QSAR methods such as Multiple Linear Regression (MLR), principal component analysis (PCA), and partial least square regression (PLS). Virtual screening of natural product derived combinatorial library also been carried out by a group of investigators based on a topological pharmacophore descriptor, CATS2D. Another two-step ligand-based virtual screening study used a charge-based descriptor to encode the active reference molecules [83]. Renner *et al.* designed a small, focused library with desired bioactivity [84].

1.4. Scope of the Present Study

Both experimental and computational approaches have played a key role in expanding the chemical space of 5-LOX inhibitors while trying to identify novel 5-LOX inhibitors without the limitations of known 5-LOX inhibitors. With the increase in the quantity and complexity of these structure-activity data, a comprehensive cheminformatic analysis of chemical space of 5-LOX inhibitors is becoming increasingly important to understand and predict the interactions between inhibitors and 5-LOX protein by employing the methods Predictive QSAR modeling and Virtual screening. So, in this study, we have tried to locate, characterize, and visualize 5-LOX inhibitor space and have decided to extract SAR from these large datasets and to present them intuitively. Besides, we tried to develop a Rat 5-LOX protein comparative model and a comprehensive assessment of the Human 5-LOX protein structure to verify the interaction of protein-ligand via molecular docking. We also decided to develop several predictive QSAR regression and classification models using conventional linear and modern non-linear machine learning methods. Also, we decided to perform virtual screening of an extensive database to identify potential 5-LOX inhibitors.

1.5. Objectives of the Present Study

Leukotrienes (LTs) are a class of potent inflammatory mediators produced by 5-LOX from arachidonic acid. The overproduction of LT causes severe allergic and inflammatory conditions. Inhibition of LT formation, therefore, has valuable

therapeutic advantages in excessive and chronic inflammatory allergic responses like asthma, rhinitis, rheumatoid arthritis, gastroesophageal reflux disease, atherosclerosis. Moreover, the role of LTs in carcinogenesis has also been documented recently. Because traditional anti-inflammatory treatments, like treatment with NSAID, are still far from effective in many of these diseases, new and improved approaches are being actively sought to counter these conditions. Methods that inhibit the biosynthesis of LTs are, therefore, of interest as potential therapies for such diseases. The non-heme iron-containing dioxygenase enzyme 5-LOX catalyzes the first step in the leukotriene biosynthetic pathway. So, 5-LOX selective inhibition offers a definite means of reducing the effects of all leukotrienes, and such an inhibitor could form a new class of therapeutic agents.

The emerging discipline, like Cheminformatics and CADD, are finding increasing applications in the field of rational drug design. With the help of virtual screening methods, researchers now routinely screen millions of compounds in the search for some biologically active compounds. The development of knowledge-based predictive QSAR models increasing with the assembly and integration of all medicinal chemistry structure-activity information. Data mining approaches and machine learning methods can uncover underlying patterns in chemical and pharmacological property space decisive for drug discovery and development. These methods are also have played a significant role in identifying, optimizing, and understanding the biological activity of 5-LOX inhibitors at the molecular level. In particular, a significant and relevant amount of structure-activity information of 5-LOX inhibitors has been released and stored in public

databases, and the newly resolved crystal structure of stable human 5-LOX is published recently. However, to the best of our knowledge, thorough cheminformatic modeling, and evaluation of 5-LOX protein and inhibitors space are very limited in the literature. In this context, we attempted:

- Preparation, evaluation, homology modeling, and the ligand-binding site identification of 5-LOX protein.
- To locate, characterize, and visualize the chemical space of 5-LOX inhibitors, including FLAP inhibitors space.
- To extract SAR from these large datasets and to presented them intuitively by structure-activity landscape modeling.
- To build CoMFA QSAR models of redox inhibitors of 5-LOX in order to understand 3D structural features that are essential for biological activity by utilizing the smooth SAR region of the structure-activity landscape.
- To develop several non-linear QSAR classification models using extensive updated and structurally diverse data set with the help of machine learning and data mining methods.
- To conduct a comparative assessment of commonly used docking programs in support of our attempts in virtual screening of novel 5-LOX inhibitors.
- To carry out virtual screening of the ZINC15 database to identify potential 5-LOX inhibitors that could be the next lead using the best docking algorithm.

References

- [1] J.L. Medina-Franco, K. Martinez-Mayorga, M.A. Giulianotti, R.A.H. and C. Pinilla, Visualization of the Chemical Space in Drug Discovery, *Curr. Comput. Aided. Drug Des.* 4 (2008) 322–333. doi:<http://dx.doi.org/10.2174/157340908786786010>.
- [2] C. Lipinski, A. Hopkins, Navigating chemical space for biology and medicine, *Nature*. 432 (2004) 855–861. doi:10.1038/nature03193.
- [3] C.M. Dobson, Chemical space and biology, *Nature*. 432 (2004) 824–828. doi:10.1038/nature03192.
- [4] J.-L. Reymond, R. van Deursen, L.C. Blum, L. Ruddigkeit, Chemical space as a source for new drugs, *Medchemcomm.* 1 (2010) 30–38. doi:10.1039/C0MD00020E.
- [5] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 28 (1988) 31–36. doi:10.1021/ci00057a005.
- [6] J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, C. Rücker, MOLGEN-CID A Canonizer for Molecules and Graphs Accessible through the Internet, *J. Chem. Inf. Comput. Sci.* 44 (2004) 542–548. doi:10.1021/ci0304041.
- [7] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.* 29 (1989) 97–101. doi:10.1021/ci00062a008.
- [8] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, *J. Cheminform.* 7 (2015) 23. doi:10.1186/s13321-015-0068-4.
- [9] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, BA Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.* 47 (2018) D1102–D1109. doi:10.1093/nar/gky1033.
- [10] H.E. Pence, A. Williams, ChemSpider: An Online Chemical Information Resource, *J. Chem. Educ.* 87 (2010) 1123–1124. doi:10.1021/ed100697w.

-
- [11] T. Sterling, J.J. Irwin, ZINC 15--Ligand Discovery for Everyone, *J. Chem. Inf. Model.* 55 (2015) 2324–2337. doi:10.1021/acs.jcim.5b00559.
- [12] J.H. Voigt, B. Bienfait, S. Wang, M.C. Nicklaus, Comparison of the NCI Open Database with Seven Large Chemical Structural Databases, *J. Chem. Inf. Comput. Sci.* 41 (2001) 702–712. doi:10.1021/ci000150t.
- [13] J. Chen, S.J. Swamidass, Y. Dou, J. Bruand, P. Baldi, ChemDB: a public database of small molecules and related chemoinformatics resources, *Bioinformatics.* 21 (2005) 4133–4139. doi:10.1093/bioinformatics/bti683.
- [14] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.* 35 (2007) D198–D201. doi:10.1093/nar/gkl999.
- [15] K.P. Seiler, G.A. George, M.P. Happ, N.E. Bodycombe, H.A. Carrinski, S. Norton, S. Brudz, J.P. Sullivan, J. Muhlich, M. Serrano, P. Ferraiolo, N.J. Tolliday, S.L. Schreiber, P.A. Clemons, ChemBank: a small-molecule screening and cheminformatics resource database, *Nucleic Acids Res.* 36 (2008) D351–D359. doi:10.1093/nar/gkm843.
- [16] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.* 40 (2012) D1100–D1107. doi:10.1093/nar/gkr777.
- [17] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res.* 36 (2008) D901–D906. doi:10.1093/nar/gkm958.
- [18] R.S. Bohacek, C. McMartin, W.C. Guida, The art and practice of structure-based drug design: A molecular modeling perspective, *Med. Res. Rev.* 16 (1996) 3–50. doi:10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.
- [19] K.L.M. Drew, H. Baiman, P. Khwaounjoo, B. Yu, J. Reynisson, Size estimation of chemical space: how big is it?, *J. Pharm. Pharmacol.* 64 (2012) 490–495. doi:10.1111/j.2042-7158.2011.01424.x.

-
- [20] L.C. Blum, J.-L. Reymond, 970 Million Drug-like Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, *J. Am. Chem. Soc.* 131 (2009) 8732–8733. doi:10.1021/ja902302h.
- [21] A.M. Virshup, J. Contreras-García, P. Wipf, W. Yang, D.N. Beratan, Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds, *J. Am. Chem. Soc.* 135 (2013) 7296–7303. doi:10.1021/ja401184g.
- [22] K.T. Nguyen, L.C. Blum, R. van Deursen, J.-L. Reymond, Classification of Organic Molecules by Molecular Quantum Numbers, *ChemMedChem.* 4 (2009) 1803–1805. doi:10.1002/cmdc.200900317.
- [23] J. Larsson, J. Gottfries, S. Muresan, A. Backlund, ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space, *J. Nat. Prod.* 70 (2007) 789–794. doi:10.1021/np070002y.
- [24] O. Rabal, J. Oyarzabal, Biologically Relevant Chemical Space Navigator: From Patent and Structure–Activity Relationship Analysis to Library Acquisition and Design, *J. Chem. Inf. Model.* 52 (2012) 3123–3137. doi:10.1021/ci3004539.
- [25] M. Hohman, K. Gregory, K. Chibale, P.J. Smith, S. Ekins, B. Bunin, Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery, *Drug Discov. Today.* 14 (2009) 261–270. doi:https://doi.org/10.1016/j.drudis.2008.11.015.
- [26] B.F. Begam, J.S. Kumar, A Study on Cheminformatics and its Applications on Modern Drug Discovery, *Procedia Eng.* 38 (2012) 1264–1275. doi:https://doi.org/10.1016/j.proeng.2012.06.156.
- [27] T. Engel, Basic Overview of Cheminformatics, *J. Chem. Inf. Model.* 46 (2006) 2267–2277. doi:10.1021/ci600234z.
- [28] F.K. Brown, Chapter 35 - Cheminformatics: What is it and How does it Impact Drug Discovery., in: J.A.B.T.-A.R. in M.C. Bristol (Ed.), *Annu. Rep. Med. Chem.*, Academic Press, 1998: pp. 375–384. doi:https://doi.org/10.1016/S0065-7743(08)61100-8.
- [29] D.S. Wishart, Introduction to Cheminformatics, *Curr. Protoc. Bioinforma.* 18 (2007) 14.1.1-14.1.9. doi:doi:10.1002/0471250953.bi1401s18.

-
- [30] P. Willett, *Chemoinformatics: a history*, Wiley Interdiscip. Rev. Comput. Mol. Sci. 1 (2011) 46–56. doi:10.1002/wcms.1.
- [31] V. Arulmozhi, R. Rajesh, *Chemoinformatics — A quick review*, in: 2011 3rd Int. Conf. Electron. Comput. Technol., 2011: pp. 416–419. doi:10.1109/ICECTECH.2011.5942128.
- [32] AL and C. Di Giovanni, *Virtual Screening Strategies in Drug Discovery: A Critical Review*, *Curr. Med. Chem.* 20 (2013) 2839–2860. doi:http://dx.doi.org/10.2174/09298673113209990001.
- [33] SZ. Grinter, X. Zou, *Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design*, *Molecules*. 19 (2014) 10150–10176. doi:10.3390/molecules190710150.
- [34] SJY. Macalino, V. Gosu, S. Hong, S. Choi, *Role of computer-aided drug design in modern drug discovery*, *Arch. Pharm. Res.* 38 (2015) 1686–1701. doi:10.1007/s12272-015-0640-5.
- [35] A.C. Anderson, *The Process of Structure-Based Drug Design*, *Chem. Biol.* 10 (2003) 787–797. doi:https://doi.org/10.1016/j.chembiol.2003.09.002.
- [36] S. Kalyaanamoorthy, Y.-P.P. Chen, *Structure-based drug design to augment hit discovery*, *Drug Discov. Today*. 16 (2011) 831–839. doi:https://doi.org/10.1016/j.drudis.2011.07.006.
- [37] T. Rodrigues, G. Schneider, *Flashback Forward: Reaction-Driven De Novo Design of Bioactive Compounds*, *Synlett*. 25 (2014) 170–178. doi:10.1055/s-0033-1340216.
- [38] C. McInnes, *Virtual screening strategies in drug discovery*, *Curr. Opin. Chem. Biol.* 11 (2007) 494–502. doi:https://doi.org/10.1016/j.cbpa.2007.08.033.
- [39] A.D. Andricopulo, L.B.S. and D.J. Abraham, *Structure-Based Drug Design Strategies in Medicinal Chemistry*, *Curr. Top. Med. Chem.* 9 (2009) 771–790. doi:http://dx.doi.org/10.2174/156802609789207127.
- [40] P. Prathipati, A.D. and A.K. Saxena, *Computer-Aided Drug Design: Integration of Structure-Based and Ligand-Based Approaches in Drug Design*, *Curr. Comput. Aided. Drug Des.* 3 (2007) 133–148. doi:http://dx.doi.org/10.2174/157340907780809516.
- [41] A. Yosipof, R.C. Guedes, A.T. García-Sosa, *Data Mining and*
-

-
- Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category, *Front. Chem.* 6 (2018) 162. doi:10.3389/fchem.2018.00162.
- [42] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, *AI Mag.* 17 (1996) 37. doi:10.1609/aimag.v17i3.1230.
- [43] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, in: *Proc. Second Int. Conf. Knowl. Discov. Data Min.*, AAAI Press, 1996: pp. 82–88. <http://dl.acm.org/citation.cfm?id=3001460.3001477>.
- [44] A. Feelders, H. Daniels, M. Holsheimer, Methodological and practical aspects of data mining, *Inf. Manag.* 37 (2000) 271–281. doi:[https://doi.org/10.1016/S0378-7206\(99\)00051-8](https://doi.org/10.1016/S0378-7206(99)00051-8).
- [45] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: Current issues and guidelines, *Int. J. Med. Inform.* 77 (2008) 81–97. doi:<https://doi.org/10.1016/j.ijmedinf.2006.11.006>.
- [46] T.K. Shameera Ahamed, V.K. Rajan, K. Muraleedharan, QSAR modeling of benzoquinone derivatives as 5-lipoxygenase inhibitors, *Food Sci. Hum. Wellness.* 8 (2019) 53–62. doi:<https://doi.org/10.1016/j.fshw.2019.02.001>.
- [47] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.* 110 (1988) 5959–5967. doi:10.1021/ja00226a005.
- [48] M. Clark, R.D. Cramer, D.M. Jones, D.E. Patterson, P.E. Simeroth, Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases, *Tetrahedron Comput. Methodol.* 3 (1990) 47–59. doi:[https://doi.org/10.1016/0898-5529\(90\)90120-W](https://doi.org/10.1016/0898-5529(90)90120-W).
- [49] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, C. Duraiswami, Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism, *J. Am. Chem. Soc.* 119 (1997) 10509–10524. doi:10.1021/ja9718937.
- [50] A. Vedani, M. Dobler, Multidimensional QSAR: Moving from three- to five-dimensional concepts, *Quant. Struct. Relationships.* 21 (2002) 382–390. doi:10.1002/1521-3838(200210)21:4<382::AID-QSAR3
-

82>3.0.CO;2-L.

- [51] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I. Igor, M. Cronin, J. Dearden, P. Gramatica, YC. Martin, V. Consonni, V.E. Kuz, R. Cramer, *QSAR Modeling: Where have you been? Where are you going to?*, 57 (2015) 4977–5010. doi:10.1021/jm4004285.QSAR.
- [52] A. Golbraikh, X.S. Wang, H. Zhu, A. Tropsha, *Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment* BT - *Handbook of Computational Chemistry*, in: J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, M. K. Shukla (Eds.), Springer International Publishing, Cham, 2017: pp. 2303–2340. doi:10.1007/978-3-319-27282-5_37.
- [53] W. Karcher, J. Devillers, *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology*, Springer Science & Business Media, 1990.
- [54] E. Burello, A.P. Worth, *QSAR modeling of nanomaterials*, *Wiley Interdiscip. Rev. Nanomedicine Nanobiotechnology*. 3 (2011) 298–306. doi:10.1002/wnan.137.
- [55] K. Roy, *Advances in QSAR Modeling*, in: *Appl. Pharm. Chem. Food, Agric. Environ. Sci.*, Springer, 2017: p. 555.
- [56] T. Puzyn, J. Leszczynski, M.T. Cronin, *Recent advances in QSAR studies: methods and applications*, Springer Science & Business Media, 2010.
- [57] J. Verma, V.M. Khedkar, E.C. Coutinho, *3D-QSAR in Drug Design - A Review*, *Curr Top Med Chem*. 10 (2010) 95–115. doi:10.2174/156802610790232260.
- [58] S. Yousefinejad, B. Hemmateenejad, *Chemometrics tools in QSAR/QSPR studies: A historical perspective*, *Chemom. Intell. Lab. Syst.* 149 (2015) 177–204. doi:https://doi.org/10.1016/j.chemolab.2015.06.016.
- [59] A. Lavecchia, *Machine-learning approaches in drug discovery: Methods and applications*, *Drug Discov. Today*. 20 (2015) 318–331. doi:10.1016/j.drudis.2014.10.012.
- [60] W.P. Walters, M.T. Stahl, M.A. Murcko, *Virtual screening—an*

- overview, *Drug Discov. Today*. 3 (1998) 160–178. doi:[https://doi.org/10.1016/S1359-6446\(97\)01163-X](https://doi.org/10.1016/S1359-6446(97)01163-X).
- [61] B.K. Shoichet, Virtual screening of chemical libraries, *Nature*. 432 (2004) 862–865. doi:10.1038/nature03197.
- [62] A. Dhasmana, S. Raza, R. Jahan, M. Lohani, J.M. Arif, Chapter 19 - High-Throughput Virtual Screening (HTVS) of Natural Compounds and Exploration of Their Biomolecular Mechanisms: An In Silico Approach, in: M.S. Ahmad Khan, I. Ahmad, D.B.T.-N.L. to P. Chattopadhyay (Eds.), Academic Press, 2019: pp. 523–548. doi:<https://doi.org/10.1016/B978-0-12-814619-4.00020-3>.
- [63] J.Z. Haeggström, C.D. Funk, Lipoxygenase and leukotriene pathways: Biochemistry, biology, and roles in disease, *Chem. Rev.* 111 (2011) 5866–5896. doi:10.1021/cr200246d.
- [64] J. Ghosh, C.E. Myers, Arachidonic acid stimulates prostate cancer cell growth: critical role of 5-lipoxygenase., *Biochem. Biophys. Res. Commun.* 235 (1997) 418–23. doi:10.1006/bbrc.1997.6799.
- [65] LG Melstrom, DJ Bentrem, M.R. Salabat, T.J. Kennedy, X.Z. Ding, M. Strouch, S.M. Rao, R.C. Witt, C.A. Ternent, M.S. Talamonti, R.H. Bell, T.A. Adrian, Overexpression of 5-lipoxygenase in colon polyps and cancer and the effect of 5-LOX Inhibitors in vitro and in a murine model, *Clin. Cancer Res.* 14 (2008) 6525–6530. doi:10.1158/1078-0432.CCR-07-4631.
- [66] R. Hennig, P. Grippo, X.Z. Ding, S.M. Rao, M.W. Buchler, H. Friess, M.S. Talamonti, R.H. Bell, T.E. Adrian, 5-Lipoxygenase, a marker for early pancreatic intraepithelial neoplastic lesions, *Cancer Res.* 65 (2005) 6011–6016. doi:10.1158/0008-5472.CAN-04-4090.
- [67] Y. Chen, Y. Hu, H. Zhang, C. Peng, S. Li, Loss of the Alox5 gene impairs leukemia stem cells and prevents chronic myeloid leukemia, *Nat Genet.* 41 (2009) 783–792. doi:10.1038/ng.389.
- [68] Alan R. Brash, Lipoxygenases: Occurrence, Functions, Catalysis, and Acquisition of Substrate, *J. Biol. Chem.* 274 (1999) 23679–23682.
- [69] C.D. Funk, Prostaglandins and leukotrienes: advances in eicosanoid biology., *Science* (80). 294 (2001) 1871–1875.

- [70] S. Mitra, Insights into 5-Lipoxygenase Active Site and Catalysis, 2015.
- [71] R.N. Young, Inhibitors of 5-lipoxygenase: A therapeutic potential yet to be fully realized?, *Eur. J. Med. Chem.* 34 (1999) 671–685. doi:10.1016/S0223-5234(99)00225-1.
- [72] S.E. Wenzel, A.K. Kamada, Zileuton: The first 5-lipoxygenase inhibitor for the treatment of asthma, *Ann. Pharmacother.* 30 (1996) 858–864.
- [73] M.C. Liu, L.M. Dube, J. Lancaster, M. Blumenthal, P.B. Boggs, N. K., Clinical aspects of allergic disease Acute and chronic effects of a 5-lipoxygenase inhibitor in asthma : A 6-month randomized multicenter trial, *J. Allergy. Clin. Immunol.* 98 (1996) 859–871.
- [74] D. Steinhilber, B. Hofmann, Recent advances in the search for novel 5-lipoxygenase inhibitors, *Basic Clin. Pharmacol. Toxicol.* 114 (2014) 70–77. doi:10.1111/bcpt.12114.
- [75] NC. Gilbert, S.G. Bartlett, M.T. Waight, D.B. Neau, W.E. Boeglin, A.R. Brash, ME Newcomer, The structure of human 5-lipoxygenase, *Science.* 331 (2011) 217–219. doi:10.1126/science.1197203.
- [76] Y. Wu, C. He, Y. Gao, S. He, Y. Liu, L. Lai, Dynamic Modeling of Human 5-Lipoxygenase–Inhibitor Interactions Helps To Discover Novel Inhibitors, *J. Med. Chem.* 55 (2012) 2597–2605. doi:10.1021/jm201497k.
- [77] P. Aparoy, R.N. Reddy, L. Guruprasad, M.R. Reddy, P. Reddanna, Homology modeling of 5-lipoxygenase and hints for better inhibitor design, *J. Comput. Aided. Mol. Des.* 22 (2008) 611–619. doi:10.1007/s10822-008-9180-0.
- [78] G. Eren, A. MacChiarulo, E. Banoglu, From molecular docking to 3D-quantitative structure-activity relationships (3D-QSAR): Insights into the binding mode of 5-Lipoxygenase inhibitors, *Mol. Inform.* 31 (2012) 123–134. doi:10.1002/minf.201100101.
- [79] P. Aparoy, G.K. Suresh, K.K. Reddy, P. Reddanna, CoMFA and CoMSIA studies on 5-hydroxyindole-3-carboxylate derivatives as 5-lipoxygenase inhibitors : Generation of homology model and docking studies, *Bioorg. Med. Chem. Lett.* 21 (2011) 456–462. doi:10.1016/j.bmcl.2010.10.119.

- [80] M.A. Babu, N. Shakya, P. Prathipati, S.G. Kaskhedikar, A.K. Saxena, Development of 3D-QSAR Models for 5-Lipoxygenase Antagonists: Chalcones, 10 (2002) 4035–4041.
- [81] M.F. Andrada, E.G. Vega-hissi, M.R. Estrada, J.C. Garro, Chemometrics and Intelligent Laboratory Systems Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors, *Chemom. Intell. Lab. Syst.* 143 (2015) 122–129. doi:10.1016/j.chemolab.2015.03.001.
- [82] J. Ruiz, C. Pérez, R. Pouplana, QSAR study of dual cyclooxygenase and 5-lipoxygenase inhibitors 2,6-di-tert-butylphenol derivatives, *Bioorganic Med. Chem.* 11 (2003) 4207–4216. doi:10.1016/S0968-0896(03)00449-8.
- [83] B. Hofmann, L. Franke, E. Proschak, Y. Tanrikulu, P. Schneider, D. Steinhilber, G. Schneider, Scaffold-Hopping Cascade Yields Potent Inhibitors of 5-Lipoxygenase, *ChemMedChem.* 3 (2008) 1535–1538. doi:10.1002/cmdc.200800153.
- [84] S. Renner, W.A.L. van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofmann, S. Wetzal, A. Schuffenhauer, P. Ertl, T.I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann, Bioactivity-guided mapping and navigation of chemical space, *Nat. Chem. Biol.* 5 (2009) 585–592. doi:10.1038/nchembio.188.

2

THEORETICAL AND METHODOLOGICAL OVERVIEW

2.1. Molecular Representation

In chemistry, like all other disciplines, digital data is growing exponentially in all shapes and sizes. However, unlike all other fields, chemistry needed to store and manipulate 'molecular structure' other than numbers, images, and text. So, studies on computational representation and manipulation of the chemical structure that includes bonds and atoms are most important. It provides a method of understanding structural properties and characteristics belonging to a particular molecule.

The sophistication of molecular representations varies depending on how they derive and how much data they gather. One-dimensional (1D) molecular representation, such as the molecular formulas, is the simplest among them. Nevertheless, molecular formulas that only give atomic composition are not enough to define a molecule. So, 'molecular graphs' that are commonly used to preserve chemical structures on a computer are the best and most used molecular representation method. Molecular graphs are two-dimensional (2D) depictions of chemical structures, with nodes

corresponding to atoms and edges corresponding to bonds. Hydrogen atoms are most often excluded. The molecular graph only shows the topology of a molecule, such as the connection between the nodes (or atoms). The molecular graph method requires a way to interact with the molecular graph in and out of the computer. This process would be achieved by connection tables and linear notation methods. Atomic coordinates (Information about the atoms XY or XYZ coordinates), their hybridization states, and bond orders are provided in connection tables. The MDL Information Systems, MDL SDF (structure data file), Mol2, *etc.*, are the example of this representation.

A linear or line notation represents a molecule as a single-line string of alphanumeric and special characters. WLN (Wisswesser Line Notation), SMILES (Simplified Molecular Input Line Entry System) [1,2], and InChI [3] are examples of this notation. In recent years, linear notations have become more commonly used that enable us to represent, store, and transfer large numbers of molecules in a compact and simple form. Canonicalization of a linear notation is a way of bringing unique ordering of the atoms for a given molecular graph. That is, a linear notation can choose one “blessed” representation from among the many, and the representation is called canonical representation. It is kind of comparable to the name of the International Union of Pure and Applied Chemistry (IUPAC) for a chemical structure.

The spatial arrangements of atoms and bonds, defining the steric and electronic properties of a molecule, cannot be represented by 2D molecular graph representations. These properties are often depending on how atoms are positioned in space to produce 3D

structures or conformations. Most molecules can have more than one low-energy conformation, and the number of accessible structures is vast in many cases. Therefore, efficient ways of taking account of conformational flexibility are needed. Three-dimensional (3D) representations like molecular surface and volume depict the molecule's essential conformational characteristics. The standard file formats like Mol2, PDB, CIF, *etc.*, are enclosed 3D information of the compound. The data stored in a 3D database is usually the result of an experiment or a computational calculation.

In this thesis, all types of representations had to be used in different contexts. For example, the hierarchical scheme for the representation of zileuton with different content of the structural information is given in Figure 2.1.

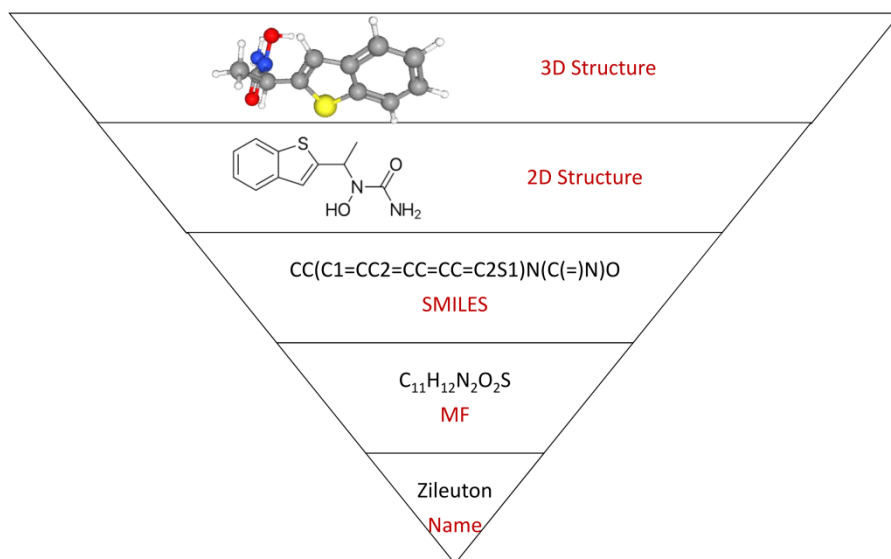


Fig. 2.1 Hierarchical scheme for representation of zileuton with different content of the structural information.

2.2. Molecular Modeling

Molecular modeling is the method of numerically representing molecular structures and simulating their behavior using quantum or classical mechanics equations with a wide variety of theoretical and computational methods [4]. For calculating molecular properties, energy minimized structure is needed. Energy minimization is the process of finding a conformer where the net inter-atomic force on each atom is acceptably nearly zero. It involves successive iterative computations where an initial conformation is submitted to full geometry optimization.

Molecular Mechanics (MM) is a commonly used molecular modeling method for molecules of large size and for a situation where the number of molecules is thousands or more [5]. Here, atoms considered as point charges with an associated mass, and the bonds between molecules are considered as springs. Classical mechanics (Newtonian mechanics) are used to calculate the potential energy surface for a specific arrangement of atoms. The potential energy function is the sum of individual functions for bond stretching, angle bending, torsional energies, and non-bonding interactions. The potential energy of all systems in MM is calculated using force fields. The force field is a collection of equations and associated constants designed to reproduce molecular geometry and selected properties of tested structures, which approximates the quantum mechanical energy surface to a classic mechanical model, thereby reducing the computational cost of large system simulations by magnitude orders.

Different force fields use different forms for the various interactions within and between molecules. CHARMM, AMBER, MMFF, GROMOS, OPLS, UFF, MM2, MM3, MM4, OPLS, *etc.*, are the examples for the force field. For chemical space analysis, predictive modeling and virtual screening studies, energy minimization or optimization of large numbers of the molecule have to be done. For these, molecular mechanics methods are used because they can reduce the computational cost and time.

Quantum mechanical based molecular modeling involves the use of a sophisticated quantum mechanical method such as Hartree-Fock (HF), Post Hartree-Fock (correlation method), Density Functional Theory (DFT), and Semi-Empirical (SE) method to mimic the behavior of molecules [6]. It accounts for the electronic nature of molecules and computes energy by solving complex Schrodinger equations. For a study involving small numbers of molecules like CoMFA, we have used the DFT method for the geometry optimization of ligands.

2.3. Molecular Descriptors and Fingerprints

A chemical compound can be described in two main ways: either using global descriptors or using fingerprints descriptors. Descriptors are a mathematical representation of a molecule that contains different sources of chemical information transformed and coded to deal with chemical, biological, and pharmacological features of various agents or small molecules. The definition of the descriptor by Todeschini and Consonni is, [7]: "The molecular descriptor is the

final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." These are commonly used to develop QSAR and QSPR models. Over the years, a broad range of descriptors has been reported capturing chemical structural characteristics and properties. These descriptors can be classified based on different criteria.

Classification based on the origin of descriptors:

1. **Experimental measurements:** these are the descriptors obtained from a standardized experiment such as log P, molar refractivity, dipole moment, Abraham's H-bond parameters, solvent parameters, NMR shift, polarizability, and, in general, physicochemical properties
2. **Theoretical calculations:** These are obtained from chemical information contained in the molecule through mathematical and computational procedures.

Classification based on the dimensionality of structure representation:

The majority of theoretical descriptors can be classified according to "dimensionality" of the structural representation. This yields three classes, such as 1, 2, and 3 D molecular descriptors.

1. **1D:** Derived from a molecular formula, also known as constitutional descriptors. *e.g.*, atom and bond counts

2. **2D:** Calculated from 2D molecular graphs such as topological index, fragment counts, and molar refractivity.
3. **3D:** Obtained from the molecular geometry, depend on conformations of molecules, such as shape, molecular surface area & fields, 3D pharmacophore keys, quantum chemical descriptors, *etc.*

Classification based on the described object:

1. **Global descriptors:** derived from the whole molecule, such as molecular volume, molecular surface, dipole moment, topological indices, *etc.*
2. **Local descriptors:** derived from particular atoms or molecular fragments, *e.g.*, atomic charges, bonds polarizabilities, CATS descriptors, ISIDA descriptors
3. **Field descriptors:** describing molecular fields in the area surrounding the molecule like electrostatic potential, CoMFA descriptors, *etc.*

Classification based on “nature” of the descriptors:

1. **Constitutional:** fragment additive, reflecting the composition and general properties of the compound without any geometrical information.
2. **Topological:** calculated using the mathematical graph theory applied to the scheme of atoms connections of the structure.

3. **Geometrical:** different kinds of conformationally dependent descriptors based on the molecular geometry
4. **Electronic:** derived from the spatial distribution of the electrons in the molecule
5. **Quantum-chemical:** calculated directly from orbital energies of the optimized geometries

To get descriptors, usually, molecular structures of compounds are first either drawn with the aid of some software such as ChemDraw, ACD/ ChemSketch, GaussView, Chemcraft, Maestro, *etc.* or downloaded from different databases like PubChem, ChEMBL, ZINC, BindingDB, *etc.* They were then optimized using computational chemistry software either with semi-empirical methods such as AM1 or PM6 or with the DFT method. Then, the optimized molecular structures are inputted into online or offline descriptor calculation software such as e-DRAGON, PowerMV, OCHEM, ODESSA, Canvas to calculate the descriptors. This process generally enables the retrieval of hundreds or thousands of descriptors.

All molecular descriptors mentioned above can be used to highlight different chemical aspects of a molecular structure, but only a few of them can be used for similarity assessment of compounds. In order to compare them, molecules must be available in a numerically well-defined form. Fingerprint descriptors are quite well-established for this purpose because they are effortless to compute, efficient to store, and easy to manipulate. Fingerprints are typically encoded as

binary bit strings whose settings produce, in different ways, a bit “pattern” characteristic of a given molecule. Every bit may reflect some feature's absence (0) or presence (1). Comparing the bit strings is easier than comparing the molecules. The length and complexity of molecular fingerprints vary significantly from simple representations of limited topological distances or functional group occurrences to complicated multi-point 3D-pharmacophore arrangements [8]. Most popular molecular fingerprints can be grouped into the following classes.

1. **Topological fingerprints:** these fingerprints depict the paths of molecular characteristics, typically atoms, linear to a certain number of bonds. It is useful for clustering compounds, no specific meaning to an individual bit, *e.g.*, Daylight, atom pairs [9].
2. **Structural keys:** these are based on substructure features; each bit represents presences or absence of predefined functional groups, substructure motifs, or fragments, *e.g.*, MACCS (166 bit and 320 bit) [10], BCI, PubChem [11].
3. **Circular fingerprints:** it records radial environments connections of each atom with increased shells of atom connections, No specific meaning to an individual bit, *e.g.*, Molprint2D, ECFP.
4. **Pharmacophore fingerprints:** these are fingerprints associated with structural or chemical properties of compounds

that are thought to be responsible for a specific pharmacological action, *e.g.*, CAT descriptors, 3pt, and 4pt 3D fingerprints.

2.4. Feature Selection Methods

Not all of the calculated molecular descriptors are needed for representing features between inhibitors and non-inhibitors. Noisy, redundant, or irrelevant descriptors should be removed without much loss of information, thereby reducing the risk of overfitting. A selection criterion is required to remove irrelevant descriptors that can measure the relevance of each selected descriptor with the output of any classifier [12]. This selection process can be achieved by employing feature selection methods. In this thesis, for QSAR modeling, feature Selection methods have performed for the selection of suitable descriptors from a massive number of raw descriptors contain little information or are correlated with other descriptors without incurring much loss of information. The commonly used feature selection methods can be classified into three main categories: filter, wrapper, and embedded approaches [13]. In the filter approach, the procedure of subsets selection of descriptors is generally a pre-processing step, independently of a learning algorithm, this makes filter methods are too simple, quick, requiring little computational time, and independent of the classifier used. The wrapper method selects the best features subset based on the learning algorithm used to train the model itself, as a result of this wrapper method is too expensive in terms of computational complexity and time and have a

risk of over fitting to the model. Embedded approaches are performed feature selection as part of the learning algorithm. The algorithm which we used to run it decides which attribute to consider and which to eliminate while the model is being created.

CFS (Correlation-based Feature Selection) [14] is a highly effective feature selection method in reducing dimensionality, removing redundant descriptors, increasing learning accuracy, and improving the performance of machine learning algorithms. The heart of CFS (Correlation-based Feature Selection) is a heuristic for evaluating the worth or merit of a subset of features to correlated with the classification, yet uncorrelated with each other. The preceding equation provides the merit of a feature subset (S) comprising of k features (Equation 2.1):

$$Merit_s = \frac{\overline{kr_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (2.1)$$

Here, the average value of all feature-classification correlations is represented by r_{cf} is, and the average value of all feature-feature correlations is represented by r_{ff} is. The equation is, in fact, Pearsons correlation where all variables have been standardized.

Information gain [15] is another filtering method which is initially used for obtaining splitting criteria for the decision tree, now largely used as a feature selection algorithm. Information gain measures ‘Shannon entropy’ of a given feature to decide how

important that feature is. For a data set S with n class labels, the overall entropy ' I ' is defined as (Equation 2.2)

$$I(S) = -\sum_{i=1}^c p_i \log_2 p_i \quad (2.2)$$

Where p_i the portion of instances that belong to class i . The information gain of an attribute is determined by the entropy reduction, which is stored by learning an attribute A (feature A) by Equation 2.3:

$$IG(A) = I(S) - \sum_i \frac{S_i}{S} I(S_i) \quad (2.3)$$

Where $I(S)$ is the entropy of the given dataset and $I(S_i)$ is the entropy of the i^{th} subset generated by partitioning S based on feature A . Information gain rank each feature in terms of decreasing entropy, greater the decrease in entropy, the more significant each feature for prediction. This method tries to give a feature with high information gain has a higher rank than other features because it has more substantial power in classifying the data.

2.5. Molecular Similarity

Molecular similarities provide a means of grouping compounds based on descriptors or fingerprints derived from different structural, biological, or physicochemical characteristics. Even though molecular descriptors encode the chemical information, these measurements determine the level of similarity. Similarity measures were extensively used in the calculation of the similarity or dissimilarity between two

dataset samples. It plays a significant role in lead discovery and compound optimization.

2.5.1. Similarity Metrics

The molecular similarity among compounds portrayed as bit strings could be quantified with similarity metrics. To date, there have been several similarity metrics in several scientific disciplines [16,17]. In cheminformatics, the Tanimoto coefficient (Tc) is perhaps the most frequently used similarity metric that quantifies overlap between fingerprints or descriptors. The Tc is the ratio between the number of features stored in both fingerprints of the molecules and the total number of features of each molecule in either fingerprint. If a and b are the number of features present in compound A and B respectively, and ' c ' is the number of features shared by both compounds, the Tc between the two compounds shall be as follows (Equation 2.4):

$$Tc = \frac{c}{a + b - c} \quad (2.4)$$

Tc=0, when there is no single common feature in two compounds. As the number of common features increases, so does the Tc value. Tc=1, when two sets of features are the same, i.e., all features the presence of identical between two compounds. The Tc similarity value between two compounds, therefore, always lies within the interval [0, 1].

2.6. Diversity Analysis

The extrapolation of the structural diversity of compound databases is of high importance in drug discovery programs. For example, database diversity is highly recommended for the construction of reliable and effective QSAR predictive models. If the goal of an HTS camp is to identify impact compounds with a desired polypharmacological profile, it is beneficial to screen a highly diverse compound collection. If the screening campaign's purpose is to build a focused library further, a compound dataset with high internal similarity (low diversity) should be screened. The diversity metric can measure the diversity of the chemical library. Molecular representation is a key aspect of the analysis of diversity. Molecular descriptors (including physicochemical properties and molecular fingerprints) and chemical scaffolds are the most common means of representing molecules that can be used to measure diversity. Details are provided in Chapter 4.

2.7. Chemical Space Analysis

For medicinal chemistry and cheminformatics, the idea of chemical space is of great importance because understanding and extracting useful information from this multidimensional data helps in the optimization of the descriptor space of active molecules, and the process is known as multidimensional compound optimization. This method improves efficiency at an early stage and reduces attrition at a later stage of drug development. Besides, the recent innovation of database fingerprints (DFPs) facilitated the charting of multiple target-

focused libraries in the chemical space, thus offering insights into polypharmacology [18]. There are no standardized methods for the representation of chemical space. A widely used method encompasses the calculation of similarity matrices, which collect all pairwise comparisons. All matrices are squared with n columns and rows, with n being equal to the number of compounds in the dataset. Visualization of chemical space by compressing relevant information; therefore, important and is considered as the next step. Principal Components Analysis (PCA) and Self-Organizing Maps (SOM) are the two important techniques of visualization that is used to decrease the multi-dimensional space into a two or three-dimensional graph.

2.7.1. PCA

PCA is a linear projection technique that finds underlying variables called principal components (PC), which are the eigenvectors of the variance-covariance matrix of the multidimensional space matrix. PCA is useful to compress most of the relevant information in a few variables. The first two or three dimensions taken from a PCA can be used to explain much of the variation in the data set of the original multi-dimension space. This makes it possible to obtain visualizations of the chemical space.

2.7.2. SOM

SOM, which is also called Kohonen Map, is a popular and robust artificial neural network algorithm that is trained using unsupervised learning to build a two-dimensional map of a complex

high-dimensional input space by preserving the local features of the input data. The SOM has gained the researcher's interest due to its ability to interpret complex multidimensional information intuitively. Typical applications of Kohonen SOMs in cheminformatics were the classification of molecules through biological activity, the collection of data sets covering wide diversity, the interpretation of the effects of high-throughput screens or other screens, and the identification of possible SARs, and the generation of molecular descriptors by mapping molecular components. SOM can cluster compounds by assigning similar compounds to the same neuron. No. of neuron selected for each chemical space is calculated using the following Equation 2.5:

$$M = 5 \times \sqrt{N} \quad (2.5)$$

Here, M is the number of neurons, and N is the number of observations. A rectangular region of smoothly evolving property space depicts the final grid of reference vectors, in which each input vector can be allocated to a similar reference vector.

2.8. 3D-QSAR

3D-QSAR methods have emerged as a natural extension of 2D-QSAR to improve the predictive accuracy of 2D methods by exploiting the three-dimensional properties of ligands. 3D methods are computationally more dynamic and challenging than 2D approaches. Two types of 3D-QSAR methods are typically present: alignment-dependent methods and alignment-independent methods. In the

alignment-dependent method, the training set must be superimposed (aligned) over a template either based on experimental data (bioactive conformations) or based on molecular superimposition software.

2.8.1. CoMFA

CoMFA [19,20] is a promising new approach to 3D-QSAR. CoMFA describes two fields, such as steric and electrostatic fields; these fields provide all necessary information for understanding the biological properties of a set of compounds via partial least squares (PLS) analysis. Open3DQSAR software [21] has used to perform CoMFA analysis. We have used Van der Waals and electrostatic fields that are calculated from classical molecular mechanics equations using the Merck force field.

2.8.2. Molecular Alignment

Proper alignment of the compounds relative to one another is one of the most important steps in 3D-QSAR analysis for obtaining a valid molecular interaction field model. The mixed alignment procedure in the combination of the atom-based fashion and pharmacophore-based fashion was performed using Open3DALIGN software (version 2.27), which is an open-source tool capable of carrying out the multi-conformational, unsupervised rigid-body alignment of 3D molecular structures [22]. The alignment procedure was executed by using all available molecules as possible templates. For each alignment, the O3A score derived from the source code of the Open3DALIGN program is computed, which indicates the quality of

the superimposition. The best alignment is either obtained using most active molecules as a template or using the alignment corresponding to the highest cumulative O3A score.

2.8.3. CoMFA Procedure

The best alignment is placed in a 3D cubic lattice with a 2 Å grid size and a 5.0 Å out gap. The steric fields were computed using an sp^3 hybridized carbon atom probe with a +1 charge. Similarly, electrostatic fields were computed using a volume-less probe. These steric and electrostatic interaction energies were considered as independent variables (CoMFA descriptors). Before creating of CoMFA model, following pre-treatment operations were carried out to reduce the noise hidden in the PLS matrix and hence reduced the computational time: 1) the minimum and maximum energy values of steric and electrostatic were set to a cutoffs value -30.0 and +30.0 kcal/mol, respectively. This process avoids the infinity of energy values inside the molecule. 2) Low energy values were set (< 0.05 kcal/mol) to zero in both fields. 3) Standard deviation set to < 0.1 in order to improve the signal-to-noise ratio. 4) N-level variables that are variables that assume only N values across the training set were removed, most of which distributed on a small number of objects. This operation avoids overweighting the importance of particular substituents present in a single molecule. Otherwise, it might negatively affect the whole model. 5) The whole block of X or Y variables scaled by block unscaled weighting (BUW) technique.

Predictivity of the CoMFA model can be significantly improved by appropriate variable clustering and selection procedures such as Smart Region Definition (SRD) procedure, Fractional Factorial Design (FFD), and Uninformative Variable Elimination (UVE) technique. These variable selection techniques selectively remove noisy variables with no predictability. The SRD procedure carries out variable grouping based on their closeness in 3D space in order to reduce the redundancy arising from the existence of multiple nearby descriptors, which encode the same kind of information [23]. FFD aims at selecting the variables which significantly increase the predictive ability (using the LOO, LTO or LMO paradigms), and can operate on both single variables or groups identified by a previous SRD run, thereby removing uninformative variables groups as performed in GOLPE [24]. UVE procedure removes the least informative variables, characterized by small PLS pseudo-coefficients.

The partial least square analysis was employed to obtain a correlation between the descriptors derived by CoMFA (independent variables) and pIC_{50} values (dependent variable). Open3DQSAR generates a PLS model through the NIPALS algorithm [25]. The statistical parameters like the coefficient of determination (R^2), Standard Deviation Error in Calculation (SDEC), Standard Deviation Error in Prediction (SEDP), and F-ratio test were used to compute the overall significance of model (Equations 2.6-2.8) moreover. The CoMFA color contour maps are derived for the steric and electrostatic fields.

$$SDEC = \left[\frac{\sum_{i=1}^n (y_{obs,i} - y_{calc,i})}{n} \right]^{\frac{1}{2}} \quad (2.6)$$

$$SDEP = \left[\frac{\sum_{i=1}^n (y_{obs,i} - y_{pred,i})}{n} \right]^{\frac{1}{2}} \quad (2.7)$$

$$F (n, n-p-1) = \frac{(n-p-1)R^2}{p(1-R^2)} \quad (2.8)$$

Where $y_{obs,i}$ is the experimental activity, $y_{calc,i}$ is the estimated y in the calibration step and $y_{pred,i}$ predicted activity of the test set. The value corresponding to n and p is the number of samples in the training set and the number of components in the PLSR model, respectively.

2.8.4. Validation of CoMFA PLSR Model

Three main Cross-Validations (CV) techniques such as Leave-One-Out (LOO), Leave-Two-Out (LTO), and Leave-Many-Out (LMO) were used to explore the reliability of statistical models. In LOO CV, each time one compound is removed from the original training set, and a new model is built based on the rest of the set and this model is used to predict the activity of the omitted one. This procedure is repeated for the whole compounds of the data set. In LTO CV, two compounds are removed instead of one, and the remaining procedure repeated as same as that of LOO. In the LMO method, each time, 20% of compounds were removed randomly, and the procedure was repeated 20 times and predicted their activities via the reduced model. Golbarikh and Tropsha

reported that the LMO CV is much more robust than LOO CV, and also, a high value of Q^2 is essential and important but not adequate for a predictive model. The cross-validated R^2 (Q^2) is given in Equation 2.9.

$$Q^2 = 1 - \frac{PRESS}{SSY} \quad (2.9)$$

The term PRESS is the sum of the squared difference between experimentally observed activity and the activity predicted by a regression model estimated when the i^{th} sample was left out from the training set, and the SSY is the sum of squared differences between the experimental activity and the average experimental activity. According to Hawkins *et al.*, a valid statistical model should have a high Q^2 value ($Q^2 > 0.5$) and is evidence of the high predictive ability of the model [26]. The predictive power of the generated model was further evaluated by using an external test set. The predictive correlation coefficient (R^2_{pred}) was determined according to Equation 2.10.

$$R^2_{\text{pred}} = \left(\frac{SD - PRESS}{SD} \right) \quad (2.10)$$

where Standard Deviation (SD) is defined as the sum of the square deviation between the experimentally observed activity of the test set compounds and the mean activity of the training set molecules.

2. 8. 5. Model Validation through Progressive Scrambling

Y-Scrambling (Y-randomisation) was applied to exclude the probability that our CoMFA model performance could have occurred by chance. The Y-vector of the compounds in the training set is sorted according to decreasing Y value and then blocked into bins according to the value of the max_bins parameter. Afterward, attempts were made to scrambling Y values inside each bin many times to get its own permuted solution. The shuffling within blocks is repeated 20 times at each binning level. For each scrambling, a PLS and a CV model are computed.

2.9. Machine Learning Modeling

Machine learning techniques show outstanding performance in QSAR modeling, where the relationship between structure and activity is often complex and nonlinear [27]. A wide range of machine learning algorithms has been used to build QSAR classification models from an input data set of molecular descriptors and activity labels. For example, decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and naive Bayes classifiers, *etc.*, are different machine learning techniques to solve a classification problem. Each technique adopts a learning algorithm to identify a model that best fits the relationship between the descriptor set and the class label of the input data.

2.9.1 Machine Learning Algorithms

1. Support vector machine

Support vector machines (SVM) [28] are a bunch of supervised learning methods mostly used for classification and regression challenges. Classification of data using a support vector machine involves looking for a hyperplane in high dimensional space of independent variables that separate positive and negative data at an optimal distance using a nonlinear kernel function. SVM methodology purely laid on maximizing the margin between a small subset of training instances (the support vectors) and the hyperplane. SVM methods are one of the most popular machine-learning methods in chemoinformatics. The choice of a correctly configured kernel function is an important parameter to a successful SVM model. Polynomial kernel and Radical basis function (RBF) kernel are the two widely used kernels for solving classification problems. We have used both kernels in this study.

2. k-nearest neighbors

k-nearest neighbors (k-NN) [29] is a non-parametric lazy learning algorithm mainly used for classification and regression problems. The principle behind nearest neighbor methods is to make predictions for instances by searching through the entire training set for the k most similar instances (k-nearest neighbor) where k is a positive integer, or it may be chosen via cross-validation. In chemical applications, instances are mainly molecules and are described as

position vectors in the high dimensional feature space. Neighbors are identified based on distance in the feature space. This distance can, in general, be any metric measure; a commonly used distance metric for continuous variables is 'Euclidean Distance.' The optimal choice of k depends upon the input data. The higher the k values, the lower the noise effect on the classification, but it make boundaries between classes less distinct. Various heuristic techniques can select a good k .

3. Logistic regression

Logistic regression analysis calculates the probabilities of the outcome of a dependent variable using a logistic function to study the association between a categorical dependent variable and a set of independent (explanatory) variables. When we are using the logistic distribution, we usually make an algebraic conversion to arrive at our usual linear regression equation. It is a classification method mostly used for binary classification, although multinomial logistic regression is usually reserved for the case when the dependent variable has two or more unique values.

4. Decision Tree (J48)

Decision Trees (DTs) are simple and widely used supervised learning approaches used in statistics, data mining, and machine learning for classification problems. This model can predict the value of dependent variables by learning simple decision rules deduced from the data features. Decision tree builds classification models in the form of a tree structure includes a root node, branches, and leaf nodes,

where leaf nodes represent class labels, branches represent the outcome of a test, and root node represents the entire population.

2.9.2. Model Performance Evaluation

Performance evaluation of the classification model is an obvious task for understanding the accuracy of the model, to purify the model, and for choosing the appropriate model. So, this study evaluated the performance of the model by using 5-fold cross-validation and external set prediction. The performance evaluation of classification models can be based on a confusion matrix and receiver operating characteristic curve (ROC). All models were appraised by the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Further, Classification Accuracy (CA), sensitivity (SEN), and specificity (SPE) were also extracted from the confusion matrix by the following Equations 2.11-2.13:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2.11)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2.12)$$

$$CA = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (2.13)$$

The overall CA is the total percentage of both active and inactive compounds that were correctly predicted. Sensitivity is the

proportion of actual active compounds that are correctly identified among all the compounds. Specificity is the proportion of actual inactive compounds that are correctly identified among all the compounds. For a model to be good, the value of all these parameters should be close to zero. Another parameter, the Area Under the ROC Curve (AUC), is often used as a measure of the quality of the classification models. The value of AUC normally ranges from 0.5 to 1. A random classifier has an AUC of 0.5 (no discriminative power); on the other hand, AUC for a perfect classifier is equal to 1.

2.10. Homology Modeling

Homology modeling, also known as comparative modeling or template-based modeling (TBM), is a process of constructing a protein 3D model (target) from its amino acid sequence using a known experimental 3D structure of a homologous protein (template) [30]. Homology modeling is expanding the druggable genome's structural coverage [31]. The possible motive behind homology modeling is the lack of knowledge of 3D structures and the possible delay in the experimental elucidation of protein structure due to crystallization-related difficulties.

The step involved in the process of homology modeling are as follows: (i) identification of proper template using BLAST search, (ii) amino acid sequence alignment the of the target protein with that of the selected template, (iii) alignment corrections, (iv) backbone generation, (v) loop modeling, (vi) side chain generation and optimization, (vii) Refinement and optimization of the model (viii)

model verification using quality criteria. Homology modeling is a simple process, but it can also be quite challenging. Today, homology modeling is one of the most common techniques that can be used to build accurate 3D structural models of proteins.

2.11. Molecular Docking

Molecular docking methodology has been used to predict the best binding orientation of ligand molecule in the active site of receptor targets, and recently it is the most used computational tool for drug designing and virtual screening. A valid three-dimensional structure of the molecular target obtained either in the form of the experimentally resolved crystal structure (e.g., X-ray crystallography or NMR) or computationally generated models (e.g., homology model) is a mandatory requirement for molecular docking. Understanding the position of the ligand-binding site before docking processes improves the efficacy of docking significantly. In many instances, the binding site is known from the co-crystallized ligand. If there is no co-crystallized ligand in protein (apoprotein), information about the binding sites can be obtained by comparing the target protein with a family of proteins sharing a similar function. If no knowledge of the binding sites is available, cavity detection programs or online servers can be used for the identification of probable active sites within proteins such as POCASA [32], SiteMap [33], PASS [34] and so on. Docking without any assumption of a binding site is called blind docking. Based on the types of ligand, docking can be classified as protein-small molecule docking, protein-nucleic acid docking, and

protein-protein docking. Protein–small molecule (ligand) docking represents a more straightforward and most commonly used.

2.11.1. Theory of Docking

The objective of molecular docking is to predict the preferred relative orientation of a ligand when bound to the active site of a protein to form a stable complex in such a way that the free energy of the entire system is minimized. Two steps can achieve this process, namely sampling the ligand conformation at the active site of the protein, and then ranking these conformations through a scoring function.

2.11.2. Search / Sampling Algorithms

There is a huge number of possible binding modes between the two molecules while taking into account all translational and rotational conformational degrees of freedom. The size of the search space is increasing exponentially as the size of the molecules increases. In practice, however, it is impossible to explore all the search space in depth using current computational resources. Different sampling algorithms have been developed and widely used in molecular docking software that can list an optimal number of ways to put together two molecules such as Simulated Annealing, Fast shape matching, Incremental construction, Particle Swarm Optimization (POS) and Evolutionary algorithms.

2.11.3. Scoring Functions

Scoring functions are mathematical functions that have been used to measure and rank the ligand-receptor complex in docking. The efficiency of molecular docking depends on the accuracy of the adopted scoring method that can direct and decide the pose of the ligand while producing thousands of possible poses of the ligand. It also needs to give every ligand the appropriate relative rank in the database. Ideally, the score should directly correspond to the binding affinity of the ligand for the protein so that the top-scoring compounds are also the best binders. Three main types of scoring functions are used by various docking programs: force field, empirical and knowledge-based.

1. **Force field-based:** Force-field method is based on molecular mechanics in which the potential energy of a system is considered as a combination of bonded (intramolecular) and nonbonded (intermolecular) components. The scoring functions of the docking software GOLD, Autodock, and DOCK, are Force field-based.
2. **Empirical scoring:** These scoring functions are obtained by summing various empirical energy terms such as van der Waals, electrostatic, hydrogen bond, solvation, entropy, hydrophobicity, *etc.*, which are weighted by coefficients gained through regression analysis utilizing known binding affinity data of experimentally determined structures by the least-

squares fitting. Examples of empirical scoring functions include ChemScore, Glide SP/XP score, and ID-Score.

3. **Knowledge-based:** these scoring functions are built based on the statistical analysis of interacting atom pairs of the protein-ligand complexes from the accessible three-dimensional structures. Such pairwise-atom data are translated to a pseudopotential, also known as a mean force potential describing the desired geometries of the pairwise atoms of the protein-ligand. Examples include PMF and DrugScore

The fourth type of scoring function recently introduced that incorporates the nonlinear regression machine-learning method.

2. 12. Softwares

The backbones of this study are some software that is used to perform all the computations. Therefore, in this section, a detailed explanation of the most important software used in this study is given. Software that is not mentioned here can be viewed in the respective chapters.

2.12.1. *Molecular Modeling Software*

1. **Gaussian**

Gaussian is a computational chemistry software package developed by John Pople and his research group at Carnegie Mellon University as early as in the 1970s in the name Gaussian 70. This software has been updated continuously since then. Different methods

of calculations available are simple molecular mechanics (such as Amber force field), semi-empirical methods (such as CNDO), Hartree-Fock (restricted and unrestricted), Moller-Plesset perturbation theory of order $n=2, 3, 4$ (MP n), Configuration-Interaction (CI), Coupled-Cluster (CC), Multi-configurational SCF (such as CAS-SCF) and various DFT methods. Gaussian software helps to do electronic-structure calculations and quantum chemical calculations.

2.12.2. Molecular Editors and Visualization Tools

1. Open Babel 2.3

Open babel [35] is an open-source collaborative chemical toolbox which allows to search, store, analyze or convert chemical data for research in the area such as cheminformatics, bioinformatics, organic chemistry, materials science, and computational chemistry. Open Babel version 2.3 interconverts over 110 formats. This software has been used throughout this work for file type conversion.

2. MayaChemTools

MayaChemTools [36] is an open-source collection of Perl and Python scripts and modules designed to facilitate a variety of everyday computational discovery requirements such as data manipulation and analysis, 2D fingerprint generation, similarity search, and physicochemical properties calculation. This software was mainly used in this study to split large-sized SDF files and to generate 2D fingerprints for analyzing chemical similarity.

3. GaussView 5

GaussView is the Gaussian graphical user interface. It helps to create Gaussian input files, allows the user to run Gaussian calculations from a graphical interface without having to use a command-line instruction, and helps to interpret Gaussian output. This study used GaussView to generate inputs for the optimization of compound geometry.

4. Maestro

Maestro is the user interface for all applications from Schrödinger. It can be used for long-standing visualization, design, generation, and analysis of compounds. It also facilitates the organization and analysis of chemical data by researchers. In this thesis, Maestro is used to generating a 2D line diagram and 3D images of molecules and protein-ligand complexes.

5. AutoDockTools

AutoDockTools (ADT) [37] is a free graphical user interface of AutoDock or AutoDock vina used for setting up, running, and analyzing AutoDock docking. It can also be useful for grid generation, computing molecular surfaces, displaying secondary structure ribbons, computing hydrogen bonds, *etc.* We used ADT for all the purposes mentioned above.

6. UCSF Chimera

UCSF Chimera [38] is an open-source molecular modeling software extensively used for interactive visualization and analysis of molecular structures (mainly proteins) and related data, including density maps, supramolecular assemblies, multiple sequence alignments, docking results, molecular dynamic trajectories, and conformational ensembles. It can also be used to generate high-quality images and animations. The primary purpose of the Chimera in this study was the preparation, alignment, and visualization of proteins.

7. PyMOL

Warren Lyford DeLano created PyMOL as an open-source molecular visualization tool. This program can generate high-quality 3D images of small molecules and biological macromolecules. Schrödinger, Inc is currently marketing it. However, Linux distributions continue to provide their builds of the open-source code. In this study, PyMOL is used to visualize the molecular alignment and CoMFA contours.

8. LIGPLOT

LIGPLOT [39] is a program that generates schematic 2D ligand-protein interaction diagrams automatically when we provide the ligand-protein complex. It runs from an intuitive java interface that enables the plots to be edited on-screen through mouse click-and-drag operations. The interactions provided by LIGPLOT are those mediated by hydrogen bonds and hydrophobic contacts. In this study, LIGPLOT

was used to obtain clear and concise images of the ligand-protein complex.

2.12.3. *Datamining, Visualisation, QSAR Modeling software*

1. WEKA Computations

The WEKA (Waikato Environment for Knowledge Analysis) [40] is a collection of machine learning algorithm package for data mining tasks, developed at Waikato University, New Zealand with an open-source issued under the GNU General Public License. Weka has been used for data preprocessing, data manipulation, classification, regression, and clustering. This study used WEKA version 3.8 for the selection of suitable descriptors and best-performing algorithm. For this purpose, many algorithms available in WEKA with its default settings were initially trained, and then the algorithms that turned out to perform at best were re-trained by more finely tuning the parameter values in order to maximize the algorithm performance.

2. DataWarrior

DataWarrior [41] is a free and versatile cheminformatics program for the exploration, analysis, and visualization of extensive chemical and biological data. DataWarrior combines dynamic graphical views and interactive row filtering with chemical intelligence. This software allows the decoding of chemical descriptors and fingerprints, the measurement of molecular similarity, the clustering, the analysis of diversity, the enumeration of combinatorial

libraries. DataWarrior can be used to explore chemical space, activity landscapes, and activity cliffs interactively.

3. PUMA

The platform for Unified Molecular Analysis (PUMA) [42] is a free interactive online platform for cheminformatic-based chemical space visualization and diversity analysis. The platform connects two public online tools such as Activity Landscape Plotter to evaluate structure-activity relationships and Consensus Diversity Plots for measuring global diversity. Both DataWarrior and PUMA were widely used in this study to analyze the chemical space of 5-LOX inhibitors.

4. Canvas

Canvas is a cheminformatics computing environment with features include custom visualization, unparalleled fingerprinting capabilities with seven types of hashed fingerprints, property calculations, clustering, classification, diversity analysis, data reduction QSAR, scaffold decomposition, R-group analysis, and Ultra-fast substructure searching.

5. Open3DQSAR

Open3DQSAR is an open-source platform for pharmacophore analysis based on MM and QM Molecular Interaction Fields (MIFs) by high-throughput chemometric analysis. It can generate steric potential, electron density, and MM/QM electrostatic potential fields. Open3DQSAR is controlled through a command-line interface.

Interface to all major molecular modeling software makes Open3DQSAR a powerful tool in pharmacophore assessment and ligand-based drug design.

2.13. Computer Power

All the calculations included in this thesis are performed using Lenovo Thinkstation with processor Intel®Xeon®CPU E5-2660 v3 @2.60 GHz and 32 GB RAM.

References

- [1] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 28 (1988) 31–36. doi:10.1021/ci00057a005.
- [2] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Comput. Sci.* 29 (1989) 97–101. doi:10.1021/ci00062a008.
- [3] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, *J. Cheminform.* 7 (2015) 23. doi:10.1186/s13321-015-0068-4.
- [4] K.I. Ramachandran, G. Deepa, K. Namboori, *Computational chemistry and molecular modeling: principles and applications*, Springer Science & Business Media, 2008.
- [5] D. Young, *Computational chemistry: a practical guide for applying techniques to real world problems*, John Wiley & Sons, 2004.
- [6] C.J. Cramer, *Essentials of computational chemistry: theories and models*, John Wiley & Sons, 2013.
- [7] R. Todeschini, V. Consonni, *Handbook of molecular descriptors*, John Wiley & Sons, 2008.
- [8] I. Muegge, P. Mukherjee, An overview of molecular fingerprint similarity search in virtual screening, *Expert Opin. Drug Discov.* 11 (2016) 137–148. doi:10.1517/17460441.2016.1117070.
- [9] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure-activity studies: definition and applications, *J. Chem. Inf. Comput. Sci.* 25 (1985) 64–73. doi:10.1021/ci00046a002.
- [10] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, *J. Chem. Inf. Comput. Sci.* 42 (2002) 1273–1280. doi:10.1021/ci010132r.
- [11] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities, in: R.A. Wheeler, D.C.B.T.-A.R. in C.C. Spellmeyer (Eds.), Elsevier, 2008: pp. 217–241. doi:https://doi.org/10.1016/S1574-1400(08)00012-1.

- [12] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron., 2015: pp. 1200–1205. doi:10.1109/MIPRO.2015.7160458.
- [13] B. Kumari, T. Swarnkar, Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review, *Int. J. Comput. Sci. Inf. Technol.* 2 (2011) 1048–1053.
- [14] M.A. Hall, *Correlation-based Feature Selection for Machine Learning*, (1999).
- [15] B. Azhagusundari, A.S. Thanamani, Feature selection based on information gain, *Int. J. Innov. Technol. Explor. Eng.* 2 (2013) 18–21.
- [16] P. Willett, J.M. Barnard, G.M. Downs, Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.* 38 (1998) 983–996. doi:10.1021/ci9800211.
- [17] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, Molecular Similarity in Medicinal Chemistry, *J. Med. Chem.* 57 (2014) 3186–3204. doi:10.1021/jm401411z.
- [18] J.J. Naveja, J.L. Medina-Franco, Insights from pharmacological similarity of epigenetic targets in epipolypharmacology, *Drug Discov. Today*. 23 (2018) 141–150. doi:https://doi.org/10.1016/j.drudis.2017.10.006.
- [19] R.D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.* 110 (1988) 5959–5967. doi:10.1021/ja00226a005.
- [20] M. Clark, R.D. Cramer, D.M. Jones, D.E. Patterson, P.E. Simeroth, Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases, *Tetrahedron Comput. Methodol.* 3 (1990) 47–59. doi:https://doi.org/10.1016/0898-5529(90)90120-W.
- [21] P. Tosco, T. Balle, Open3DQSAR: A new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields, *J. Mol. Model.* 17 (2011) 201–208. doi:10.1007/s00894-010-0684-x.
- [22] P. Tosco, T. Balle, F. Shiri, Open3DALIGN: An open-source software aimed at unsupervised ligand alignment, *J. Comput. Aided. Mol. Des.* 25 (2011) 777–783. doi:10.1007/s10822-011-9462-9.

- [23] M. Pastor, G. Cruciani, S. Clementi, Smart region definition: A new way to improve the predictive ability and interpretability of three-dimensional quantitative structure-activity relationships, *J. Med. Chem.* 40 (1997) 1455–1464. doi:10.1021/jm9608016.
- [24] S.C. Massimo Baroni, Gabriele Costantino, Gabriele Cruciani, Daniela Riganelli, Roberta Valigi, Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems, *Quant. Struct. Relationships.* 12 (1993) 9–20. doi:10.1002/qsar.19930120103.
- [25] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.
- [26] D.M. Hawkins, S.C. Basak, D. Mills, Assessing Model Fit by Cross-Validation, *J. Chem. Inf. Comput. Sci.* 43 (2003) 579–586.
- [27] A. Lavecchia, Machine-learning approaches in drug discovery: Methods and applications, *Drug Discov. Today.* 20 (2015) 318–331. doi:10.1016/j.drudis.2014.10.012.
- [28] C. Cortes, V. Vladmir, Support Vector Networks, *Mach. Learn.* 20 (1995) 273–297. doi:10.1007/BF00994018.
- [29] D.W. Aha, D. Kibler, M.K. Albert, Instance-Based Learning Algorithms, *Mach. Learn.* 6 (1991) 37–66. doi:10.1023/A:1022689900470.
- [30] S. Abeln, K.A. Feenstra, J. Heringa, Protein Three-Dimensional Structure Prediction, in: S. Ranganathan, M. Gribskov, K. Nakai, C.B.T.-E. of B. and C.B. Schönbach (Eds.), Academic Press, Oxford, 2019: pp. 497–511. doi:https://doi.org/10.1016/B978-0-12-809633-8.20505-0.
- [31] S. Hongmao, Chapter 4 - Homology Modeling and Ligand-Based Molecule Design, in: S.B.T.-A.P.G. to R.D.D. Hongmao (Ed.), Woodhead Publishing, 2016: pp. 109–160. doi:https://doi.org/10.1016/B978-0-08-100098-4.00004-1.
- [32] J. Yu, Y. Zhou, I. Tanaka, M. Yao, Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere, *Bioinformatics.* 26 (2009) 46–52. doi:10.1093/bioinformatics/btp599.
- [33] Schrödinger Release 2018-3: SiteMap, Schrödinger, LLC, New York, NY, (2018).

- [34] G.P. Brady, P.F.W. Stouten, Fast prediction and visualization of protein binding pockets with PASS, *J. Comput. Aided. Mol. Des.* 14 (2000) 383–401. doi:10.1023/A:1008124202956.
- [35] N.M. O’Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminform.* 3 (2011) 33. doi:10.1186/1758-2946-3-33.
- [36] M. Sud, MayaChemTools: An Open Source Package for Computational Drug Discovery, *J. Chem. Inf. Model.* 56 (2016) 2292–2297. doi:10.1021/acs.jcim.6b00505.
- [37] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J. Comput. Chem.* 30 (2009) 2785–2791. doi:10.1002/jcc.21256.
- [38] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera—A visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612. doi:10.1002/jcc.20084.
- [39] A.C. Wallace, R.A. Laskowski, J.M. Thornton, LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions, *Protein Eng. Des. Sel.* 8 (1995) 127–134. doi:10.1093/protein/8.2.127.
- [40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, *SIGKDD Explor.* 11 (2009) 10–18.
- [41] T. Sander, J. Freyss, M. Von Kor, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.* 55 (2015) 460–73. doi:10.1021/ci500588j.
- [42] M. González-Medina, J.L. Medina-Franco, Platform for Unified Molecular Analysis: PUMA, *J. Chem. Inf. Model.* 57 (2017) 1735–1740. doi:10.1021/acs.jcim.7b00253.

3

PROTEIN STRUCTURE EVALUATION AND MODELING

3.1. Introduction

The three-dimensional (3D) structure of the biological target, obtained from X-ray, NMR, or computational modeling, usually binds with other small molecules (ligand) either to enhance or to inhibit a biological function. The ligand that inactivates the protein target is called Antagonists, while ligands that activate the protein target are called agonists [1]. A few essential residues are generally engaged in all these protein-ligand interactions. In order to comprehend the protein function, it is essential to define the locations of these interacting residues. For which, we need to have thorough knowledge about the selected biological target and its mechanism of action. This study has selected 5-LOX as a therapeutic target. Therefore, we should have a deep understanding of the structure and function of this protein. There may be several underlying problems related to 5-LOX crystal structure that should be examined before further research; for instance, receptor druggability, choice of the binding site, selection of the most appropriate protein structure, incorporation of receptor flexibility,

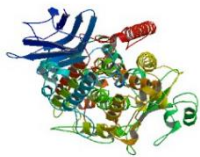

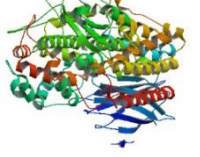

proper allocation of protonation states, and consideration of water molecules in the binding site. Also, the identification of ligand binding sites on 5-LOX is becoming increasingly important because of the lack of co-crystallized ligand in the active site. Besides, most of the *in-vitro* analysis has been carried out in *Rattus norvegicus* (Rat) 5-LOX protein because of its amino acid sequence similarity to the Human 5-LOX, and hence we have tried to develop a comparative model/ homology model of Rat 5-LOX protein. So, in this Chapter, we are going to discuss the preparation, evaluation, and identification of the ligand-binding site of 5-LOX protein by addressing all the issues mentioned above. Also, we have tried to make a comparative model for Rat 5-LOX protein.

3.2. Evaluation of Human 5-LOX Protein Crystal Structure

5-LOX protein has four 3D structures so far reported in the PDB database, including a substrate (AA) bound form and three substrate-free structural forms (Table 3.1). The substrate-free 5-LOX structures are almost similar (RMSD = 0.3 Å) and are represented as PDB ID 3O8Y, 3V92, and 3V98 [2,3]. The structure '3O8Y' is a stable 5-LOX protein; additional mutation to 3O8Y could lead to a phosphorylated mimic Ser663Asp (3V98) or a non-phosphorylated one Ser663Ala (3V92) structures. The substrate 'Arachidonic Acid (AA)' bound form of 5-LOX is represented as PDB ID, 3V99 [3] (the AA: Stable 5-LOX-Ser663Asp mutant). However, a few of the residues in the crystal structure of 3V99 do not have X-ray coordinates (flexible loops from residues 172 to 176, 190 to 198, 294 to 299, 611 to 613, α

helix from residues 414 to 429, and terminal Ile 673), and the partial density does not permit to explicitly determine the orientation of the substrate in the binding site. Although Substrate Arachidonic acid (AA) was present in the 3V99 structure, 3V99 was not used as a target structure in this study due to all the problems mentioned above. Thus, the stable 3D structure of 5-LOX (3O8Y) with a resolution of 2.389 Å was finalized as the target structure for this study.

Table 3.1 The crystal structures of the 5-LOX enzyme with basic information

Protein Structure	PDB Code	Description	Chain	Position	Resolution (Å)
	3V98	S663D Stable 5-LOX	A/B	1-674	2.070
	3V99	S663D Stable 5-LOX in complex with Arachidonic Acid	A/B	1-674	2.252
	3O8Y	Stable 5-LOX	A/B	1-674	2.389
	3V92	S663A Stable 5-LOX	A/B	1-674	2.740

The protein structure obtained from the PDB was not suitable in its native state for molecular docking. Therefore, optimization and energy minimization of the protein were necessary. It has been done using the protein preparation wizard of the Schrödinger suite.

3.3. Protein Preparation

Protein Preparation Wizard module in Schrödinger [4] helps to Protein preparation and refinement. This process includes adding hydrogens and disulfide bridges, removing crystallographic water molecules and ions, fixing bond orders, assigning partial charges with the OPLS 2005 force field [5]. Initially, H-bond was optimized, then the whole protein structure was allowed to relax, and subsequently, the receptor protein was minimized by applying OPLS 2005 force field implemented with Root Mean Square Deviation (RMSD) value of 0.30 Å° for energy minimization of protein to ensure the stability and quality for further studies.

3.4. Ligand Binding Site Identification

Due to the lack of bound ligand (co-crystallized ligand) in the crystal structure of 5-LOX, binding site identification is necessary before docking. Many reports in literature mention possible active site amino acid residues of 5-LOX, and some even try to compare the crystal structure of 5-LOX with other enzymes in the lipoxygenase family to identify the active site based on similarity [6,7]. These kinds of literature reports predicted that an elongated cavity surrounded by 5-LOX specific amino acids such as Tyr 181, Ala 603, Ala 606, His 600, and Try 364 is the substrate-binding site of the 5-LOX crystal structure, whereby they help to bind ligand. Residues like Leu- 368,

373, 414, 607, and Ile-406 are conserved in all AA-metabolizing LOX are also preserved in the 5-LOX. These amino acids function as preserved hydrophobic side chains that can handle the substrate's or ligand's hydrophobic part. We have confirmed the ligand-binding site using a binding site prediction tool POCASA (POcket and CAvity Search Algorithms). Besides, we also tried to characterize and understand the nature of the binding cleft using the SiteMap module of Schrödinger.

POCASA can generate the shapes of pockets and cavities that coincide well with the bound ligand using an algorithm 'ROLL' [8]. The key concept of ROLL is to generate a crust-like surface called the probe surface enveloping protein to identify the region between the probe surface and protein surface as a 'pocket' and the region surrounded by protein surface as a 'cavity.' The grid points occupied by protein atoms are labeled as protein points with a value of 1, and free grid points have a value of 0. In the next step, ROLL uses a rolling probe sphere to generate the 'probe surface.' Once the probe encounters the first protein point, it begins to roll along the protein surface without any overlap with protein until it returns to the starting position. The rolling direction is controlled based on the inner border tracing algorithm. By adjusting the probe radius, ROLL can predict various binding sites.

Default parameters in POCASA of grid size = 1Å, probe radius = 2Å, Single-Point Flag (SPF) = 16 and Protein-Depth Flag (PDF) = 18 were used for calculation. We have identified nine binding pockets and one cavity in the 5-LOX protein. The information of predicted pockets and cavities in the order of their rank is given in Table 3.2,

including volume, Volume-Depth (VD) value, and average VD values. Colored grid map of nine binding pockets and one binding cavity is shown in Figure 3.1A. The violet pocket is the top-ranked one followed by blue, green, yellow, orange, pink, forest green, and black, and the binding cavity is represented by cyan color. Here, 'Volume Depth (VD),' a new parameter which was designed in Roll for the more accurate ranking of pockets, is considered for the ranking. The VD value of the pocket will be determined by summing the depth of all pocket points where the depth of the pocket point is defined as the shortest distance from a pocket point to the probe surface. The cavity obtained with an average VD value of 13.015 and a VD value 1158 can be taken as the active site. This predicted cavity is found to be surrounded by active site amino acid that is already reported by Brash *et al.*, [2]. As a result, this region of the cavity defined by the grid box near the catalytic iron with a volume of 89\AA^3 could be used to find out the suitable binding poses of the ligand. However, some difficulties are observed while the ligand is coming into this cavity. Since the passage into this 5-LOX active site cavity is blocked, a conformational change in the active enzyme should occur upon ligand binding. Otherwise, ligands tend to go to the other most favorable and easily accessible binding pockets listed in Table 3.2. Therefore, we need to give extra care to the grid box assignment before conducting a docking procedure. Defining the grid box in such a way that the first goal of the ligand should be to reach this cavity instead of entering other pockets was therefore necessary. For this purpose, the known 5-LOX inhibitor 'zileuton' was docked to the binding site of 5-LOX using Autodock vina with the gradient optimization algorithm and optimizing the dimension of the grid box to $20 \times 20 \times 25 \text{\AA}$ cube at -8.374, 66.379, -

1.009 for x, y, and z respectively using AutoDockTools (ADT). This setup allows the well-fitted molecules to enter into the binding site of 5-LOX and dismiss molecules which are not well fitted. Figure 3.1B shows a close view of zileuton at the active site of 5-LOX.

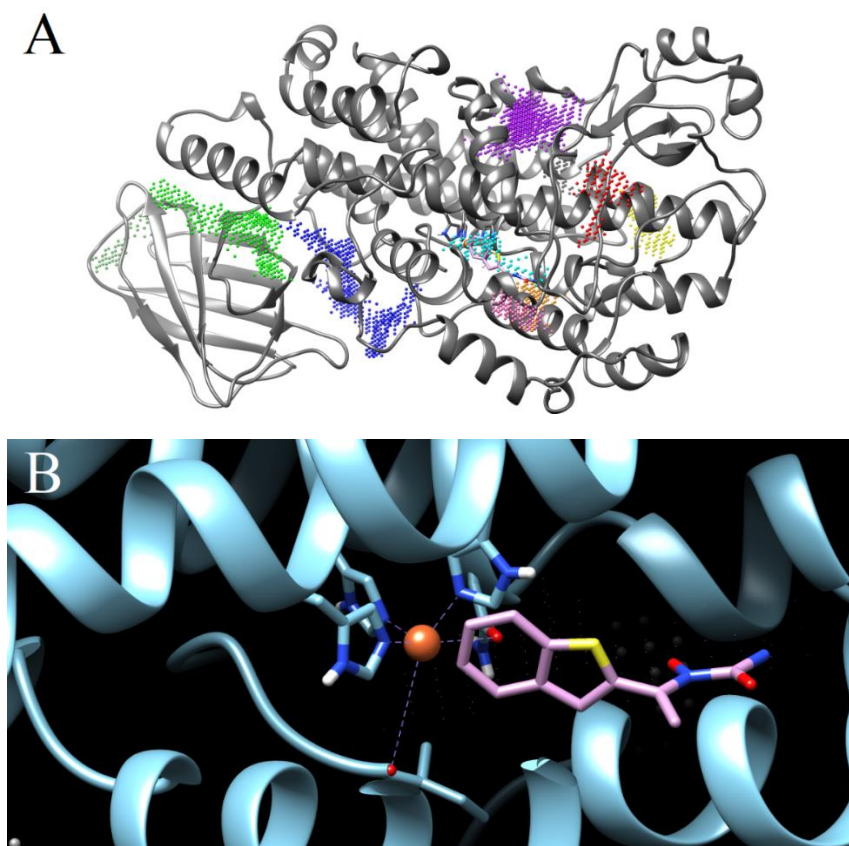


Fig. 3.1 A) The 5-LOX with top 9 ligand-binding pocket pockets and one binding cavity obtained using POCASA; the violet pocket is the top-ranked one followed by blue, green, yellow, orange, pink, forest green and black and the binding cavity is represented by cyan color. B) Close view of zileuton at the active site of 5-LOX. The orange sphere represent Fe atom at the active site.

Table 3.2 Top ten ligand-binding pocket or cavity of Human 5-LOX protein

Rank No.	Types of binding site	Pocket No./Cavity No.	VD value (Å)	Average VD (Å)
1	Pocket	267	1824	4.470
2	Pocket	724	1486	3.325
3	Pocket	103	1310	4.503
4	Pocket	881	531	3.382
5	Pocket	443	311	2.616
6	Pocket	281	242	2.444
7	Pocket	213	208	3.151
8	Pocket	77	203	2.417
9	Pocket	1002	152	2.375
1	Cavity	364	1158	13.015

3.5. Druggability Assessment of 5-LOX Protein

Hopkins & Groom had previously defined the term "druggable" as a protein capable of binding drug-like molecules or which was predicted to bind drug-like molecules with high affinity [9]. Although zileuton, a 5-LOX inhibitor, is thought to have reached the market, it has several side effects, including liver toxicity and unfavorable pharmacokinetics. This is merely a reflection of the challenge faced by medicinal chemists in the development of new 5-LOX inhibitors. In this section, the catalytic pocket of 5-LOX enzymes has been analyzed and tested for druggability in order to reach the full picture.

The SiteMap [10–12] module in Schrödinger Maestro, analyses the protein crystal structure 3O8Y to determine the druggability of the catalytic pocket or site. A "site" is characterized by a set of points on a grid that is either adjacent or bridged in the exposed regions by short

gaps. Contributions from the degree of enclosure/exposure, tightness, hydrophobic/hydrophilic character, and possibilities of hydrogen bonding were used for the assessment. Two crucial parameters SiteScore and Dscore, are derived as the weighted average of these measurements and are given as Equations 3.1 and 3.2, respectively.

$$\text{SiteScore} = 0.0733n^{1/2} + 0.6688e - 0.20p \quad (3.1)$$

$$\text{Dscore} = 0.094n^{1/2} + 0.60e - 0.324p \quad (3.2)$$

where n , e , and p respectively are the number of site points, the enclosure score, and the hydrophilic score. The two scoring functions use the same terms but with different coefficients, as shown in the equations, which makes them highly polar active sites more vulnerable to the Dscore function than SiteScore. Thus, the polar pocket may have a low score in Dscore, while it may have a high score in SiteScore. Such a pocket is more likely to promote the binding of strongly polar ligands that are not drug-like.

If a binding pocket with a SiteScore value of ≥ 0.8 , then it is considered as possible binding pocket and if a binding pocket with a SiteScore of ≥ 1.0 , then it is considered as binding sites of particular importance. In the case of Dscore, extremely druggable protein should have a Dscore of ≥ 1.0 , the druggable protein might have a Dscore in between 0.8–1.0, intermediate druggable protein should have a Dscore in between 0.7–0.8 and non-druggable protein may have a Dscore of ≤ 0.7 [13,14]. Various druggability parameters of the five probable binding pockets identified by SiteMap are shown in Table 3.3.

Table 3.3 Predicted ligand-binding pockets and the various site parameters of 5-LOX (PDB entry 3O8Y)

Probable ligand binding sites	SiteScore	Size	Dscore	Volume
Site-A	1.197	95	1.211	144.403
Site-B	1.062	453	0.971	1015.28
Site-C	1.049	151	1.070	482.944
Site-D	1.022	100	1.036	198.940
Site-E	0.913	93	0.923	208.201

Three sites are found to be druggable by considering the Dscore parameter alone, while the inclusion of parameter SiteScore for the evaluation provides four sites having excellent binding site character. However, Site A is the only member found in the "very druggable" category with both SiteScore and Dscore values greater than 1.19. Another factor measured by SiteMap is the size of the active site; the size of the active site is defined as the number of points that make up the examined pocket. 'Site A' is similar to the ligand-binding cavity obtained from the POCASA. The same amino acids are found to be present on this site too. Also, the size of the site A (= 95 spheres) is smaller as compared to other sites, and is similar to the POCASA result. This observation indicates probable ligand binding sites obtained from SiteMap and POCASA are compatible with each other and confirms that Site A (in SiteMap)/cavity 1 (in POCASA) is the ligand-binding site of the 5-LOX protein.

The colored contours of five top-ranking surface pockets identified by SiteMap on 5-LOX are shown in Figure 3.2A. We have

already discussed that 'Site A,' the deepest pocket in the protein, is considered as the primary site. This site is shown in white with a site score of 1.197, having a volume of 144.403. The contours of second, third, fourth, and fifth binding pockets are colored in green, magenta, cyan, and orange, respectively.

The SiteMap shows binding site areas appropriate for hydrophobic groups or ligand hydrogen bond donors, acceptors, or metal binding features. Distinguishing the various sub-regions of the binding site enables the evaluation of the ligand-receptor complementarity. These maps can aid in the development of better ligands in lead-discovery studies by illustrating regions where the ligand and the receptor are not complementary. Figure 3.2B shows the SiteMap analysis of the most famous inhibitor of 5-LOX. The yellow, blue, and red regions respectively represent hydrophobic, hydrogen bond donor, and hydrogen bond acceptor regions. The hydrophobic phenyl part of zileuton correctly situated in the yellow-colored region. Likewise, hydrogen bond donating -OH group located on the blue contour while hydrogen bond acceptor C=O group is placed on the red contour. This observation indicates that zileuton satisfied all the structural requirements for receptor binding. From this observation, we can easily deduce the reason for the high inhibitory activity of zileuton. The active site in 5-LOX identified using Glide is relatively accurate and matches with ADT tools and could be adopted to perform further large-scale screening.

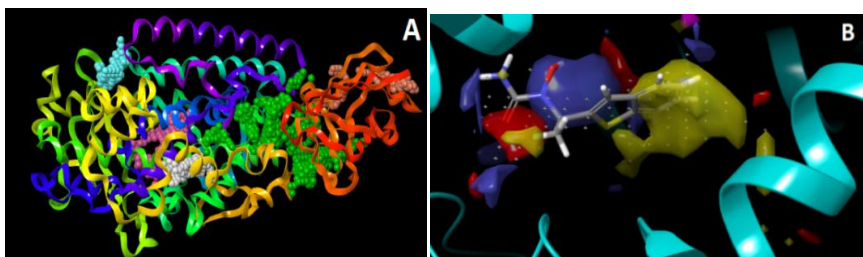
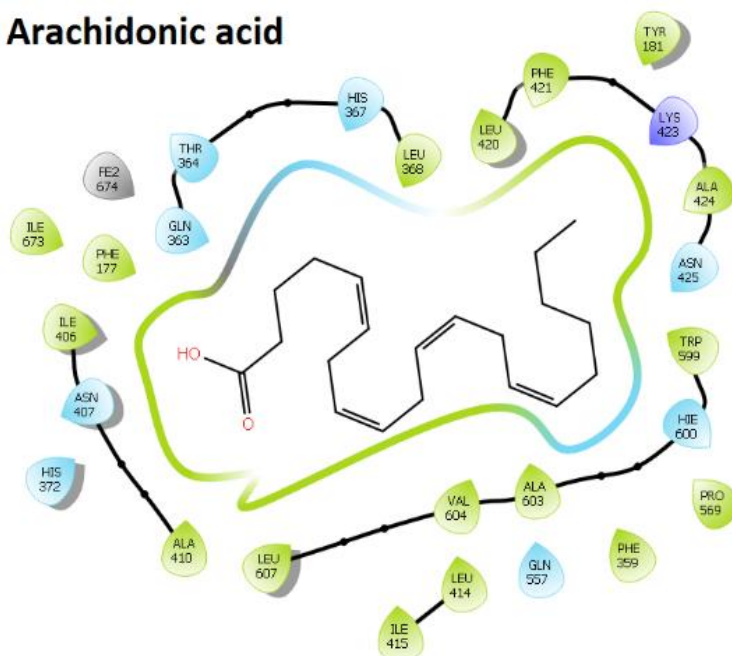


Fig. 3.2 A) The five top-ranking surface pockets identified by SiteMap on 5-LOX. Colored spheres represent the surface clefts on 5-LOX: site A: white, B: green, C: magenta, D: cyan, E: orange. B) SiteMap highlighted regions within the binding site (site A) suitable for occupancy by hydrophobic groups in yellow color or by ligand hydrogen-bond donors in blue color, acceptors in red color, or metal-binding functionality in magenta color.

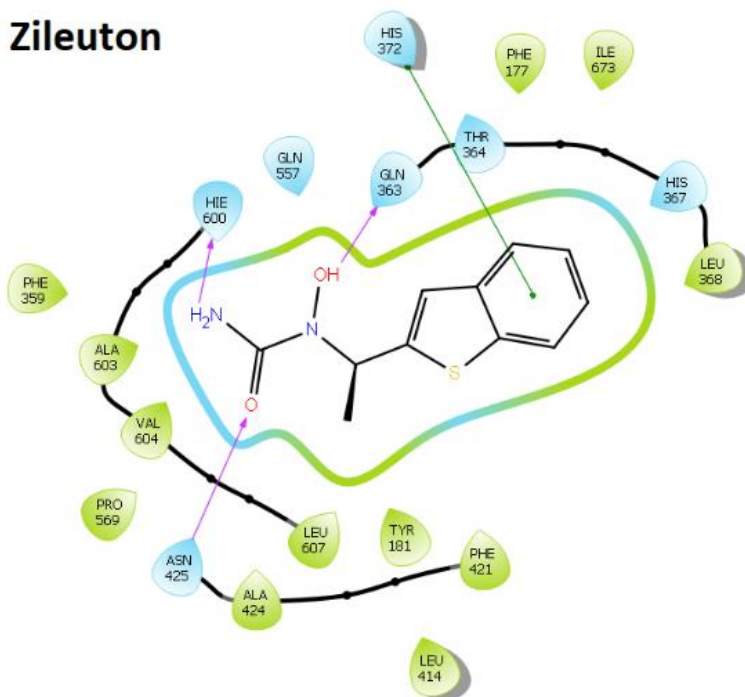
3. 6. Molecular Docking

The molecular docking was performed on a 5-LOX binding pocket using a collection of known 5-LOX antagonists to discover the predictability and binding characteristics of the 5-LOX binding pocket and to help facilitate a reasonable design of the novel 5-LOX antagonists. In order to evaluate the efficiency of docking analysis, AA was also docked. 5-LOX protein that has already been prepared was used as the target for this study. Glide module [4,15,16] of Schrödinger Suite with extra precision (XP) [17] was used as a docking algorithm. The 2D docked images of these 2 well known 5-LOX inhibitors and substrate AA at the active site of 5-LOX are shown in Figure 3.3.

Arachidonic acid



Zileuton



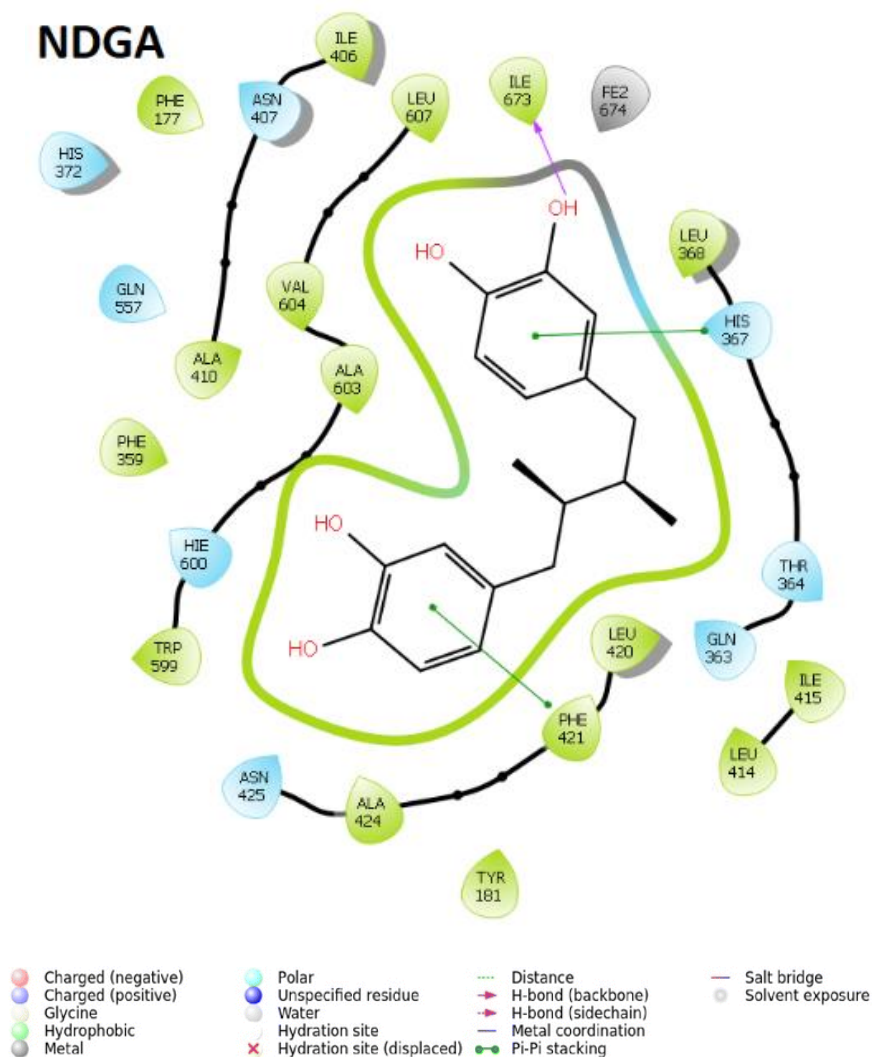


Fig. 3.3 Docking poses of the substrate arachidonic acid and 2 well known 5-LOX inhibitors (Glide Gscore of arachidonic acid, zileuton, NDGA, respectively is -8.509, -7.923, and -9.525 kcal/mol)

All molecules are well occupied in the active site of 5-LOX i.e., Site A (in SiteMap) / cavity 1 (in POCASA), and they all seem to block any access to the iron atom. The docking result also finds out that most of the interactions are hydrophobic because the 5-LOX catalytic center consists predominantly of the hydrophobic amino acids Leu 368, Ile 406, Leu 691, Ile 415, Ala 410, Leu 414, Val 604 and Leu 607 and some pi-pi stacking interactions are also observed. So, the presence of hydrophobic moieties is a mandatory characteristic of 5-LOX inhibitors, while a flexible structure capable of adapting curved conformations may facilitate stable complex formation as well. For example, the result of molecular docking shows that the Phe 177 amino acid, which also takes part in the complex stabilization of Fe, appears to interact with the carboxylic group of the AA. H-bond with amino acid-like Hie 600, Leu 420, His 367, Gln 363, Ile 673, Asn 425, His 372 in a small polar section of the pocket enabling the hydrogen bond and diminishing contact with the hydrophobic residues.

3.7. Homology Modeling Studies

In the last few decades, 5-LOX of Rat basophilic leukemia cells has been utilized for the screening of LOX inhibitors because of the high similarity between the Human and Rat enzymes. However, as far as we know, the three-dimensional crystal structure of 5-LOX of Rat or its homology model is unknown. In this situation, this study also intended to design a homology model of 5-LOX of Rat and use this three-dimensional model to carry out the binding studies by rigid-flexible docking.

The homology modeling of Rat 5-LOX was performed by utilizing the web-based homology modeling suit SWISS-MODEL [18]. The rigid fragment assembly modeling approach is used in SWISS-MODEL for homology modeling [18,19]. The amino acid sequence of Rat 5-LOX was collected from the databank in the National Center for Biotechnology Information, NCBI (NCBI Reference Sequence: NP_036954.1). The program Blasts [20] and HHblits [21] are used to select suitable templates structure for homology modeling. It automatically selects the Human 5-LOX crystal structure with a protein data bank (PDB) ID 3O8Y (chain A) from ExPDB as suitable templates that share 92.40% identity with a Rat 5-LOX as shown in Figure 3.4. Local pair-wise alignment of the target sequence to the main template structures has been calculated, followed by a heuristic step to improve alignment for modeling purposes [18]. The core of the model was generated by averaging the backbone atom positions of the template structure. Reconstruction of the model side chains was done based on the weighted positions of corresponding residues in the template structure [18]. The GROMOS96 force field is used in the final step of modeling for the steepest energy minimization. The quality of the resulting homology model was verified using the protein structure verification tool WHAT_CHECK module of the WHAT IF online server. False statistics of bad nonbonded interactions within the structure model was checked by ERRAT [22]. The model was further evaluated by visualizing energetically allowed regions for backbone amino acid residues of protein by Ramachandran plot provided by RAMPAGE.

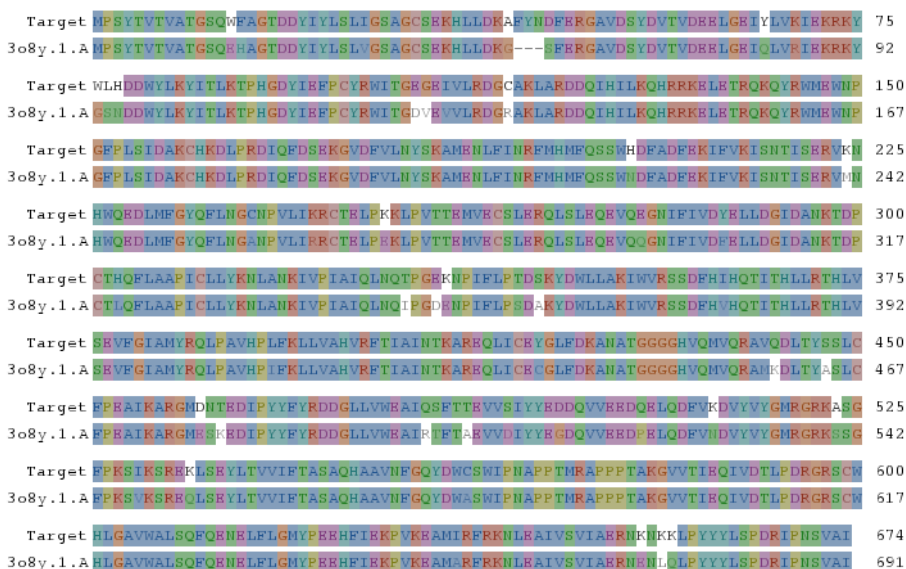


Fig. 3.4 Alignment of 3O8Y.1.A (Human 5-LOX) and target (Rat 5-LOX)

3.7.1. Analysis of the Homology Models

The 3D structure of the Rat 5-LOX was built by homology modeling based on the template protein. The superimposition of the developed model over the template 3O8Y was done in chimera to determine the accuracy of the alignment of the residues of two structures. As per the results of chimera, the RMSD value for these two structures was found to be 0.101 Å, which implies that the developed model superimposes well with the template. Model structure and superimposition of template and model are shown in Figures 3. 5A and 3. 5B, respectively.

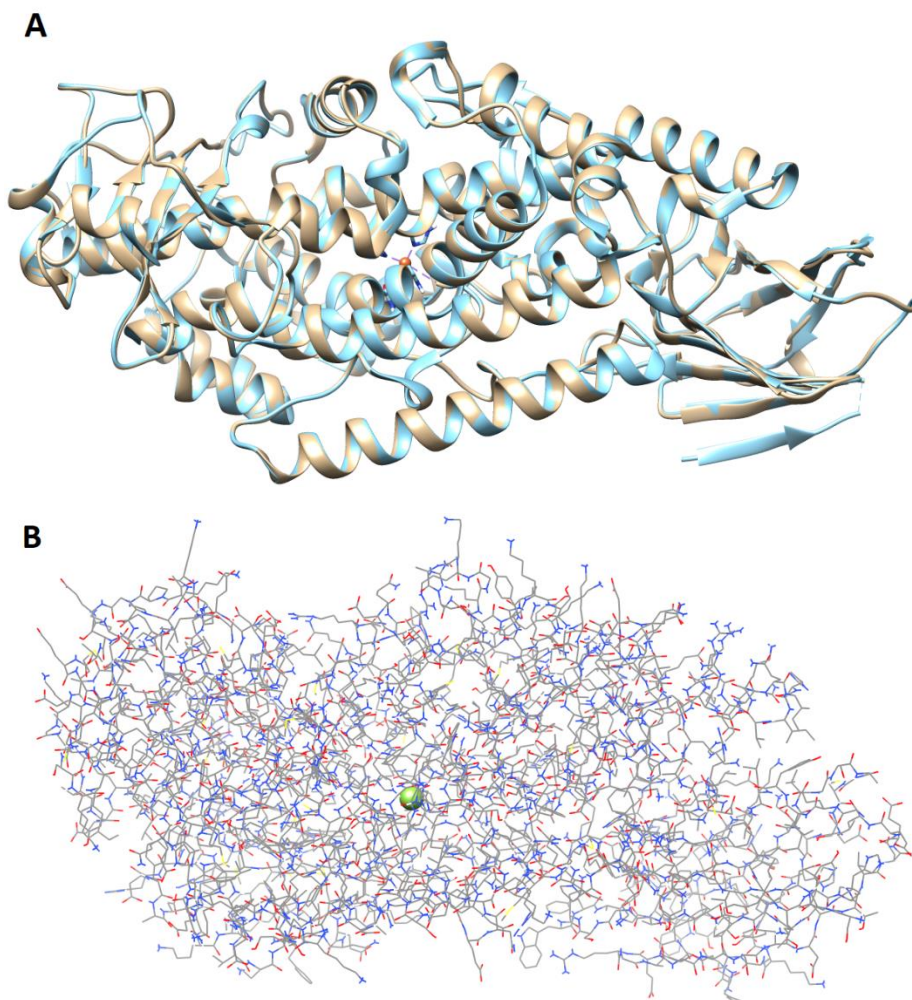


Fig. 3.5 A) The superimposition of template Human 5-LOX (3O8Y.1.A sky blue color) and Homology model of Rat 5-LOX (gray color) and B) Homology model of Rat 5-LOX.

The developed model is evaluated by the 'WHAT IF' algorithm of the WHAT-CHECK program. It checks the local abnormalities in stereochemistry and detects gross errors in protein structures. 'WHAT IF' produces a detailed report on the overall quality of the structure compared to the current, reliable structures portrayed in the form of

RMS Z-scores. Table 3. 4 presents the summary output of 'WHAT IF' criteria for the generated homology structures. RMS Z-scores for bond lengths and bond angles as determined by 'WHAT IF' 0.369 and 1.046, respectively, which is almost equal to 1.0, suggesting the high quality of the model.

Table 3.4 Summary evaluation information produced by the WHAT_CHECK package for the generated homology structure

Structure Z-scores (positive is better than average)	
1st generation packing quality	-0.617
2nd generation packing quality	-1
Ramachandran plot appearance	0.076
chi-1/chi-2 rotamer normality	0.858
Backbone conformation	0.495
RMS Z-scores (close to 1.0 is good)	
Bond lengths	0.369
Bond angles	1.046
Omega angle restraints	1.201
Side-chain planarity	1.165
Inside/Outside distribution	1.045
Improper dihedral distribution	0.974

The homology model is further evaluated by ERRAT, a protein structure verification algorithm. Protein verification was performed by comparing it against a database of trustworthy high-resolution structures premised based on the statistics of bad nonbonded interactions within the structure model. ERRAT plot shown in Figure

3. 6 shows the value of the error function vs. position of a 9-residue sliding window. From the plot, structure error at each residue in the protein can be identified. An error value exceeding 99% confidence level indicates poorly modeled regions. The developed model exhibits an overall quality factor of 93.67% and is within the accepted range, which means the excellent quality of the developed homology model. The number of residues that are in the allowed or disallowed regions of the Ramachandran plot further determines the quality of the model. Conformationally unreasonable residues generally fall in the disallowed regions of the statistical Ramachandran map. The percentage of residues in the allowed regions was expected to be more than 90% for a good model. The Ramachandran plot analysis (Figure 3.7) of the developed model shows that 656 (97.9%) residues are found to be in the favored, 14 (2.1%) in the allowed, and none were in the disallowed or outlier region. All these validation results indicate that the developed homology model is excellent for carrying out further studies.

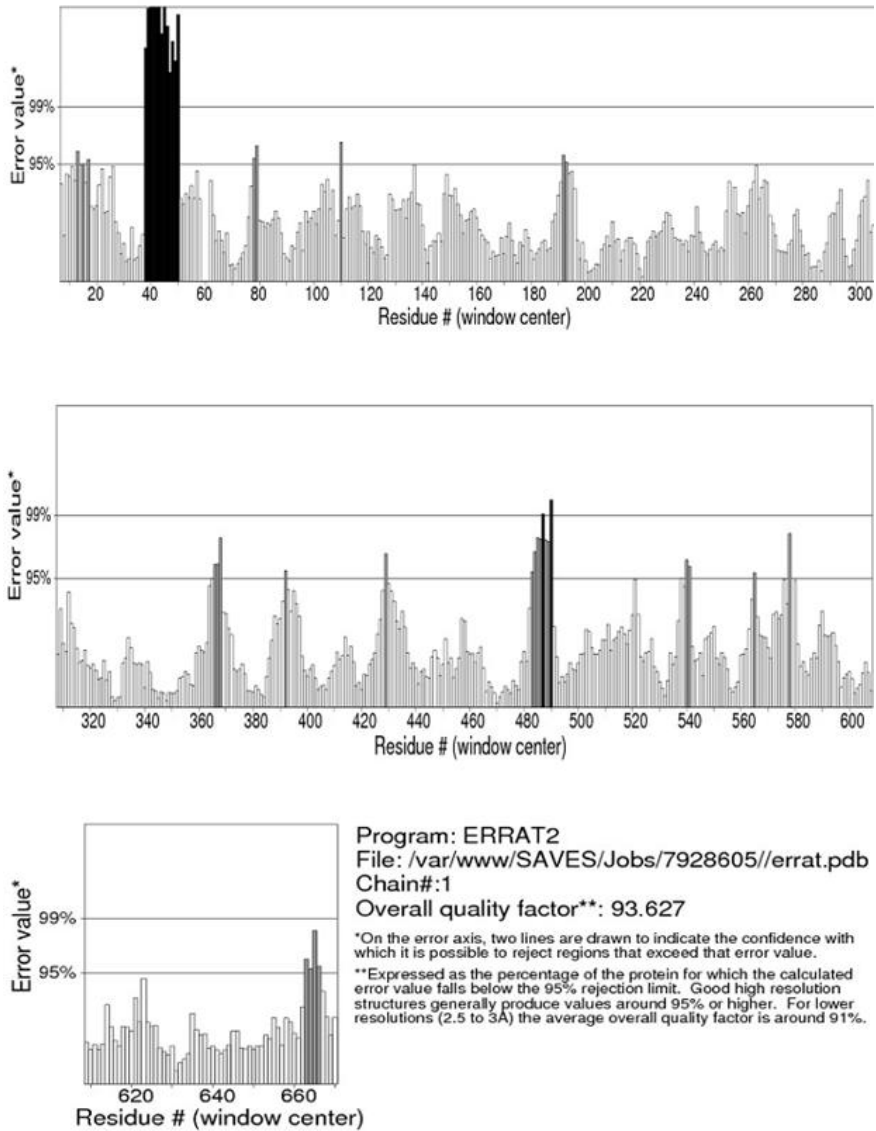


Fig. 3.6 ERRAT plot of Rat 5-LOX model

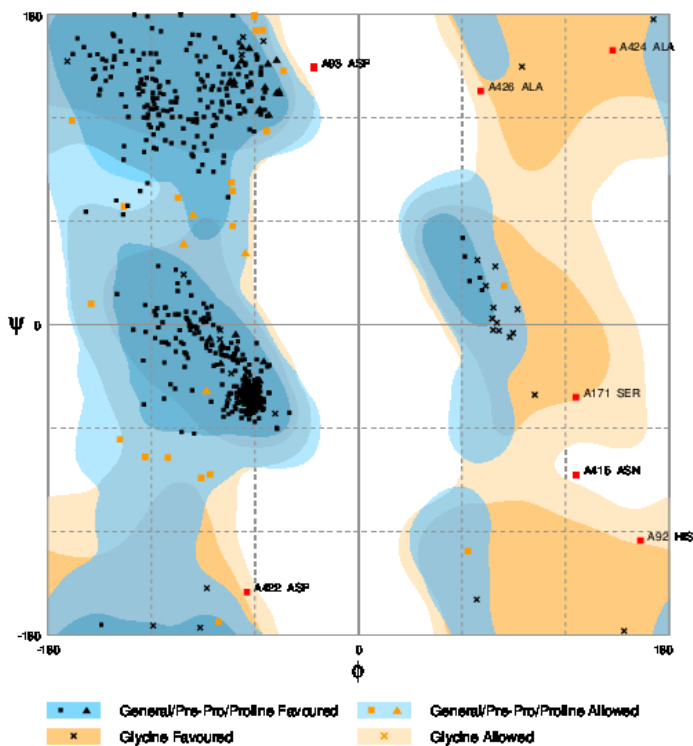


Fig. 3.7 Ramachandran plot of Rat 5-LOX model

3.7.2 Binding Site Investigation

The iron-binding site of both the Rat 5-LOX and Human 5-LOX obtained from chimera is given in Table 3. 5. The iron-binding site of both models follows a similar structural pattern. The catalytic Fe is held in place in the developed homology model by coordination with an imidazole nitrogen atom of three preserved histidine residues His 551, 368, and 373, Asn 555 carboxylic oxygen, and C-terminal Ile

674 carboxylate oxygen. Iron is in distorted octahedral position with a distance of Fe – O with Ile 674 being 2.14 Å and Fe-N with His 373, His 551 and 368 is 2.01, 2.18, and 2.23 Å, respectively.

Table 3.5 Iron binding site information obtained for LOXs (Human) and 5-LOX model

Protein structure	Coordinator	Distance (Å)	Distance RMSD	Best Geometry
Rat 5-LOX model	HIS 373.A NE2	2.01	0.000	N/A
	ILE 674.A O	2.14	0.065	trigonal bipyramid
	HIS 551.A NE2	2.18	0.208	octahedron
	HIS 368.A NE2	2.23	0.246	octahedron
	ASN 555.A OD1	3.10	0.376	octahedron
5-LOX Human crystal structure	HIS 372.A NE2	2.10	0.000	N/A
	ILE 673.A O	2.24	0.171	tetrahedral
	HIS 550.A NE2	2.21	0.101	trigonal bipyramid
	HIS 367.A NE2	2.24	0.293	octahedron
	ASN 554.A OD1	3.22	0.312	octahedron

The binding site of Rat 5-LOX homology model was in compactable with the Human 5-LOX crystal structure reported in the literature, and both are characterized by elongated cavity as the active site, which is surrounded by 5-LOX specific amino acids Tyr 181, Ala 603, Ala 606, His 600 and Try 364 wherein they contribute to ligand binding. Residues like Leu- 368, 373, 414, 607, and Ile-406 are conserved in all AA-metabolizing LOX are also retained in the developed homology model. These amino acids form a structural constellation of conserved hydrophobic side chains that can accommodate the hydrophobic part of the substrate for catalysis observed by Nathaniel C *et al.*, in Human 5-LOX [2]. Small side

chains of Ala 603 and Ala 606 also appeared in the developed model as Ala 604 and Ala 607, which is to be required for the confirmation of Tyr181. So, this homology model could be used for docking of compounds that are experimented for Rat 5-LOX protein inhibition.

3.8. Conclusion

As a part of structure-based CADD, it is necessary to understand protein structure, function, nature of ligand binding site, and druggability. So, in this study, we have performed a complete analysis of 5-LOX protein structural evaluation. Among all crystal structures available for 5-LOX, the stable 3D structure of 5-LOX (3O8Y) with a resolution of 2.389 Å was finalized as the target structure for the entire study. Ligand binding sites obtained from SiteMap and POCASA are compatible with each other, and the Site A (in SiteMap) /cavity 1 (in POCASA) is the ligand-binding site of the 5-LOX protein with a site score of 1.197 having a volume of 144.403. The docking of 11 known 5-LOX and substate AA was performed using established ligand binding sites. The outcome also shows that most of the interactions are, in fact, hydrophobic and that some pi-pi stacking interactions are also there. The H-bond interactions with amino acids in the polar section of the binding pocket diminishes the contact with the hydrophobic residues. The molecular docking results show that zileuton, a known 5-LOX inhibitor, satisfied all the structural requirements for receptor binding.

Moreover, a 3D model of the rat 5-LOX receptor was also constructed, and refined by energy minimization. The validation of the

modeled protein was done using the different protein structure verification tool. RMS Z-scores for bond lengths and bond angles are 0.369 and 1.046, respectively, which is almost equal to 1.0, suggesting high model quality. ERRAT plot exhibits an overall quality factor of 93.67% and is within the accepted range, which indicates the excellent quality of the developed homology model. The Ramachandran plot analysis of the developed model shows that 656 (97.9%) residues are found to be in the favored region, and 14 (2.1%) residues are in an allowed region with none were in the disallowed region. The metal binding site and the ligand/substrate-binding site of the developed homology model were also identified. In conclusion, in this study, we have prepared Human 5-LOX protein and Rat 5-LOX model that can be used under appropriate conditions such as SBVS or molecular docking.

References

- [1] T. Kenakin, New Concepts in Drug Discovery: Collateral Efficacy and Permissive Antagonism, *Nat. Rev. Drug Discov.* 4 (2005) 919–927. doi:10.1038/nrd1875.
- [2] N.C. Gilbert, S.G. Bartlett, M.T. Waight, D.B. Neau, W.E. Boeglin, A.R. Brash, M.E. Newcomer, The structure of human 5-lipoxygenase, *Science*. 331 (2011) 217–219. doi:10.1126/science.1197203.
- [3] N.C. Gilbert, Z. Rui, D.B. Neau, M.T. Waight, S.G. Bartlett, W.E. Boeglin, A.R. Brash, M.E. Newcomer, Conversion of human 5-lipoxygenase to a 15-lipoxygenase by a point mutation to mimic phosphorylation at Serine-663, *FASEB J.* 26 (2012) 3222–3229. doi:10.1096/fj.12-205286.
- [4] Schrödinger Release 2018-3: Glide, Schrödinger, LLC, New York, NY, (2018).
- [5] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J.Y. Xiang, L. Wang, D. Lupyan, M.K. Dahlgren, J.L. Knight, J.W. Kaus, D.S. Cerutti, G. Krilov, W.L. Jorgensen, R. Abel, R.A. Friesner, OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins, *J. Chem. Theory Comput.* 12 (2016) 281–296. doi:10.1021/acs.jctc.5b00864.
- [6] P. Aparoy, R.N. Reddy, L. Guruprasad, M.R. Reddy, P. Reddanna, Homology modeling of 5-lipoxygenase and hints for better inhibitor design, *J. Comput. Aided. Mol. Des.* 22 (2008) 611–619. doi:10.1007/s10822-008-9180-0.
- [7] S. Mitra, Insights into 5-Lipoxygenase Active Site and Catalysis, 2015.
- [8] J. Yu, Y. Zhou, I. Tanaka, M. Yao, Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere, *Bioinformatics*. 26 (2009) 46–52. doi:10.1093/bioinformatics/btp599.
- [9] A.L. Hopkins, C.R. Groom, The druggable genome, *Nat. Rev. Drug Discov.* 1 (2002) 727–730. doi:10.1038/nrd892.
- [10] T.A. Halgren, Identifying and Characterizing Binding Sites and Assessing Druggability, *J. Chem. Inf. Model.* 49 (2009) 377–389. doi:10.1021/ci800324m.

- [11] T. Halgren, New Method for Fast and Accurate Binding-site Identification and Analysis, *Chem. Biol. Drug Des.* 69 (2007) 146–148. doi:doi:10.1111/j.1747-0285.2007.00483.x.
- [12] Schrödinger Release 2018-3: SiteMap, Schrödinger, LLC, New York, NY, (2018).
- [13] M.A. Ghattas, N. Raslan, A. Sadeq, M. Al Sorkhy, N. Atatreh, Druggability analysis and classification of protein tyrosine phosphatase active sites, *Drug Des. Devel. Ther.* 10 (2016) 3197–3209. doi:10.2147/DDDT.S111443.
- [14] L.R. Vidler, N. Brown, S. Knapp, S. Hoelder, Druggability analysis and structural classification of bromodomain acetyl-lysine binding sites, *J. Med. Chem.* 55 (2012) 7346–7359. doi:10.1021/jm300346w.
- [15] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.* 47 (2004) 1739–1749. doi:10.1021/jm0306430.
- [16] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening, *J. Med. Chem.* 47 (2004) 1750–1759. doi:10.1021/jm030644s.
- [17] R.A. Friesner, R.B. Murphy, M.P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, P.C. Sanschagrin, D.T. Mainz, Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes, *J. Med. Chem.* 49 (2006) 6177–6196. doi:10.1021/jm051256o.
- [18] T. Schwede, J. Kopp, N. Guex, M.C. Peitsch, SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res.* 31 (2003) 3381–3385. doi:10.1093/nar/gkg520.
- [19] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting , position-specific gap penalties and weight matrix choice, 22 (1994) 4673–4680.

- [20] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402. doi:10.1093/nar/25.17.3389.
- [21] A.H. & J.S. Michael Remmert, Andreas Biegert, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nat. Methods.* 9 (2012) 173–175.
- [22] C. Colovos, T.O. Yeates, Verification of protein structures: Patterns of nonbonded atomic interactions, *Protein Sci.* 2 (1993) 1511–1519. doi:10.1002/pro.5560020916.

4

CHEMICAL SPACE CHARACTERIZATION AND SAR ANALYSIS

4.1. Introduction

Not only the target 5-LOX but also the thousands of compounds that are experimented for 5-LOX inhibitory potency could provide a great deal of information regarding the design of novel inhibitors. This large amount of structure and activity data that are obtained as a result of general screening, high-throughput screening (HTS) and combinatorial chemistry is collected and deposited not only in research papers but also in public domain databases like ChEMBL [1], PubChem [2], BindingDB [3], etc. These compound collections contribute to the development of 'Biologically Relevant Chemical Space (BRCS) of 5-LOX. Chemoinformatic characterization of these chemical spaces is a significant step towards virtual or experimental testing to recognize novel biologically active molecules [4–6]. Besides, data mining of these chemical spaces could help the development of models that can be beneficial to predict the activity of a new compound [7].

So, visualization and characterization of 'biologically relevant chemical space of 5-LOX and their compound distributions are very important to medicinal chemists as it can assist in understanding molecular features that are pharmaceutically important [4,8]. Different computational data mining and visualization methods and machine learning algorithm that originally developed for computer science is now largely used for the understanding the chemical space of biological interest [6,9–12]. Standard statistical and classification techniques can also be used to organize datasets and evaluate the chemical neighborhood of potent hit [13–16]. With the help of these methods, chemoinformatic characterization of chemical space of inhibitors of various targets or disease has been reported for the last few decades by many eminent scientists, especially the team lead by Jose L. Medina-Franco [5,17–22].

In recent years, several SAR, QSAR, and other predictive models of 5-LOX and its activating protein 'FLAP' inhibitors have been developed [23–27]. However, there are no systematic chemoinformatic studies on the structural diversity, activity landscape, and chemical space distribution analysis of 5-LOX and FLAP inhibitor's chemical space. Therefore, exploring and navigating the biologically relevant 5-LOX and FLAP inhibitor's chemical space is of the utmost importance. It can provide an opportunity for the analysis and enumeration of the compound in the chemical space of 5-LOX and FLAP. It also helps their library design, Structure-Activity Relationships (SAR), landscape studies, and virtual screening.

So, in this work, we have tried to locate the chemical space of 5-LOX and FLAP currently stored in a major public database ‘ChEMBL’ and characterized these spaces using multiple criteria including physicochemical properties (PCP), structural fingerprints, and molecular scaffolds. We have also decided to extract SAR from these large datasets and presented them intuitively by Land-Scape modeling based on systematic pair-wise comparisons of similarity between structure and activity. Also, this study made an effort to identify activity cliffs in these landscapes. In all these studies, we have decided to compare the chemical space of 5-LOX, and FLAP inhibitors to the dataset of ‘FDA approved drugs from drug bank’ because the FDA approved drug database is a commonly used reference compound database in drug discovery campaigns. We have also tried to understand and validate the virtual LOX library created and offered by the Enamine database by comparing it to the 5-LOX dataset.

4.2. Compound Databases and Bioactivity Representations

The study of the complexity and diversity of large molecular databases allows for the creation of high-quality leads and thus increases the performance of real and virtual drug design. Here we too tried to understand the complexity and diversity of 5-LOX and FLAP chemical space. ChEMBL dataset of bioactive molecules with drug-like properties that are active against 5-LOX and FLAP has been selected for this purpose. The chemical structures and activity data of 5-LOX and FLAP were downloaded from the ChEMBL_25 database to DataWarrior software [28] by specifying the search criteria

'biological targets'. DataWarrior allows querying the ChEMBL database. Each of the crude data containing 3498 and 4781 compounds was tested respectively for 5-LOX and FLAP inhibitory potency.

Bioactivities are often reported in ChEMBL as units of K_i , K_d , IC_{50} , and EC_{50} , etc., along with the different conditions of the assay such as cell line, tissue, or organism used. Figures 4.1A and 4.1B show similarity maps of non-curated 5-LOX's and FLAP's inhibitors clustered using Skelsphere descriptors based on structural similarity. This view depicted the chemical space of all inhibitors. Each marker of these maps represents a compound in the dataset. Marker's color, shape, and size respectively show the type of bioactivity data, organism, and cluster number. These maps indicate how diverse each dataset is. Also, it can be easily understood from the map that most of the bioactivity test of 5-LOX inhibitors are done in both Human and Rat 5-LOX protein. However, in this work, we have focused only on Human 5-LOX protein inhibitors. In the case of FLAP data, most of the *in-vitro* analysis is carried out using Human protein, so we have selected Human FLAP inhibitors for further study.

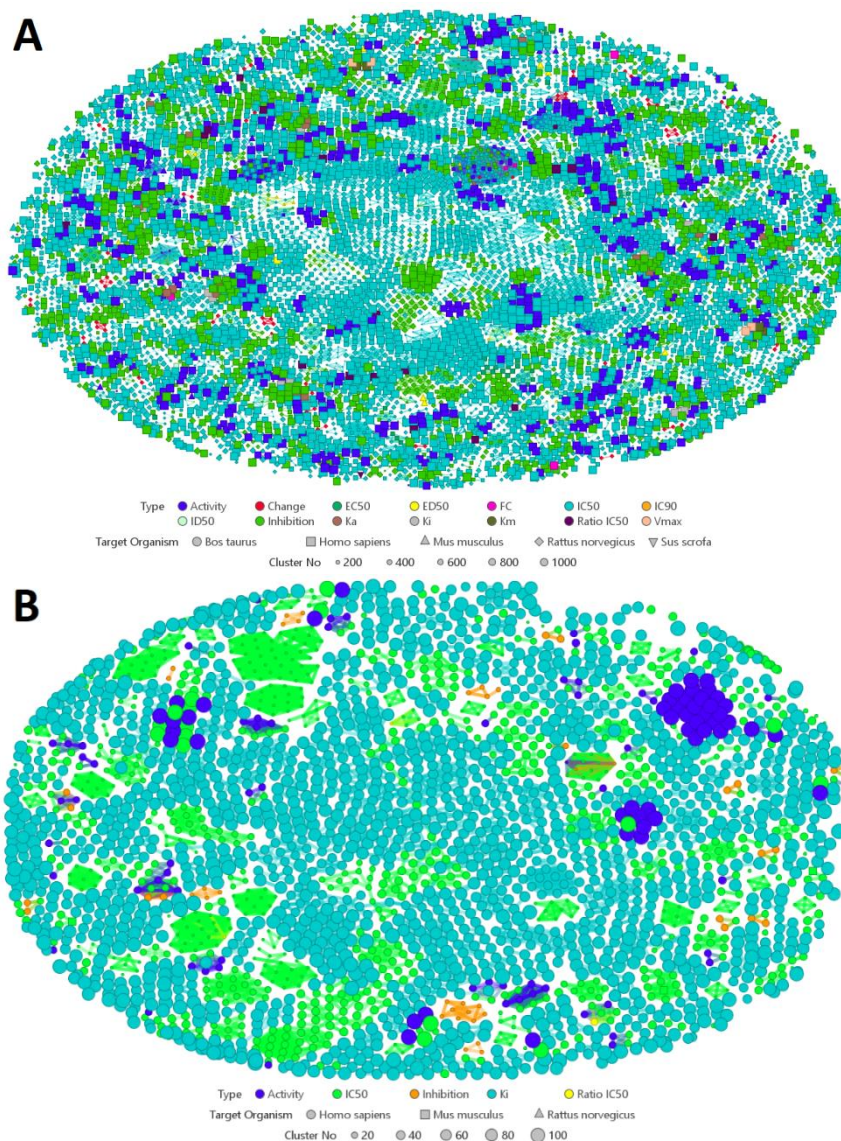


Fig. 4.1 Similarity map of 5-LOX and FLAP noncurated inhibitors space.

Although the ChEMBL database recorded the bioactivity data of the molecule in different units, it is mostly in IC_{50} format. Figure 4.2 shows the Pi-diagram of the classification of biological activity

based on their unit. 64% of activity data of 5-LOX were expressed in IC_{50} , while in the case of FLAP, 61% of activity data is expressed in K_i . In this study, therefore, only compounds with reported IC_{50} values obtained in the enzymatic inhibitory assay were included in the 5-LOX dataset, whereas compounds with reported K_i values obtained in the enzymatic inhibitory assay were included in the FLAP dataset. Then data curation was done by removing duplicate compounds and compounds which have no specified activity details etc. For compounds with the same chemical structure with contradictory bioactivity reports, the lowest activity value was kept. After data curation, the datasets had 1373 unique molecules for 5-LOX and 1379 for FLAP.

Also, in each study conducted in this Chapter, molecules from the 5-LOX and FLAP databases were compared to FDA approved drugs from the drug database 'DrugBank' version 5.1.5 and are represented by 'App_drug.' Furthermore, the comparison of the similarity and diversity of the virtual library of the lipoxygenase inhibitors (LOX_lib) containing 1387 compounds provided by the Enamine database was carried out by comparing it with the 5-LOX dataset. The vendors' Enamine used two different methods to build the Target focused virtual LOX library. The first one is docking-based in silico screening, and the second one is a similarity search based on 2D linear fingerprints and 3D pharmacophore features of the reported lipoxygenase inhibitors. The former method recognized compounds of novel scaffolds, and later method provided molecules with new side chains. So, there are four compound databases (datasets/chemical

space/inhibitor space) is used in the study and are represented by 5-LOX (target inhibitors set), FLAP (target inhibitors set), LOX_lib (virtual LOX library set) and App_drug (reference datasets).

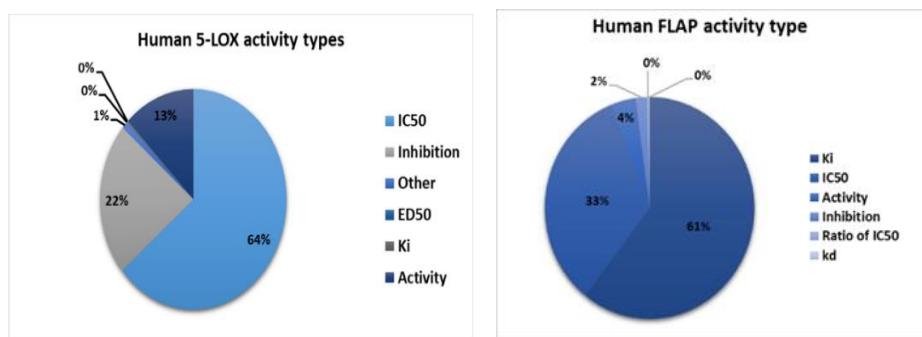


Fig. 4.2 Activity chart of human 5-LOX and FLAP inhibitors

4.3. Physicochemical Properties of 5-LOX Chemical Space

The appropriate physicochemical and molecular properties of drug candidates are supposed to increase their market chances and to cut the attrition rate [29]. Tracking the property space for large dataset enables to identify the pharmacokinetic and pharmacodynamic nature of the chemical space covered by ligands in the set. So, in this section, a thorough physicochemical property (PCP) analysis of 5-LOX and FLAP inhibitor's chemical space has been carried out and compared with App_drug and LOX_lib database. For this, we have used basic molecular property descriptors of pharmaceutical relevance such as octanol/water partition coefficient (Clog P), molecular weight (MW), number of rotatable bonds (RB), number of hydrogen bond donors (HBD) and number of hydrogen bond acceptors (HBA). These are the five PCP associated with the oral bioavailability of drugs suggested by

Lipinski et al. Additional to this, the topological polar surface area (TPSA) descriptor is also used. Together, these six descriptors can provide the important properties of size, flexibility, and molecular polarity of the chemical space of 5-LOX and FLAP.

The property distribution was evaluated with SPSS box plots. Figure 4.3 shows the box and whisker plots spread out of the basic PCP of all four datasets. The primary portion of the chart is a box, which indicates the interquartile range, i.e., it shows where the central portion of the data is. The first quartile, Q1 (25% mark), and the third quartile, Q3 (75% mark), are situated at lower and upper ends of the box, respectively. The 'median' represents a horizontal bar in the middle of the box. The smallest and largest number in the dataset, respectively represented by a horizontal line situated far below and far above the graph. The open circles on the graph denote the outliers, and the asterisks or stars are represented as extreme outliers. Summary statistics of the distributions are presented in Table 4.1.

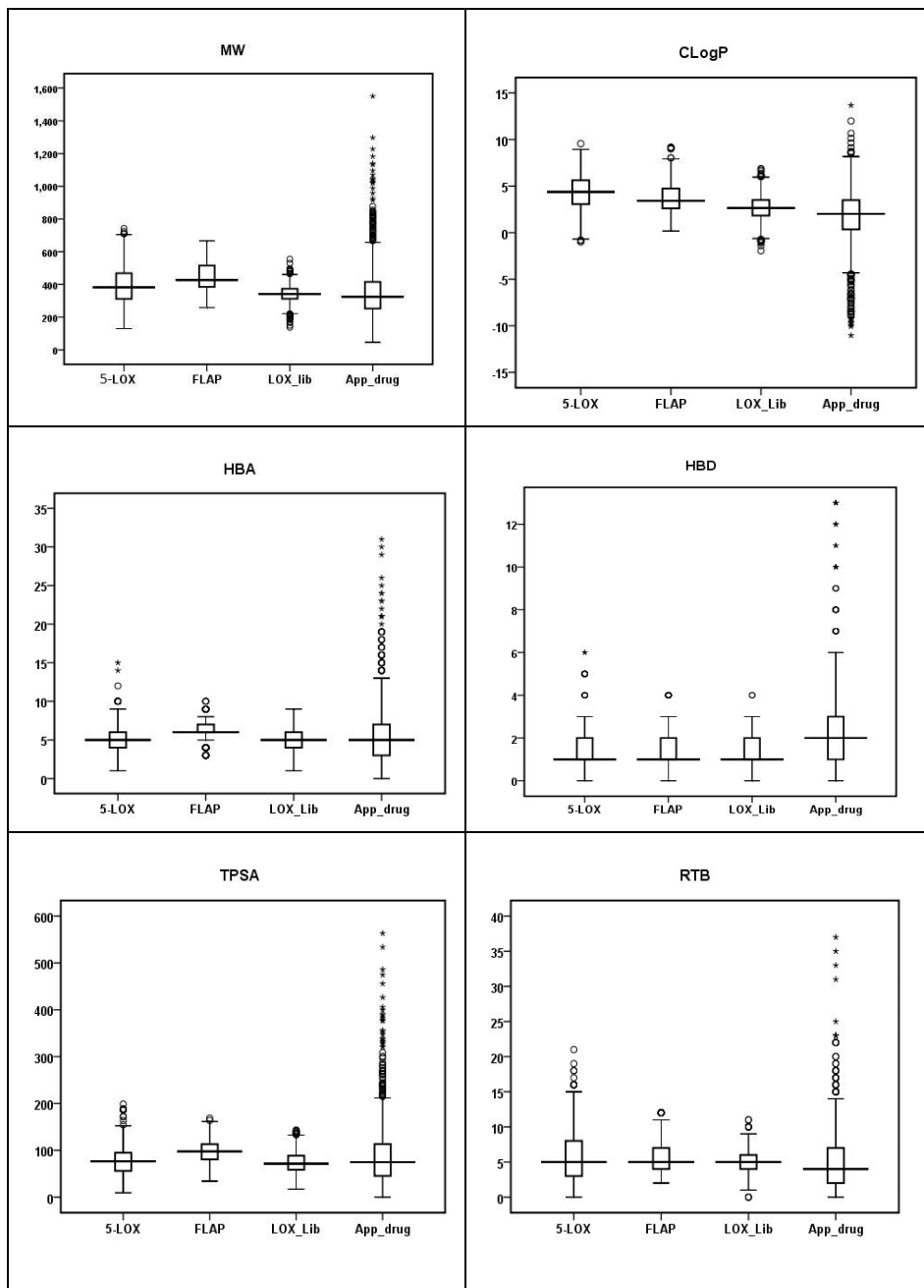


Fig. 4.3 Box and whisker plots of four databases with six physicochemical characteristics.

The box plot result shows that among four datasets, the approved drug dataset shows larger ranges in all PCP, which indicates wider distribution, i.e., more scattered data. Also, the distribution of PCP values of compounds in the LOX virtual library falls within the range of distribution of PCP values of inhibitors of 5-LOX, i.e., there is an overlap in spreads. This result indicates that there is no significant difference between these two sets, i.e., 75% of the molecules having similar PCP values. Besides, all three datasets have almost similar distributions of HBD, which is less than that of approved drug and have HBA and TPSA distributions that are comparable to the approved drug. These results indicate that the compounds screened as inhibitors of 5-LOX and FLAP are, in general, less or comparable polar than approved drugs. The MW box plot of the LOX library is comparatively short, indicates compounds in this virtual library have a short-range of MW, which is between 138.16 and 553.84. Short boxes mean their data points consistently hover around the center values. The distribution of RTB of all the three datasets was similar to the approved drug dataset. This observation indicates that the flexibility of compounds in these databases is similar to the drugs in the drug database.

Table 4.1 Summary statistics of the distributions of each dataset

Statistic	MW				CLogP			
	5-LOX	FLAP	LOX_Lib	App_drug	5-LOX	FLAP	LOX_Lib	App_drug
Mean	393.76	446.38	341.71	352.80	4.26	3.71	2.66	1.68
Median	382.43	426.52	341.34	324.42	4.38	3.42	2.65	2.01
Variance	14000.00	7191.00	2723.00	27060.00	3.84	2.32	1.65	8.27
Std.Dev	118.33	84.80	52.18	164.51	1.96	1.52	1.28	2.88
Minimum	130.15	258.30	138.16	46.07	-0.99	0.17	-1.94	-11.04
Maximum	742.28	666.70	553.84	1550.16	9.56	9.17	6.89	13.69
Range	612.14	408.40	415.67	1504.09	10.55	8.99	8.83	24.73
Int. Range	157.64	131.47	61.12	163.58	2.56	2.14	1.66	3.16
Skewness	0.41	0.49	-0.11	1.83	-0.27	0.56	0.03	-0.91
Kurtosis	-0.19	-0.55	0.68	6.17	-0.26	0.00	0.28	2.51
Q1	311.36	384.84	312.34	252.27	3.07	2.62	1.84	0.35
Q3	468.59	515.61	373.46	415.57	5.62	4.75	3.50	3.50
Statistic	HBA				HBD			
	5-LOX	FLAP	LOX_Lib	App_drug	5-LOX	FLAP	LOX_Lib	App_drug
Mean	4.87	6.39	5.34	5.76	1.18	1.50	1.46	2.14
Median	5.00	6.00	5.00	5.00	1.00	1.00	1.00	2.00
Variance	2.81	1.56	2.31	15.42	0.90	0.48	0.60	4.71
Std. Dev	1.68	1.25	1.52	3.93	0.95	0.69	0.77	2.17
Minimum	1.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00
Maximum	15.00	10.00	9.00	31.00	6.00	4.00	4.00	16.00
Range	14.00	7.00	8.00	31.00	6.00	4.00	4.00	16.00
Int. Range	2.00	1.00	2.00	4.00	1.00	1.00	1.00	2.00
Skewness	0.66	-0.25	-0.21	1.96	1.16	1.16	0.25	2.63
Kurtosis	1.73	0.25	-0.20	5.95	3.00	0.94	-0.24	10.55
Q1	4.00	6.00	4.00	3.00	1.00	1.00	1.00	1.00
Q2	6.00	7.00	6.00	7.00	2.00	2.00	2.00	3.00
Statistic	TPSA				RTB			
	5-LOX	FLAP	LOX_Lib	App_drug	5-LOX	FLAP	LOX_Lib	App_drug
Mean	77.66	97.97	73.88	89.70	6.11	5.69	4.67	5.20
Median	76.55	97.56	71.53	74.73	5.00	5.00	5.00	4.00
Variance	850.83	494.03	447.44	4539.00	12.28	5.01	2.84	16.99

Chemical space Characterization and SAR Analysis

Std. Dev	29.17	22.23	21.15	67.37	3.50	2.24	1.69	4.12
Minimum	9.23	34.15	17.07	0.00	0.00	2.00	0.00	0.00
Maximum	198.48	168.32	143.24	563.45	21.00	12.00	11.00	37.00
Range	189.25	134.17	126.17	563.45	21.00	10.00	11.00	37.00
Int. Range	38.97	32.53	30.02	68.00	5.00	3.00	2.00	5.00
Skewness	0.66	0.15	0.37	2.20	0.76	0.68	0.29	2.06
Kurtosis	0.84	-0.29	-0.07	7.89	0.16	-0.49	0.20	8.51
Q1	55.84	80.48	58.37	45.34	3.00	4.00	4.00	2.00
Q3	94.80	112.91	88.39	113.24	8.00	7.00	6.00	7.00

None of the distributions displayed a normal distribution as calculated using the SPSS-based Shapiro-Wilk method. Also, two quantitative statistical tests obtained from SPSS, such as skewness and excess kurtosis, can be used to evaluate how the normality of the distribution of each dataset has changed. Skewness measures the asymmetry of the probability distribution of a random variable about its mean. Kurtosis is a measure of the distribution's peakness. A normal distribution has zero skewness and zero excess kurtosis, so if the distribution is close to zero, it is likely to be close to normal. Negative values for the skewness indicate data that are skewed left, and positive values indicate data that are skewed right. Datasets with negative kurtosis tend to have light tails or lack of outliers, while datasets with positive kurtosis indicate a "heavy-tailed" distribution with more outliers. West et al. (1996) suggested as an absolute skew value > 2.1 and absolute kurtosis (proper) value > 7 are the measures of significant deviation from normality [30]. The 'excess' kurtosis obtained by subtracting three from the kurtosis (proper). Most of the dataset is moderately skewed. That is, the skewness of these sets is

between -1 and $-\frac{1}{2}$ or between $+\frac{1}{2}$ and $+1$. Most of the cases, kurtosis is <3 indicates the distribution is shorter, and tails are thinner than the normal distribution or lack of outliers. We may conclude that all the datasets seem not to follow the presumption of normality concerning Table 4.1 and Figure 4.3. The approved drug datasets show a greater deviation from normality measured in terms of skewness and kurtosis, which is even greater than the range proposed by West et al.

4.4. Visual Representation of the Property Space

This section aims to obtain an initial overall assessment of the chemical space coverage and distribution of compounds tested for the 5-LOX and FLAP inhibitory activity in terms of selected PCP. The Principal Component Analysis (PCA) [31] is the best technique to convert many highly correlated variables into a smaller number of less correlated variables that still contains most of the information in the large set. Besides, it can be used to visualize various patterns of data hidden in datasets. In order to verify the chemical space of 5-LOX and FLAP inhibitors, this study explores the possibility of PCA using six PCP descriptors mentioned earlier.

The results show that the first five principal components (PC) have Eigenvalues greater than 1. By following the Kaiser criterion [32], we could have used all the PC with Eigenvalues that are greater than 1. However, in this case, the first three-component explains the maximum percentage of variation in all the data, so this study used the first three PC only. The first two PCs (PC1 and PC2 or PCa and PCb) retrieved 80.32, 84.38, 70.57, and 86.03% of the variance of 5-LOX

inhibitors, FLAP inhibitors, LOX library and Approved drug property space respectively. The first three PCs captured 90.51, 93.38, 83.79, and 93.68% of the variance of 5-LOX inhibitors, FLAP inhibitors, LOX library, and Approved drug property space, respectively. The magnitude and direction of the coefficients of the original variables help to interpret each PC. When the absolute value of the coefficient is large, the associated variable becomes even more important in the component calculation. The loadings for the first three PCs of the property space of the three datasets are summarized in Table 4.2. The loading plot (Figure 4.4) visually shows the results for the first two components of each inhibitors property space.

Table 4.2 Loadings for the first three PCs of the property space of the 5-LOX set, FLAP set, LOX_lib, and App_drug

Databases	Variable Name	MW	RTB	PSA	HBD	HBA	CLogP
5-LOX	PC1	0.501	0.439	0.454	0.249	0.476	0.248
	PC2	-0.255	-0.324	0.355	0.472	0.307	-0.623
	PC3	0.128	-0.069	0.189	-0.826	0.366	-0.355
FLAP	PC1	0.350	0.387	-0.413	-0.409	-0.313	0.540
	PC2	-0.529	-0.463	-0.444	-0.145	-0.530	-0.083
	PC3	0.068	0.115	-0.045	0.865	-0.409	0.256
LOX_lib	PC1	0.383	0.334	0.522	0.339	0.539	-0.254
	PC2	-0.559	-0.443	0.083	0.295	0.115	-0.619
	PC3	0.060	0.131	-0.188	-0.778	0.320	-0.485
App_drug	PC1	0.390	0.320	0.497	0.443	0.496	-0.237
	PC2	-0.470	-0.462	0.115	0.254	-0.008	-0.698
	PC3	0.404	-0.827	0.155	0.032	0.179	0.309

From Table 4.2, it can be seen that MW and HBA had large positive contributions to the first PC, while CLogP and HBD had the highest negative contribution to the second and third PC, respectively while analyzing the PC loading of the 5-LOX database. In other words, the second component has large negative associations with CLogP, so this component primarily measures the molecule's octanol/water partition coefficient. In contrast, the third component has large negative associations with HBD, so this component primarily measures the molecule's hydrogen bond donating capacity. The result of FLAP's PC loadings shows that, while both MW and HBA had a large negative contribution to the second PC, CLogP and HBD had a large positive contribution to the first PC and third PC, respectively. PCA results of LOX virtual library shows property contributions of PC as similar to that of 5-LOX. In approved drug datasets, unlike the other two (5-LOX and FLAP), PSA and HBA contribute mostly to the first PC, and like the other two, CLogP shows large negative contributions to the second PC. The third component had large negative associations with RTB indicates this component primarily measures the number of the molecules of rotating bonds.

In short, except for the approved drug dataset, the HBD of all other datasets either positively or negatively contributes to the third PC with a magnitude greater than 0.7. Also, except for the FLAP set, the ClogP of all other datasets contributes negatively to the second PC with a magnitude greater than 0.6, while HBA contributes positively to the first PC with a magnitude greater than 0.4. Note that all of these properties are associated with the polarity of the compound.

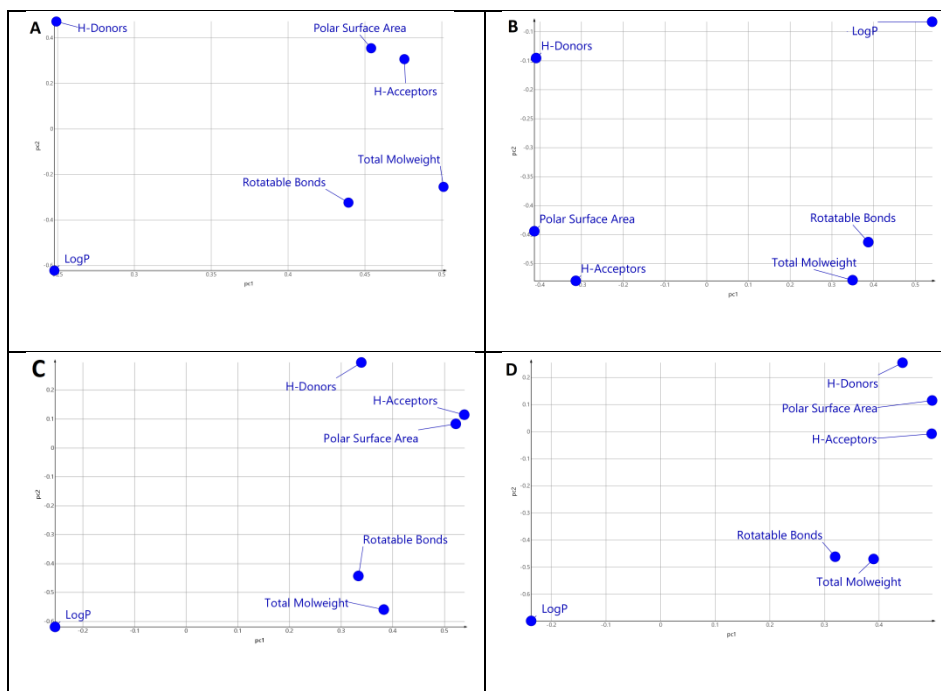


Fig. 4.4 Loading plot showing variables that have the largest effect on the first two-component of PCA of 5-LOX (A), FLAP (B), App_drugs (C), and LOX_library (D) property space.

Figures 4.5 and 4.6 shows PCA 2D and 3D plots of the distribution of the six PCP descriptors calculated for the two target inhibitors set (5-LOX and FLAP), one LOX virtual library (LOX library) and one reference datasets (Approved drug). 2D visual representation of the property space shows that the approved drug (green) covers the vast area of the property space as expected and is also the database with the largest diversity in PCP. In contrast, the LOX library (violet) encompasses more limited space, which is occupied within the property space of all other three databases. Most of the 5-LOX inhibitors space is within the drug database property space while some area is not. It indicates that 5-LOX inhibitors cover

most of the medicinal property space. However, some compounds broaden traditional medicinal space.

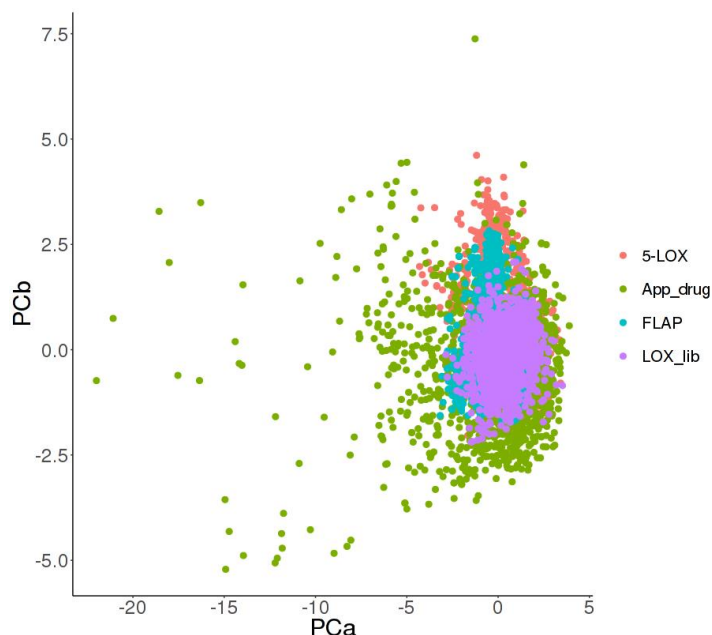


Fig. 4.5 2D visual representations of the chemical space of 5-LOX, FLAP, LOX_lib, and App_drug in the form of a PCA plot generated using six PCP.

The 3D PCA plot of all datasets was constructed (Figure 4.6), and HBD was used to mark the color of the points in all datasets (except the approved drug database) due to its contribution to the third component. For the PCA plot of the approved drug dataset, RTB is used to mark the color of the points. The distinct areas of equal color are the evidence for the separation of chemical space that a PCA can achieve, although the dataset consists of somewhat different structures.

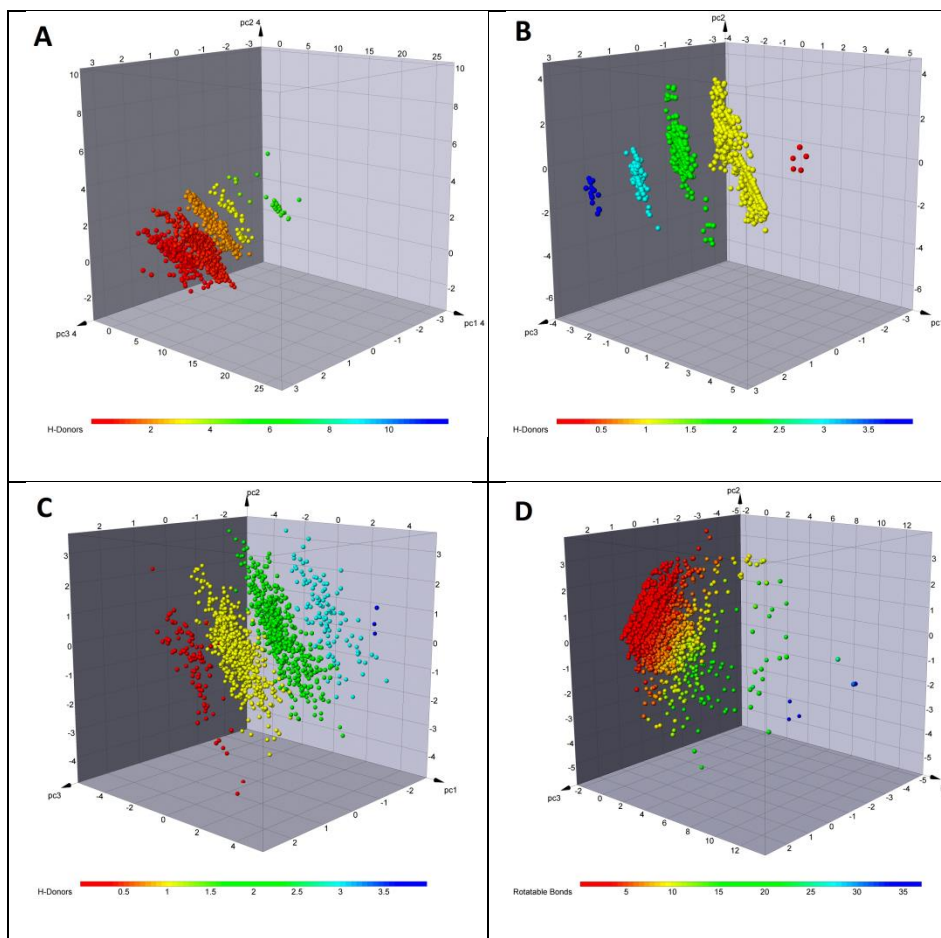


Fig. 4.6 3D visual representations of the chemical space 5-LOX (A), FLAP (B), LOX_lib (C), and App_drug (D) in the form of PCA plot generated using six PCP.

4.5. Diversity Analysis

In drug discovery, it is important to assess the structural diversity of compound databases in order to explore novel regions within the biologically relevant chemical space or to establish a balance between diversity and novelty. It is, therefore, important to carry out a comprehensive cheminformatic analysis of the diversity of

5-LOX inhibitor space. Since molecular diversity depends on molecular representation, the diversity of the 5-LOX, FLAP, Approved drugs, and LOX library is evaluated by employing molecular scaffolds, structural fingerprints, and physicochemical properties.

4. 5. 1. Diversity based on PCP

Six PCP descriptors were used for the profiling of the diversity of the datasets by several researchers [33–35]. Based on systematic pairwise comparisons of compound distances, the molecular diversity of datasets can be explained. This study computed pairwise inter and intra database chemical property diversity using six PCP descriptors with Euclidean distance. The principle behind this is that the two objects are said to be dissimilar when the distance between them is bigger, and two objects are said to be similar or closer when the distance between them is smaller [36]. The Euclidean distance is the most common distance measure, and it is the square root of the sum of all squared distances between corresponding data points. A distance matrix is generated with the mean inter and intra database Euclidean distances values and is shown in Figure 4.7. The matrix is color-coded from grey (low values) to red (high values). Largest inter-and intra-dataset Euclidean distances are marked in dark red, and the shortest Euclidean distances are marked in white.

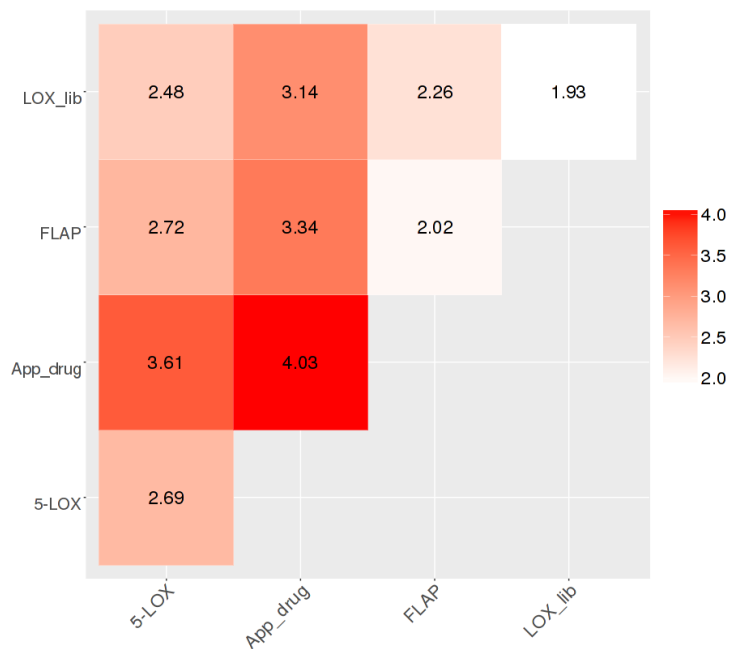


Fig. 4.7 Heat map showing Euclidean distance of datasets constructed based on six pharmaceutically relevant PCP descriptors

According to these values, with the obvious exception of approved drugs, the 5-LOX inhibitor set shows the largest intra-dataset distance as compared to other datasets. This result suggests that this dataset has a larger chemical diversity. The approved drug shows significant inter dataset distance from all the databases. Interestingly, the 5-LOX dataset shows the lowest inter dataset distance with LOX library, indicating the database property similarity of 5-LOX with LOX library.

4. 5. 2. Fingerprint Diversity

This study has thus far analyzed the distribution of six molecular property descriptors of the pharmaceutical relevance of each

dataset in order to understand, visualize, and compare the property space. However, for the more accurate description of molecular structures and their similarity, we normally use Molecular Fingerprints. A comparison of fingerprints allows the resemblance between two molecules. It is achieved by converting these molecules into a sequence of bits, which can then easily be compared between molecules. Three binary molecular fingerprints, such as extended connectivity fingerprints ((ECFP), 166-bit molecular access system (MACCS) keys, and 881-bit PubChem, were used to quantify the structural diversity (including side Chains) of datasets. ECFP is a circular topology fingerprint with a variable diameter distance. MACCS is a dictionary-based representation that matches pre-defined fragments from a list with the structure of the molecule. The structural similarity was computed using the Tanimoto similarity coefficient and generated a similarity matrix. For each similarity matrix, random samples of 5000 similarity values off the diagonal were extracted to calculate statistics like mean, median, interquartile distances, and standard deviation and analyzed with the cumulative distribution function (CDF). Figure 4.8 illustrates the cumulative distribution function of the pairwise intraset similarity values calculated with ECFP4/Tanimoto and MACCS keys/Tanimoto, and corresponding summary statistics of the similarity distributions are depicted in Table 4.3.

Table 4.3 The statistical values of the similarity of the Tanimoto coefficient with ECFP4 and MACCS fingerprints.

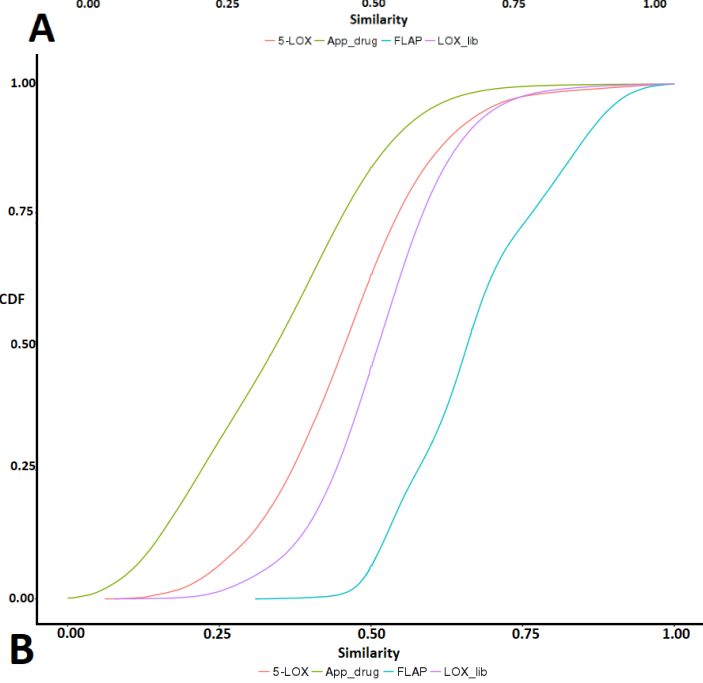
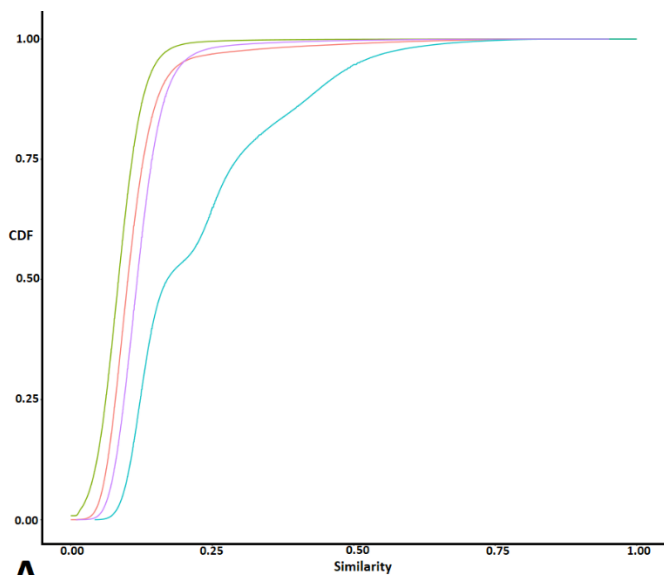
FP	Dataset	Min	1st Qu	Median	Mean	3rdQu	Max	Std.Dev
ECFP4	5-LOX	0.000	0.079	0.101	0.114	0.128	1.000	0.073
	App_drug	0.000	0.062	0.084	0.087	0.108	1.000	0.042
	FLAP	0.038	0.123	0.171	0.229	0.294	1.000	0.139
	LOX_lib	0.000	0.094	0.117	0.124	0.144	1.000	0.052
PubChem	5-LOX	0.061	0.370	0.459	0.461	0.545	1.000	0.139
	App_drug	0.000	0.221	0.347	0.344	0.456	1.000	0.155
	FLAP	0.293	0.577	0.661	0.674	0.765	1.000	0.124
	LOX_lib	0.071	0.442	0.514	0.514	0.586	1.000	0.116
MACCS	5-LOX	0.000	0.273	0.346	0.360	0.430	1.000	0.132
	App_drug	0.000	0.224	0.310	0.317	0.400	1.000	0.130
	FLAP	0.198	0.405	0.473	0.521	0.632	1.000	0.156
	LOX_lib	0.035	0.342	0.424	0.427	0.508	1.000	0.126

Results show that the degree of similarity values varies in each representation of fingerprints. Overall, similarity values measured with PubChem for any given compound database had the highest values with a mean similarity between 0.34 and 0.67, followed by MACCS keys with a mean similarity between 0.31 and 0.52 and then ECFP4 with a mean similarity between 0.08 and 0.22. That is, for a given dataset, the relative order of the similarity values decreased in the order PubChem > MACCS > ECFP4. This result indicates that among the three fingerprints used, the very low similarity values are obtained with ECFPs, which is the indication of its high resolution. But it provides comparable similarity values for each dataset except for FLAP, so it is not helpful to identify and classify datasets by their structural diversity. As a result, we have selected the MACCS keys as the best choice to differentiate these four datasets.

The Approved drug dataset shows more diversity as it displays the lowest similarity value with all the fingerprints. This may be because the drugs approved by the DrugBank cover a wide range of molecular targets, and each target has its own mechanism of action so that the structure of ligand of each target may also vary. The median similarity values reported for approved drugs are 0.087, 0.344, and 0.317, respectively, when using ECFP4, PubChem, and MAACS key fingerprints.

Furthermore, the distributions of the similarity values show that, in general, 5-LOX inhibitors are structurally diverse as compared to FLAP inhibitors. The least diversity of FLAP inhibitor is well understood from the CDF *vs.* similarity graph with ECFP4 fingerprints. This curve is shown to be far apart from that of all other curves (shifted to the right). These results suggest that new chemical structures need to be created as FLAP inhibitors by covering the large area of chemical space. Finally, LOX library compounds show the diversity that is comparable to the diversity of 5-LOX inhibitors space. This result shows that this virtual library could provide a more novel scaffold that can give positive 5-LOX inhibitory potency in *in-vitro* testing.

Chemical space Characterization and SAR Analysis



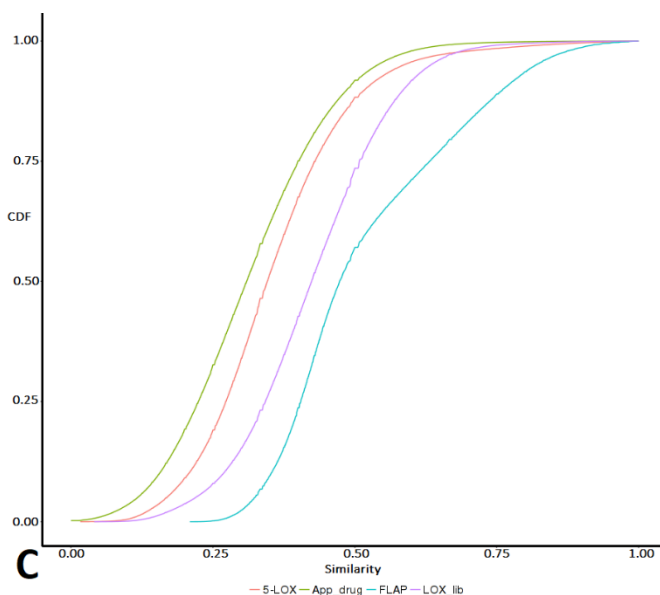


Fig. 4.8 Cumulative Distributions Function (CDF) of pairwise Tanimoto similarity values computed for all datasets A) ECFP4, B) PubChem, and C) MAACS key fingerprints.

4.5.3. Molecular Scaffolds and Scaffold Diversity

The phrase 'Scaffold' is often used to characterize a compound's core structure that is connected to functional groups. To access the diversity of a dataset, you can base the diversity of the scaffold. The dataset contained a more diverse scaffold that indicates its diversity. To enumerate the diversity of 5-LOX and FLAP inhibitors, space 'Murcko framework method' is used, which systematically extracts side chains from the molecules to convert molecules into cyclic systems. Murcko scaffolds were calculated with the program Molecular Equivalent Indices (MEQI). According to this methodology, the cyclic

systems are part of the specific chemotypes, and for each cyclic system, a combination of an equivalencing function with the naming function will produce a simple molecular equivalence index (MEQI). Such codes can be translated and articulated visually, thereby providing a new means to visually depict the diversity of the substances, albeit still unexplored. The cyclic-system and functional group features can be important in analyzing drug-likeness.

1. Scaffold Diversity

The number of cyclic systems (N_c) was reported along with singletons (N_s) were computed to account for the scaffold diversity of each dataset and are given in Table 4.4. Singletons (N_s) are cyclic systems with only one compound. N_T represents the total number of molecules in the database. The fraction of cyclic systems relative to the size of the dataset (N_c/N_T), the fraction of singletons relative to the size of the dataset (N_s/N_T) and the fraction of singleton relative to the number of cyclic systems (N_s/N_c) were also computed for each dataset and depicted in Table 4. 4. The percentage of N_s/N_c and N_s/N_T of 5-LOX dataset is 72.4, and 55.9, respectively, are indicative of the large scaffold diversity of this dataset. Both these values of FLAP are lesser than 5-LOX implies FLAP inhibitor's chemical space is not as diverse as 5-LOX inhibitor's. Comparable N_s/N_c and N_s/N_T values of LOX library and 5-LOX dataset indicate the library designed is contained diverse scaffold as same as that of 5-LOX inhibitor's chemical space. Also, numbers provided in Table 4.4 for Approved drug datasets are comparable to the equivalent fractions of cyclic systems reported in

many kinds of literature. So, our main focus is on the other three datasets.

Table 4.4 Summary table for metrics of scaffold diversity of 5-LOX inhibitors, FLAP inhibitors, LOX library, and Approved drug space

Databases	N_T	N_C	N_C/N_T	N_S	N_S/N_C	N_S/N_T	AUC	F_{50}
5-LOX	1373	1061	0.773	768	0.724	0.559	0.894	0.033
FLAP	1379	744	0.539	480	0.645	0.348	0.904	0.054
LOX_Lib	1387	947	0.683	806	0.851	0.581	0.890	0.028
App_drug	1657	966	0.583	791	0.819	0.477	0.715	0.157

The scaffold diversity of the 5-LOX and FLAP datasets was evaluated by plotting a fraction of cyclic systems on the X-axis and the fraction of compounds containing cyclic systems on the Y-axis. This curve is called cyclic systems recovery (CSR) curves. It measures the fraction of cyclic systems contained in a given fraction of the database. It provides two important terms AUC (Area under the curve) and F_{50} . The term F_{50} is the fraction of cyclic systems that contain 50% of the inhibitors. CSR curves of 5-LOX, FLAP, and LOX virtual library datasets are shown in Figure 4.9, and corresponding AUC and F_{50} values of each dataset are given in Table 4.4. The CSR curve represented by a diagonal with an AUC of 0.5 indicates the maximum diversity dataset. These datasets might have different chemotypes for each compound. As the AUC value increases, the diversity of the dataset decreases. The CSR curves show that the 5-LOX inhibitor dataset has more variety in scaffold content with an AUC value of 0.894 and an F_{50} value of 0.033 as compared to the FLAP. But

generally, the diversity of examined datasets is lower than other datasets in the literature.

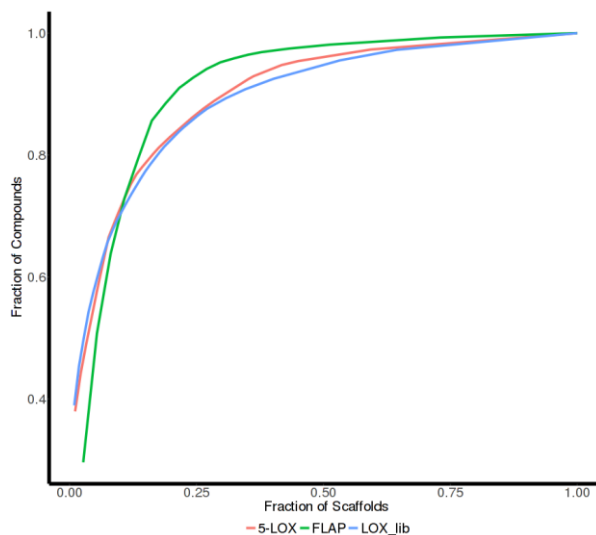


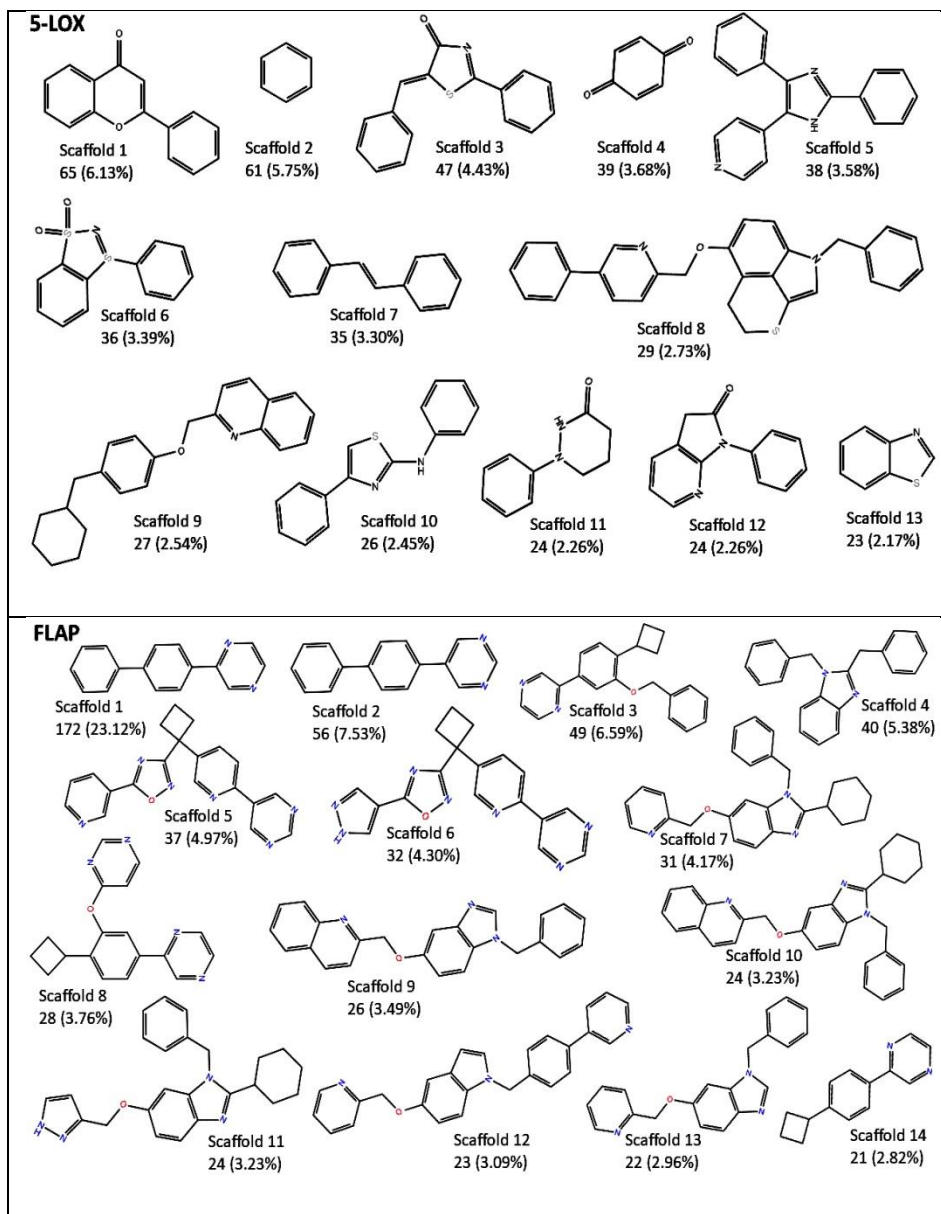
Fig. 4.9 Cyclic system recovery (CSR) curves for the 5-LOX, FLAP, and LOX library datasets.

2. Scaffolds Content in the Databases

The most frequent cyclic systems of the 5-LOX, FLAP, LOX library, and Approve drug sets are shown in Figure 4. 10, along with the cyclic system frequency and percentage. The most frequent scaffold in the 5-LOX set is a flavonoid core and has a frequency of 65 (6.13%) compounds followed by the cyclic benzene system with 61 (5.8%) molecules. These results support the fact that in earlier days of development of 5-LOX inhibitors, investigations are majorly based on the optimization of the compound with redox activity (antioxidant/redox inhibitors) like flavonoids (scaffold 1), chalcones, quinones (scaffold 4), etc. However, due to the lack of oral

bioavailability, poor selectivity, methemoglobin formation, etc., of the redox inhibitors, later, there has been an increased interest in developing non-redox inhibitors (scaffold 6) or iron chelators (a hydroxamic acid derivative of scaffold 10). The most frequent scaffold in the FLAP inhibitor's chemical space is the biphenyl pyrazine system with 172 (23.12%) compounds followed by biphenyl pyridine derivative with 56 (7.53%) compounds. It is evident that the most populated cyclic system of both 5-LOX and FLAP inhibitor set has no common scaffold. This indicates designing molecules with customized polypharmacological profiles may be worthless on these two proteins. Another important finding is that the 'benzene ring,' which can be seen as a most frequent scaffold in approved drugs, is also found highly frequent in the 5-LOX dataset however is not present as a cyclic system in the FLAP set. Because, most of the inhibitors of the FLAP contains complex cyclic systems like in the drug AM-679. The AM-679 is a drug act as a selective inhibitor of FLAP. Here a simple cyclic system like benzene has no role to play. Also, the most frequent scaffold of LOX library with 67 (6.21%) compounds, is a scaffold of the most frequent cyclic system in the 5-LOX set, which is scaffold 10. Besides, some novel scaffolds which are not present in the 5-LOX cyclic system can be seen in the LOX virtual library cyclic system. These scaffolds could be used to design a potent 5-LOX inhibitor in the future.

Chemical space Characterization and SAR Analysis



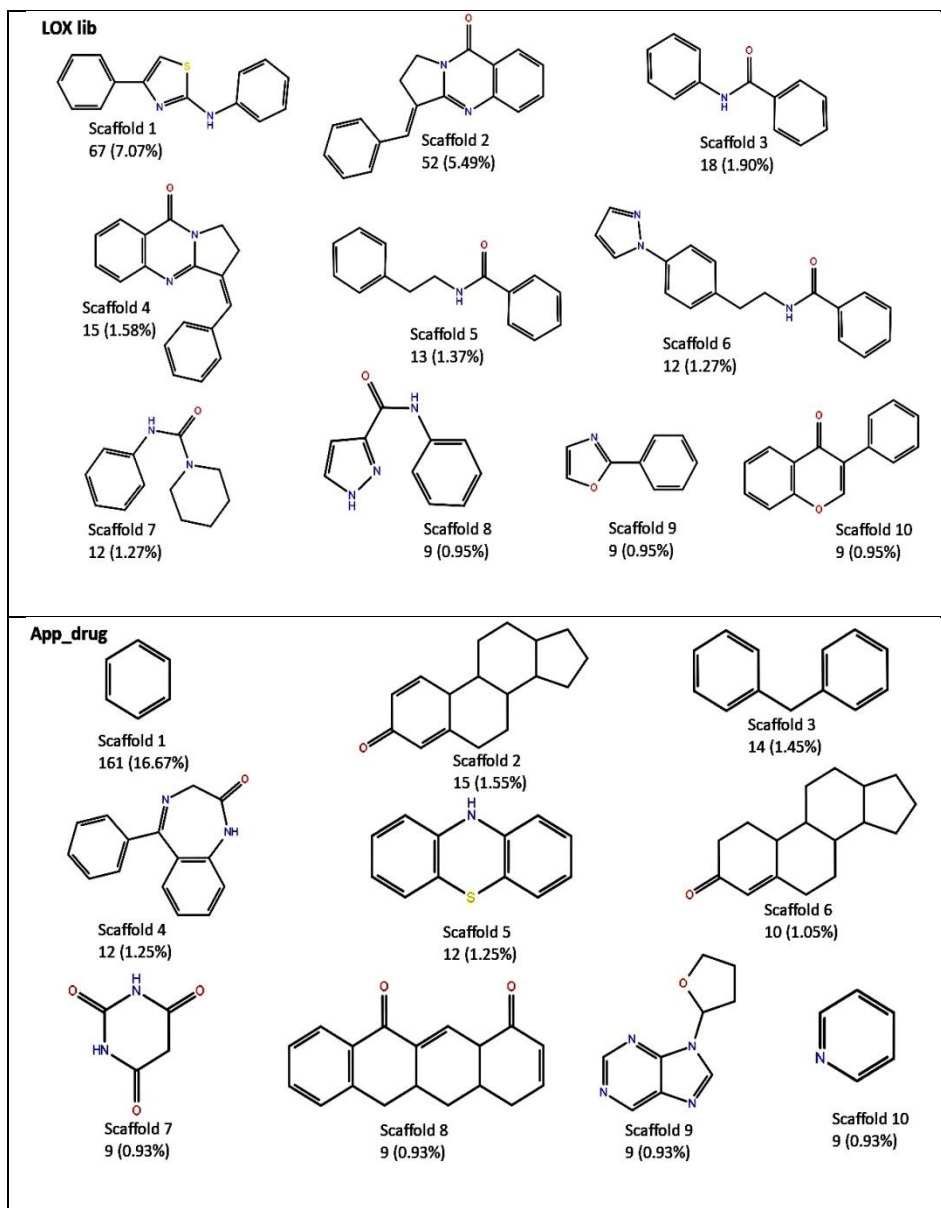


Fig. 4. 10. Most frequent cyclic systems identified in the dataset 5-LOX, FLAP, LOX_lib, and App_drug. For each scaffold, the frequency and relative percentage are indicated in parenthesis.

4.5.4. Scaled Shannon Entropy (SSE)

This method is used to calculate the diversity of the scaffold compounds in the n most populated scaffolds by normalizing Shannon entropy (SE) to the different n . So, the method is called Scaled Shannon entropy (SSE). The Equation 4.1 is represented by the SE of a population of P compounds distributed in n systems and normalization of it to the different n give us Equation 4. 2 which is represented by SSE

$$SE = -\sum_{i=1}^n P_i \log_2 p_i \quad (4.1)$$

$$SSE = \frac{SE}{\log_2 n} \quad (4.2)$$

Where P_i is the estimated probability that a given chemotype i exists in a population of P -compounds that contain a total of n acyclic and cyclic systems. SSE has the value of a real number in the range of 0 to 1. If SSE values are zero, all the molecules in the database contain only one chemotype (minimum diversity), and if it is 1, it indicates each chemotype contains only one compound (maximum diversity). In this study, we calculated the SSE for values ranging from $n = 10$ to 60. Table 4.5 presents the SSE for each dataset's first 60 most popular chemotypes. It demonstrates that compound in the 5-LOX set is much more diverse than those of the compounds in the FLAP set, with SSE levels ranging from 0.982 to 0.956. The top 30 most common

chemotype's distribution and SSE values of compounds are shown in Figure 4.11.

Table 4.5 SSE of the Most Populated Scaffolds ranging from n = 10 to n = 60

dataset	SSE10	SSE20	SSE30	SSE40	SSE50	SSE60
5-LOX	0.982	0.97	0.966	0.963	0.961	0.956
FLAP	0.871	0.884	0.883	0.88	0.876	0.873
LOX_lib	0.832	0.847	0.855	0.858	0.86	0.865
App_drug	0.987	0.994	0.996	0.998	0.998	0.998

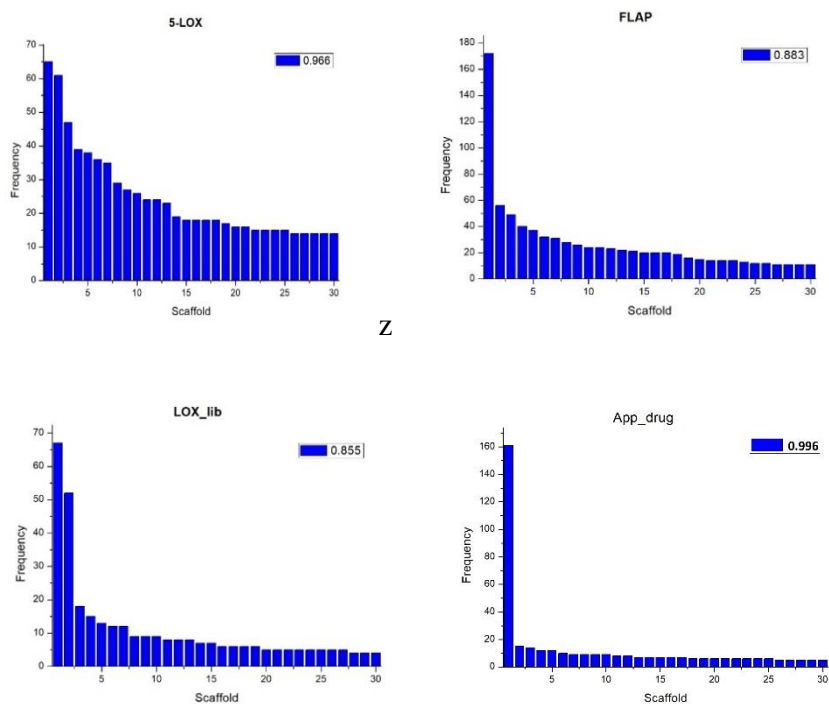


Fig. 4.11 Frequency plot of 30 Most frequent scaffolds of 5-LOX, FLAP, LOX_lib, and App_drug datasets.

4.5.5. Consensus Diversity Plot

We have already discussed the diversity of all the datasets in terms of PCP descriptors, structural fingerprints, and molecular scaffolds. However, each representation has its own advantages and disadvantages. Pharmacological importance and easy interpretation make PCP is a valuable tool for dataset analysis, but in some cases, it is not possible to distinguish compounds using these properties alone. Structural fingerprints normally capture the entire structure's information but are more difficult to interpret. Molecular scaffolds are easy to interpret, but they capture only a part of the chemical structure, it does not contain information from side chains. On this occasion, Consensus Diversity Plots (CDPs) is introduced, which considers these three different structural representations simultaneously to compare the diversity of dataset. This study constructed the CDP of 5-LOX, FLAP, LOX library, and approved drug set in order to compare the diversity of each dataset by considering multiple criteria simultaneously. The summary of the data used for constructing CDP is given in Table 4.6, and the corresponding CDP is shown in Figure 4.12.

Table 4.6 Summary of the Diversity Study

DataSet	5-LOX	FLAP	LOX_lib	App_drug
Size	1373	1379	1387	1657
MACCS	0.36	0.521	0.427	0.317
ECFP	0.114	0.229	0.124	0.087
Fingerprint	0.114	0.229	0.124	0.087
N _C /N _T	0.773	0.539	0.683	0.583
AUC	0.894	0.904	0.89	0.715
F50	0.033	0.054	0.028	0.157
SSEn (n=10)	0.982	0.871	0.832	0.987
PCP	2.69	2.02	1.93	4.03

In CDP (Figure 4.12), each point represents a single dataset. The mean MACCS /Tanimoto, which represent the Fingerprint-based diversity of the dataset was plotted on the X-axis while AUC of the scaffold recovery curves which represent the Scaffold diversity of the dataset was plotted on the Y-axis. The color of each data point represents the diversity of molecular properties, which is based on the mean Euclidean intra dataset distance of six PCP of pharmaceutical relevance. A continuous color scale from red (more diversity) to orange/brown (intermediate diversity) to green (less diversity) reflects these distances. CDPs can be divided into four quadrants by color to assist in the interpretation of the plots, which classify datasets as high/low diverse considering both fingerprints and scaffolds. Red quadrants classify compound datasets with high fingerprint and scaffold diversity. White quadrants identify datasets with relatively low fingerprint diversity and lower scaffold diversity; blue find quadrants with high fingerprint diversity, but low scaffold diversity and yellow quadrants recognize compound libraries with low fingerprint diversity.

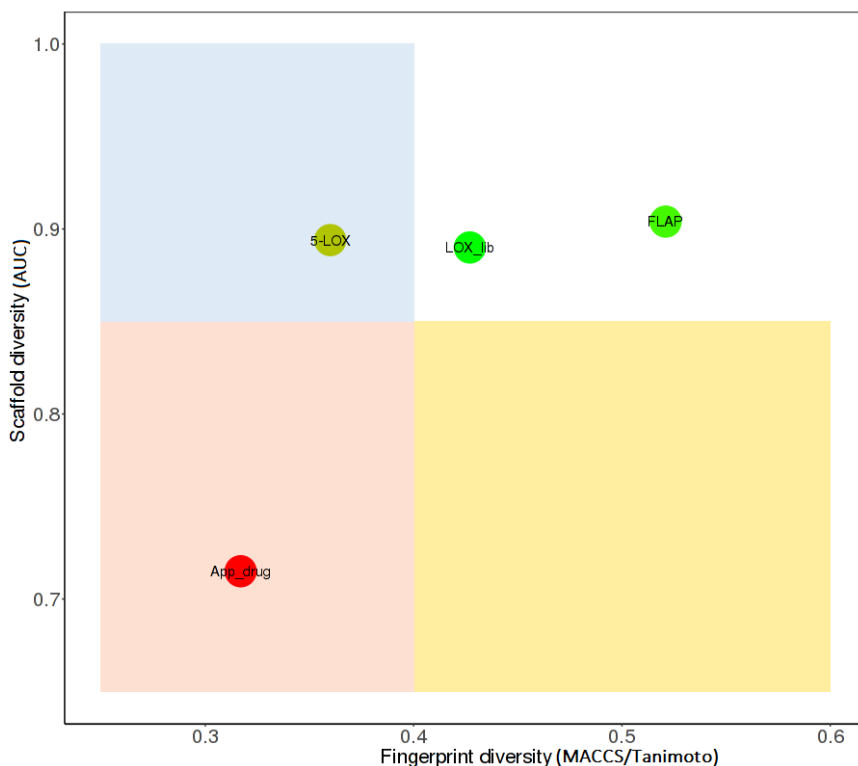


Fig. 4.12 Consensus Diversity Plot

The CDP indicates that, overall, the set of approved drugs is the most diverse, while the set of FLAP inhibitors is the least diverse in all of these four metrics. Compared to all other datasets, the 5-LOX inhibitor set is the second most diverse set, which shows less diversity in terms of scaffolds, average diversity in terms of PCPs, and high diversity in terms of fingerprints.

4. 6. Structure-Activity Relationship

The vast amount of structure-activity data publicly available for 5-LOX and FLAP provides an opportunity to mine Structure-Activity

Relationships (SAR). To characterize SAR, Activity Landscape approaches have been increasingly used because of its special ability to handle SAR of the datasets quickly. Activity Landscape can be considered as the chemical space with an extra dimension of biological activity [38]. Thus, any graphical representation that combines similarity and activity relationships between compounds sharing a specific biological activity can be used to describe an Activity Landscape [39]. Structure-Activity Similarity (SAS) maps and Structure-Activity Landscape Index (SALI) are the two methods that we have used for this study among the many visual and quantitative Activity Landscape approaches that have been developed so far.

4. 6. 1. Structure-Activity Similarity (SAS) Maps

SAS maps are 2D graphs that plot the relationship between the structural similarity with activity similarity for all possible pairs of compounds in the dataset. The similarity of the structure is shown on the X-axis, and the difference in activity is plotted on the Y-axis. In this work, the structural similarity was determined by Tanimoto coefficients using ECFP4 fingerprint. A schematic representation of a SAS map is shown in Figure 4. 13. The map of SAS is divided into four areas: the first one is 'Similarity Cliffs,' i.e., the lower left region, which includes pairs of compounds with low molecular similarities and low activity differences. The second region is 'Smooth SAR'; this is the right bottom area, where you can find pairs of compounds with high molecular similarity and a low activity difference. The third zone is the top-right area containing 'Activity Cliffs,' i.e., pairs of compounds with

high molecular similarity and high activity difference. The top left is the 'Non-Descriptive' region, containing pairs of compounds with low structural similarity and high activity difference.

The detection of activity cliffs is one of the main applications of activity landscape methods. Activity cliffs can be quantified by using the Structure-Activity Landscape Index (SALI). SALI value between two molecules is defined as the ratio between the difference in biological activity (ΔpIC_{50}) to the dissimilarity (1–similarity) of the pair. The researchers have discussed the duality of activity cliffs in medicinal chemistry and computational approaches in drug design. Activity cliffs prevent the effective development of predictive models in predictive computational modeling, such as QSAR. The 'activity cliff generators,' i.e., compounds frequently found in activity cliffs, are not suitable to be used as query molecules in similarity-based virtual screening. Activity cliffs, on the other hand, have a positive effect in medicinal chemistry because they provide key information to pharmacophoric regions and can, therefore, be used for lead optimization. An interesting word from the review by Maykel Cruz-Montegudo et al. is, “For instance, whereas medicinal chemists can take advantage of regions in chemical space rich in activity cliffs, QSAR practitioners need to escape from such regions” [40].

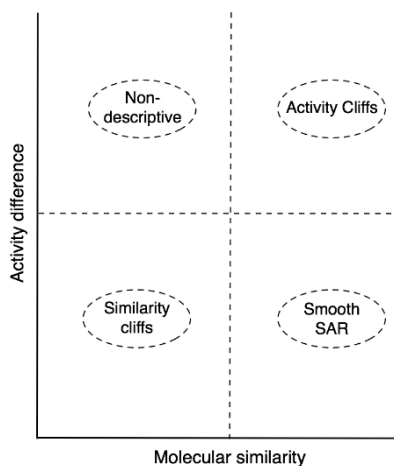


Fig. 4.13 The general form of a Structure-Activity Similarity (SAS) map.

Figure 4.14 shows the SAS map of 939112 and 948496 compound pairs of the 5-LOX and FLAP datasets, respectively. Figure 4.14 separates the four main quadrants (I-IV) by dotted lines. Different criteria have been used to set up a threshold for partitioning the plot. In this work, the threshold for the structure similarity was defined as the median of the distribution of the pairwise similarity values of all compounds plus two standard deviations, and the threshold for the potency difference is taken as two log unit activity differences. So median similarities of these molecules (0.26 for 5-LOX set and 0.507 for FLAP set) taken as the threshold value for the X-axis.

A continuous color scale from red (more data dots) to grey (less) shows the number of data points in each region on the density SAS map of 5-LOX (Figure 4.14A) and FLAP (Figure 4.14B). Here high data point density can be seen in the similarity cliff region of both

the 5-LOX and FLAP SAS maps, indicating that most of the compounds in the chemical space of these proteins have a different structure but similar activity value. Figures 4.14C and 4.14D represent the color of the most active compound in the similarity pairs of 5-LOX and FLAP datasets, respectively. Where red color dots are the most active compounds in each pair, yellow-to-orange color dots are the pair's intermediate active compounds, and green color dots are the least active compound in the pair. It should be noted that the red points in the SAS map of 5-LOX and FLAP are distributed along with the whole range of the potency difference. The red dots at the top of the plot (high potency difference values) contain one active and one inactive compound. On the other hand, the red dots at the bottom of the plot (low potency difference) denote a pair of compounds where both are active.

Similarly, the SALI-SAS Map (4.14E and 4.14F) will contain green colored dots representing the lowest SALI compound pairs, the orange to yellow dots are intermediate SALI compound pairs, and the red dot indicates the highest SALI compound pairs. In both SALI-SAS maps, almost all compound pairs are in green in color and have low SALI value. This result indicates that small structural changes yield only small changes in the activity.

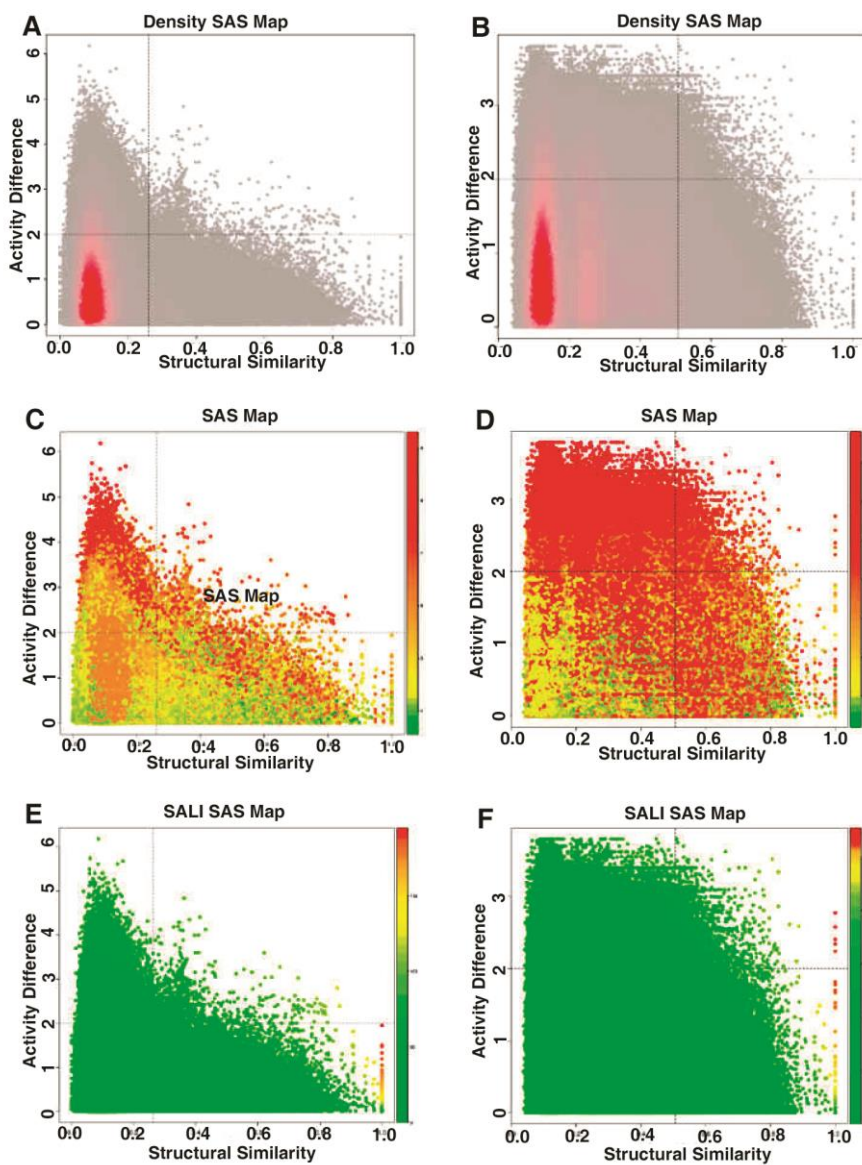


Fig. 4.14. SAS maps of the global activity landscapes of 5-LOX and FLAP inhibitors: A and B are density SAS map, C and D are maximum activity SAS map, and E and F are SALI map of 5-LOX and FLAP dataset respectively.

The number of similarity pairs in each area (I-IV) of the SAS map is summarized in Table 4.7 indicates that 5-LOX and FLAP, in particular, have a heterogeneous SAR with data points in the SAR's continuous and discontinuous regions. The quantitative analysis indicates that FLAP inhibitors dataset has the largest proportion of activity cliffs (0.38%) as compared to the 5-LOX inhibitor dataset. This indicates the rough nature of the SAR for compounds associated with FLAP. Noteworthy is that the scaffold hop/similarity cliff region has the highest data point density for both the 5-LOX and FLAP inhibitor datasets, which is 85.60% and 82.54%, respectively. This indicates that nearly 85% of both dataset's compound pairs contain quite different chemical structures with similar activity. Only 2.83 and 4.61% of the 5-LOX and FLAP pairs respectively are in a smooth SAR region. These proportions depend on the current ChEMBL content, i.e., the number may change on the publication of more activity data.

Table 4. 7 Quantitative analysis of the SAS maps

Quadrant	Region	5-LOX		FLAP	
		No. of pairs	Percentage	No. of pairs	Percentage
1	Uncertainty	107642	11.46	118262	12.47
2	Similarity cliff	803860	85.60	782928	82.54
3	Smooth SAR	26622	2.83	43689	4.61
4	Activity cliffs	988	0.11	3617	0.38
	Total	939112		948496	

4. 6. 2. Activity Cliff Generators and SAR Interpretation

This study has already mentioned the definition and importance of the 'Activity Cliff' landscape region. In the SAS map, this quadrant

contains compound pairs with high similarity in structure and a high difference in potency. Activity cliff generators are molecules that are frequently found in these activity cliff regions. It is expected that direct analysis and interpretation of these activity cliff generators will provide insight into the relevant characteristics of 5-LOX and FLAP inhibitory potency of an inhibitor. This study found eight important compounds identified as active cliff generators in the 5-LOX and FLAP inhibitor sets. The compound pairs in the 5-LOX and FLAP deep activity cliff region are respectively given Tables 4.8 and 4.9 along with their similarity, activity difference, and SALI values.

Table 4.8 Deep activity cliffs formed by compounds in the 5-LOX set

Compound Pair	Similarity	ΔpIC_{50}	SALI	Compound Pair	Similarity	ΔpIC_{50}	SALI
1341879_1341887	0.533	2.003	4.293	1758852_1426797	0.690	2.209	7.128
1341880_1341887	0.519	2.070	4.308	1758852_1758843	0.629	2.156	5.806
1341882_1341887	0.581	2.195	5.240	1758852_1758846	0.566	2.312	5.324
1341883_1341887	0.560	2.149	4.885	1758852_1758847	0.605	2.593	6.569
1341884_1341887	0.564	2.195	5.036	1758852_1758857	0.482	2.395	4.627
1341888_1341887	0.459	2.274	4.205	19081_19004	0.688	2.134	6.827
1341890_1341887	0.433	2.035	3.590	19119_19004	0.652	2.134	6.122
1341891_1341887	0.575	2.371	5.579	19168_19004	0.623	2.134	5.662
1341893_1341887	0.419	2.195	3.776	19171_19004	0.606	2.134	5.410
1368234_2076274	0.426	3.000	5.226	19187_19004	0.606	2.134	5.410
1368234_267704	0.554	2.243	5.033	232686_17576	0.586	2.347	5.675
1368234_268400	0.539	2.011	4.364	232778_17576	0.740	2.125	8.168
1368234_268497	0.462	2.176	4.041	232818_17576	0.719	2.187	7.783
1368234_83058	0.402	2.097	3.506	233037_17576	0.710	2.125	7.319
1368234_83097	0.484	2.097	4.063	233087_17576	0.558	2.240	5.068
141397_141024	0.423	2.548	4.414	247101_141937	0.413	2.230	3.799
141397_141575	0.418	2.146	3.690	247101_141992	0.411	2.097	3.558
141397_141937	0.453	2.054	3.753	247101_142183	0.411	2.114	3.590
141397_142274	0.427	3.000	5.237	247101_212260	0.479	2.439	4.682
141397_25982	0.419	2.176	3.746	247101_212684	0.487	2.477	4.832
141397_26054	0.701	2.519	8.427	267686_2076274	0.407	2.412	4.068
141856_109413	0.409	3.602	6.091	267687_2076274	0.413	2.648	4.510

Chemical space Characterization and SAR Analysis

141856_109943	0.415	2.944	5.032	267752_2076274	0.442	2.699	4.840
141856_141024	0.730	3.025	11.194	268305_2076274	0.429	2.426	4.245
141856_141355	0.442	2.114	3.789	268306_2076274	0.421	2.757	4.758
141856_141442	0.620	3.602	9.485	268853_2076274	0.425	2.426	4.216
141856_141516	0.671	3.279	9.973	268854_2076274	0.418	2.441	4.195
141856_141575	0.720	2.623	9.369	268891_2076274	0.409	2.472	4.183
141856_141890	0.733	2.230	8.364	268964_2076274	0.405	2.412	4.054
141856_141937	0.730	2.531	9.366	268965_2076274	0.427	2.187	3.818
141856_141992	0.714	2.398	8.393	269003_2076274	0.414	2.602	4.444
141856_142183	0.761	2.415	10.086	269023_2076274	0.404	2.077	3.482
141856_142495	0.768	2.398	10.341	269032_2076274	0.402	2.581	4.314
141856_26007	0.438	2.799	4.983	269035_2076274	0.426	2.789	4.859
142008_142274	0.405	2.176	3.660	275665_141093	0.430	2.125	3.727
142050_142274	0.500	2.176	4.352	275665_142274	0.409	4.000	6.765
142057_142274	0.405	2.176	3.660	275665_275101	0.432	4.125	7.257
142217_142274	0.609	2.875	7.347	275665_275107	0.455	3.301	6.052
142227_142274	0.626	2.574	6.889	275665_373008	0.815	2.279	12.332
142255_142274	0.622	2.176	5.760	275665_373025	0.817	2.204	12.058
142259_142274	0.629	2.574	6.942	275665_373033	0.800	2.442	12.210
142333_142274	0.583	2.778	6.668	275665_373040	0.817	2.380	13.021
153228_17576	0.593	2.058	5.063	275665_373053	0.809	2.409	12.582
153322_17576	0.589	2.155	5.245	275665_373062	0.784	2.452	11.327
153354_17576	0.642	2.084	5.824	275665_373071	0.863	2.392	17.481
153579_17576	0.748	2.523	10.005	275665_373079	0.730	2.404	8.903
153685_17576	0.661	2.155	6.360	275665_373080	0.716	2.426	8.533
153762_17576	0.782	2.034	9.322	275665_645447	0.413	4.407	7.508
153778_17576	0.531	2.084	4.440	41649_144242	0.418	2.058	3.537
153779_17576	0.567	2.240	5.173	41649_17620	0.413	2.854	4.859
154143_17576	0.680	2.561	8.002	41649_2076274	0.463	3.058	5.694
154145_17576	0.723	2.323	8.378	41649_267704	0.470	2.301	4.342
154158_17576	0.795	2.240	10.909	41649_83097	0.464	2.155	4.020
154256_17576	0.613	2.398	6.203	70615_153321	0.478	2.442	4.678
154342_17576	0.512	2.155	4.416	70615_153881	0.577	2.146	5.073
154406_17576	0.562	2.301	5.248	70615_17575	0.511	2.477	5.064
1610461_107254	0.423	2.271	3.937	70615_17576	0.530	3.125	6.647
1610461_1783014	0.534	2.166	4.651	70615_17620	0.555	3.222	7.235
1610461_1783028	0.527	2.014	4.259	70615_232685	0.507	2.079	4.219
1610461_1783029	0.500	2.067	4.134				
1610461_1783030	0.422	2.213	3.827				

Table 4.9 Deep activity cliffs formed by compounds in the FLAP set

Compound pairs	Similarity	ApKi	SALI	Compound pairs	Similarity	ApKi	SALI
2021582, 2021722	0.565	3.122	7.18	2030555, 2026440	0.736	3.014	11.411
2021582, 2021770	0.558	3.081	6.978	2030555, 2026445	0.639	3.097	8.587
2021582, 2026388	0.535	3.097	6.658	2030555, 2026451	0.574	3.503	8.215
2021582, 2026389	0.535	3.136	6.743	2030555, 2030441	0.661	3.62	10.678
2021582, 2030441	0.526	3.398	7.173	2030555, 2030463	0.52	3.119	6.499
2021582, 2030465	0.538	3.288	7.123	2030555, 2030465	0.619	3.509	9.212
2021582, 2030548	0.512	3.396	6.954	2030555, 2030578	0.542	3.499	7.635
2021582, 2030550	0.56	3.396	7.71	2030555, 2030586	0.609	3.349	8.573
2021582, 2030601	0.545	3.396	7.472	2030555, 2030592	0.544	3.142	6.892
2021582, 2030640	0.513	3.149	6.469	2030555, 2030640	0.672	3.371	10.291
2021582, 2034098	0.592	3.064	7.511	2030555, 2034074	0.574	3.178	7.452
2021621, 2021722	0.588	3.122	7.581	2030555, 2034085	0.6	3.215	8.037
2021621, 2021770	0.558	3.081	6.978	2030555, 2034098	0.609	3.286	8.411
2021621, 2030441	0.526	3.398	7.173	2030555, 2034102	0.557	3.195	7.214
2021621, 2030465	0.519	3.288	6.835	2030555, 2034107	0.557	3.285	7.417
2021621, 2030544	0.553	3.396	7.592	2030555, 2034110	0.52	3.269	6.811
2021621, 2030640	0.554	3.149	7.062	2030555, 2034118	0.534	3.411	7.323
2021621, 2034098	0.571	3.064	7.149	2030555, 2034202	0.569	3.318	7.703
2026279, 2026276	0.672	3.097	9.431	2030555, 2034203	0.521	3.392	7.083
2026279, 2026311	0.556	3.097	6.968	2034153, 2021770	0.519	3.303	6.86
2026279, 2026313	0.592	3.097	7.592	2034153, 2034098	0.512	3.286	6.736
2026279, 2030548	0.573	3.095	7.252	2034153, 2034147	0.566	3.025	6.975
2026279, 2030550	0.625	3.095	8.254	2034153, 2034175	0.824	3.34	19.011
2026279, 2030601	0.553	3.095	6.919	2034153, 2034202	0.538	3.318	7.175
2030490, 2021722	0.549	3.122	6.926	2034153, 2034203	0.518	3.392	7.032
2030490, 2026451	0.568	3.281	7.594	2037632, 1992153	0.69	3	9.693
2030490, 2030441	0.513	3.398	6.975	2037632, 2037550	0.517	3.062	6.343
2030490, 2030578	0.541	3.278	7.144	2037632, 2037615	0.521	3.284	6.86
2030490, 2030586	0.618	3.127	8.195	2037632, 2037624	0.52	3.228	6.732
2030490, 2030640	0.519	3.149	6.554	2037632, 2037653	0.621	3.236	8.538
2030555, 2021722	0.765	3.343	14.209	2037637, 1992153	0.616	3	7.818
2030555, 2021733	0.656	3.04	8.831	2037637, 2037550	0.518	3.062	6.348
2030555, 2021736	0.629	3.087	8.321	2037637, 2037615	0.556	3.284	7.389
2030555, 2021767	0.513	3.048	6.261	2037637, 2037624	0.521	3.228	6.738
2030555, 2021770	0.619	3.303	8.67	2037637, 2037653	0.573	3.236	7.576
2030555, 2026438	0.78	3.014	13.701				

Figure 4.15 shows the chemical structures of two representative cliff generators of 5-LOX and FLAP dataset provided with ChEMBL Molregno and activity value. All compounds in this figure have a potency difference $pIC_{50} > 2$ with respective cliff generator. Activity

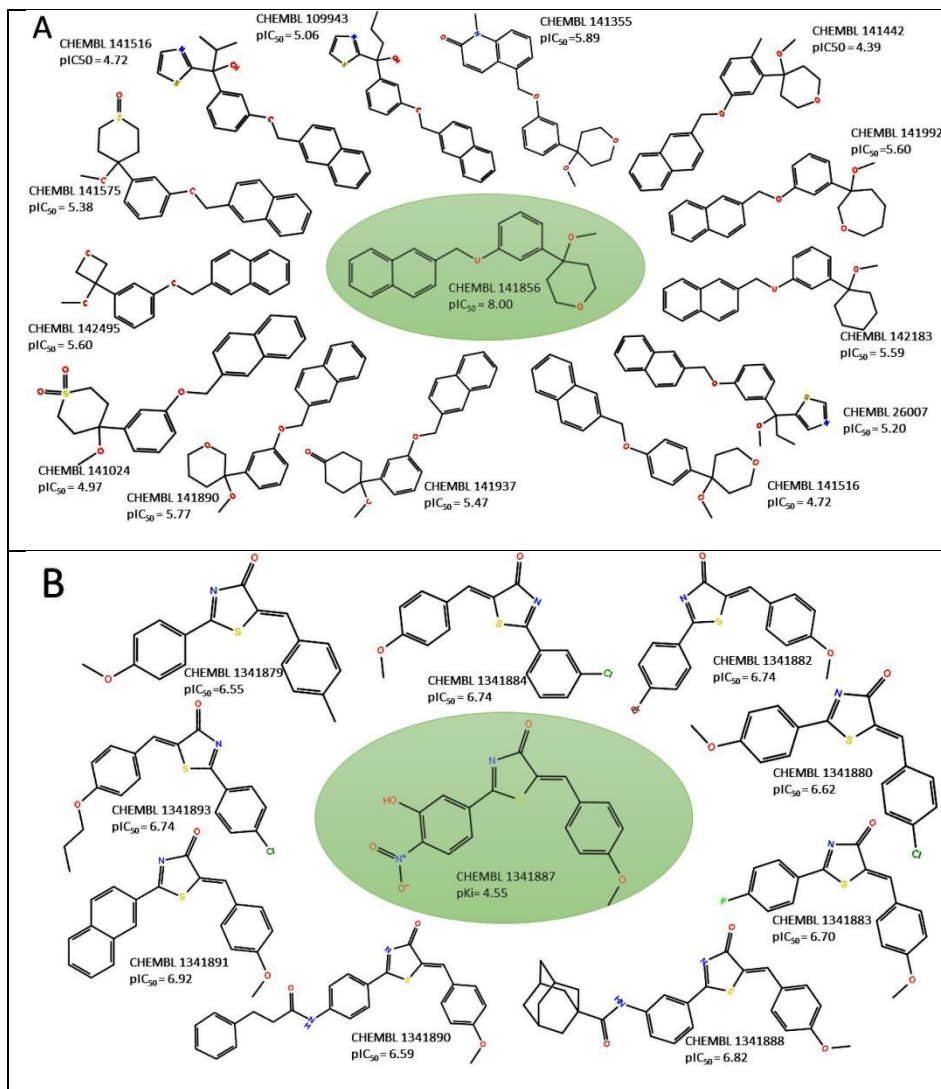
cliffs associated with 4-Methoxy-4-[3-(2-naphthylmethoxy) phenyl] tetrahydro-2H-pyran (ChEMBL Molregno 141856) highlights the relevance and sensitivity of tetrahydro-2H-pyran with a Methoxy group for binding (Figure 4.15A). Other heterocycles instead of tetrahydro-2H-pyran and any substitution in the phenyl group are found to affect the binding activity negatively. Similarly, the activity cliffs formed with the generator (5Z)-2-(3-Hydroxy-4nitrophenyl)-5-[(4-methoxyphenyl)methylidene]-1,3-thiazol-4-one (ChEMBL Molregno 1341887) is also interesting (Figure 4.15B). All substitutions except the Hydroxy group at the third position and nitro group at the fourth position of the phenyl group are found to increase the activity. Modifying the common structure to improve activity based on information from activity cliff generators is, therefore, a way to the process of lead optimization of 5-LOX. All activity cliff generators of 5-LOX, having different chemical structures. So there are no benefits for carrying out 'local SAS' study.

In the activity landscape of the FLAP inhibitor's chemical space, most of the activity cliff generators are of compounds of pyrazin-2-amine derivatives. The derivatives of the pyrazin-2-amine compound are the most promising FLAP modulators that have recently got patent. However, this study found that these are the compounds with the most dramatic changes in activity associated with a small change in the structure. Deep activity cliffs generators of two such pyrazin-2-amine related compounds are illustrated as an example. These are compounds of biphenyl pyrazin-2-amine and compounds of 3-Phenylpyridine pyrazin-2-amine. Figure 4.15C and 4.15D shows the

compound pairs that form the cliffs with them. Figure 4.15C shows the activity cliff associated with 6-[2-[4-(5-Aminopyrazin-2-yl)-3-fluorophenyl]phenoxy]pyrimidine-4-amine (ChEMBL Molregno 2030490) and the chemical structures of the six cliff-forming compounds. By analyzing the structural difference between them, it is understood that the substituent pyrimidin-4-amine abruptly increases binding affinity. Replacing phenoxy pyrimidin-4-amine with Cyclopentylsulfonylpyrimidin increases the binding affinity (Figure 4.15D), but the molecule 5-[4-(2-Cyclopentylsulfonylpyridin-3-yl)-2-fluorophenyl]pyrazin-2-amine (ChEMBL Molregno 2034153) itself can be seen as an activity cliff generator (Figure 4. 15D). This result indicates the need for a local SAS study for compounds of pyrazin-2-amine derivatives. This finding also illustrates the difficulty of performing QSAR like predictive modeling using biphenyl pyrazine-2-amine compounds. They represent a discontinuity between structure and activity, and SAR continuity provides the fundamental basis for QSAR analyses. However, chemical modifications of ChEMBL 2030490 and ChEMBL 2034153 as a lead compound, might be improved potency, selectivity, or pharmacokinetic parameters.

So, the identification of activity cliffs in compound datasets may be extremely important in guiding the construction of predictive models. Removing active cliffs from compound datasets would improve the performance of predictive models that are explicitly based on the similarity principle, such as traditional QSAR approaches. So, it is necessary to develop and test different predictive models with and

without the activity cliffs and assess the predictive power of 5-LOX, and FLAP dataset studied in this work.



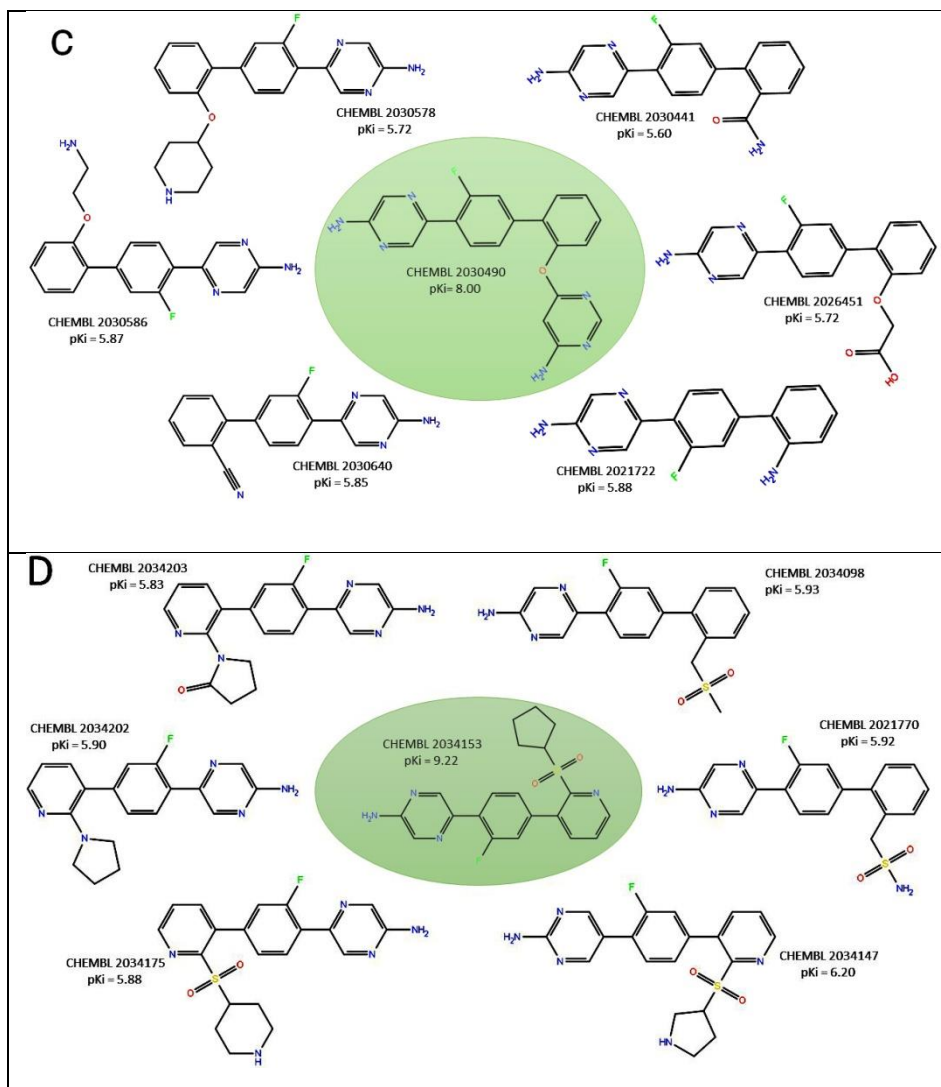


Fig. 4.15 Representative activity cliff generators and selected pairs of compounds formed with the generators (A) (4-Methoxy-4-[3-(2 naphthylmethoxy)phenyl]tetrahydro-2H-pyran and (B) (5*Z*)-2-(3-Hydroxy-4 nitrophenyl)-5-[(4-methoxyphenyl)methylidene]-1,3-thiazol-4-one in the activity landscape of 5-LOX inhibitor's chemical space and (C) 6-[2-[4-(5-Aminopyrazin-2-yl)-3-fluorophenyl]phenoxy]pyrimidin-4-amine and (D) 5-[4-(2-Cyclopentylsulfonylpyridin-3-yl)-2-fluorophenyl]pyrazin-2-amine in the activity landscape of FLAP inhibitor's chemical space.

4. 6. 3. Chemotype Enrichment

Chemical scaffolds can also be used to explore the SAR of the datasets. Each scaffold could be considered as a cluster of molecules with biological activity expressed against a particular biological target. Finding the clusters which have a higher or lower proportion of active molecules is very important to understand the SAR. This can be achieved by chemotype enrichment analysis or scaffold enrichment analysis. For this, the proportion of active compounds in a given most frequent scaffold of 5-LOX and FLAP dataset relative to the fraction of active compounds in the entire datasets were analyzed, and the measurement result provides a term called 'Enrichment Factor (EF).' The method for computing EF was given in Equation 4.3.

$$EF(C_{\lambda}) = \frac{Act(C_{\lambda})}{Act(C)} \quad (4.3)$$

$Act(C)$ is the fraction of active compounds in the database and was determined using the Equation 4. 4 while $Act(C_{\lambda})$ is the fraction of active compounds in a specific chemotype and was calculated with the Equation 4.5:

$$Act(C) = \frac{[C^*]}{[C]} \quad (4.4)$$

$$Act(C_{\lambda}) = \frac{[C_{\lambda}^*]}{[C_{\lambda}]} \quad (4.5)$$

where $[C]$ is the total number of compounds, and $[C^*]$ is the total number of active compounds. $[C_{\lambda}]$ and $[C_{\lambda}^*]$, respectively, are the total

number of compounds and active compounds in the chemotype class λ . In this study compounds with $pIC_{50} > 6$ ($IC_{50} > 1000nM$) are considered as the 'active compounds' in the 5-LOX inhibitor's chemical space while compounds with $pKi > 7$ ($Ki > 100nM$) are considered as the 'active compounds' in the FLAP inhibitor's chemical space.

The molecular scaffolds with the highest EF are the most desirable. A chemotype enrichment plots were also generated to further differentiate molecular scaffolds with the highest frequency of activity by plotting the EF on the X-axis and the cyclic system frequency on the Y-axis. Figure 4.16 shows the chemotype enrichment plot for the nineteen and twenty-one most frequent cyclic systems identified for the set of inhibitors of 5-LOX and FLAP, respectively.

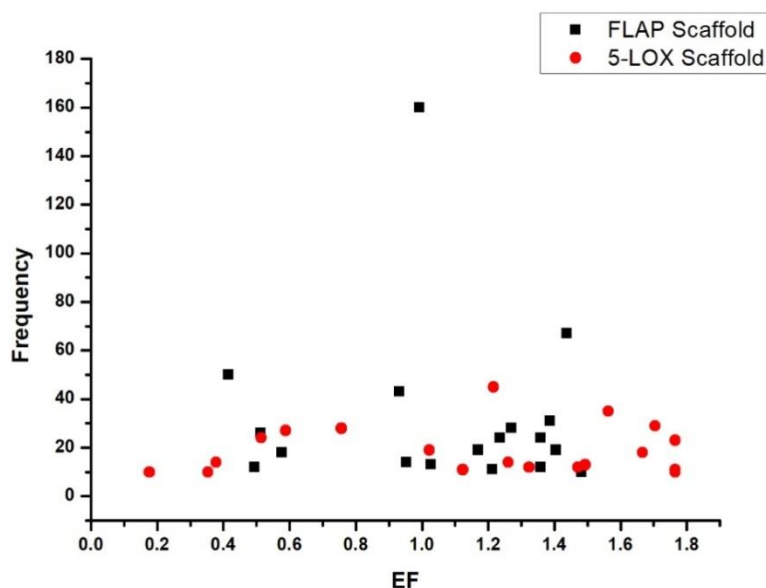


Fig. 4.16 Chemotype enrichment plot of 5-LOX and FLAP inhibitors.

High-frequency scaffolds with higher valued EFs are of getting particular interest in drug design, as they have more information about the SAR and are enriched with active compounds for the targets in which they have been tested. This enrichment plot shows that there are 13 and 14 cyclic systems with EF greater than one present in 5-LOX and FLAP inhibitor's chemical space, respectively. Out of which, 6 and 8 scaffolds of 5-LOX and FLAP inhibitors respectively have a frequency greater than 15. The most active and most frequent scaffolds of 5-LOX and FLAP inhibitors are reported in Figure 4.17. This finding indicates that the datasets examined in this study comprise of cyclic systems with a large proportion of active molecules.

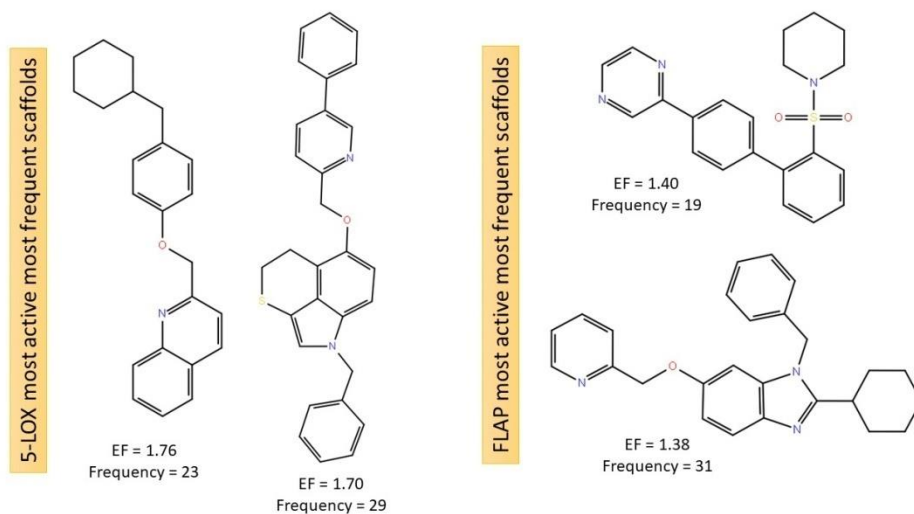


Fig. 4.17 Most active and most frequent scaffolds in 5-LOX and FLAP inhibitor's chemical space.

4.7. Conclusion

This Chapter discusses a comprehensive cheminformatic characterization of the chemical space of 5-LOX and FLAP inhibitors obtained from the ChEMBL database by comparing it with the Approved Drug space. Also, we have compared the virtual LOX library created by the Enamine database with the 5-LOX dataset. Analysis of the distributions of PCPs like HBA, HDB, and TPSA indicated that the compounds screened as inhibitors of 5-LOX and FLAP are, in general, less or comparable polar to the approved drugs in the drug database. From the distribution of RTB, it is recognized that the flexibility of compounds in these databases is similar to that of drugs in the drug database. PCA results of 5-LOX, FLAP, and LOX library sets show that properties that are associated with the polarity of the compound have a significant contribution toward each PC. The visual representation of the property space indicates most of the 5-LOX inhibitors space is within the drug database property space; however, some compounds broaden the traditional medicinal space. Also, the distinct areas of equal color in all 3D PCA maps are evidence of the separation of chemical space that a PCA can achieve, although the dataset consists of somewhat different structures.

The structural diversity of the databases is computed using complementary approaches, including PCP descriptors, molecular fingerprints, and molecular Scaffold. With the obvious exception of approved drugs, the 5-LOX dataset shows more diversity compared to FLAP and LOX library set. It has the largest intra-dataset distance,

lowest similarity value for all fingerprints, highest SSE levels (0.982 - 0.956), near-diagonal cyclic recovery curve (CSR), maximum AUC (0.894) and F_{50} (0.033) and high N_S/N_C (72.4%) and N_S/N_T (55.9%) percentage. The Consensus Diversity Plot (CDP) also supports this, and it says 5-LOX inhibitor set is the second most diverse set after approved drugs, showing less scaffold diversity, average PCP diversity, and high fingerprint diversity. Besides, LOX library compounds show the diversity that is comparable to the diversity of 5-LOX inhibitors space. FLAP inhibitor set is the least diverse set and is well understood from all the diversity matrices. The scaffold content analysis shows that the most populated cyclic system of both 5-LOX and FLAP inhibitor set has no common scaffold. Besides, some novel scaffold which is not present in the 5-LOX cyclic system can be seen in the LOX virtual library cyclic system. This scaffold could be used to design a potent 5-LOX inhibitor in the future.

SAR of the dataset is studied using activity landscape analysis and chemotype enrichment. Evaluation of the activity landscape of 5-LOX and FLAP inhibitors showed an overall heterogeneous SAR with most of the molecule are in similarity cliff region, and some are in the activity cliff region. The rough nature of the SAR for FLAP inhibitors is due to the presence of the largest proportion of activity cliffs (0.38%) as compared to the 5-LOX inhibitor dataset. We have found eight important active cliff generators in the 5-LOX, and FLAP inhibitor sets contain several pharmacophoric interactions that are substantial to determine its potency. This enrichment plot shows that there are 13 and 14 cyclic systems with EF greater than 1 present in 5-

LOX and FLAP inhibitor's chemical space, respectively. Out of which, 6 and 8 scaffolds of 5-LOX and FLAP inhibitors respectively have a frequency greater than 15. This finding indicates that the datasets examined in this study comprise of cyclic systems with a large proportion of active molecules. In short, the smooth SAR region present in the 5-LOX chemical space open up the possibility of the development of highly qualitative and robust QSAR models.

References

- [1] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.* 40 (2012) D1100–D1107. doi:10.1093/nar/gkr777.
- [2] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem Substance and Compound databases, *Nucleic Acids Res.* 44 (2016) D1202–D1213. doi:10.1093/nar/gkv951.
- [3] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.* 35 (2007) D198–D201. doi:10.1093/nar/gkl999.
- [4] J.-L. Reymond, The Chemical Space Project, *Acc. Chem. Res.* 48 (2015) 722–730. doi:10.1021/ar500432k.
- [5] E.F. Gortari, J.L. Medina-Franco, Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases, *RSC Adv.* 5 (2015) 87465–87476. doi:10.1039/C5RA19611F.
- [6] N. Singh, R. Guha, M.A. Giulianotti, C. Pinilla, R.A. Houghten, J.L. Medina-Franco, Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository, *J. Chem. Inf. Model.* 49 (2009) 1010–1024. doi:10.1021/ci800426u.
- [7] A. Yosipof, R.C. Guedes, A.T. García-Sosa, Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category, *Front. Chem.* 6 (2018) 162. doi:10.3389/fchem.2018.00162.
- [8] J.L. Medina-Franco, Interrogating Novel Areas of Chemical Space for Drug Discovery using Chemoinformatics, *Drug Dev. Res.* 73 (2012) 430–438. doi:10.1002/ddr.21034.
- [9] G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe Jr, Computational methods in drug discovery, *Pharmacol. Rev.* 66 (2013) 334–395. doi:10.1124/pr.112.007336.

- [10] C. Lipinski, A. Hopkins, Navigating chemical space for biology and medicine, *Nature*. 432 (2004) 855–861. doi:10.1038/nature03193.
- [11] Y.A. Ivanenkov, N.P. Savchuk, S. Ekins, K. V Balakin, Computational mapping tools for drug discovery, *Drug Discov. Today*. 14 (2009) 767–775. doi:https://doi.org/10.1016/j.drudis.2009.05.016.
- [12] A.S. Rifaioglu, H. Atas, M.J. Martin, R. Cetin-Atalay, V. Atalay, T. Doğan, Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases, *Brief. Bioinform.* (2018). doi:10.1093/bib/bby061.
- [13] M. Wawer, E. Lounkine, A.M. Wassermann, J. Bajorath, Data structures and computational tools for the extraction of SAR information from large compound sets, *Drug Discov. Today*. 15 (2010) 630–639. doi:https://doi.org/10.1016/j.drudis.2010.06.004.
- [14] A. Golbraikh, X.S. Wang, H. Zhu, A. Tropsha, Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment BT - Handbook of Computational Chemistry, in: J. Leszczynski (Ed.), Springer Netherlands, Dordrecht, 2016: pp. 1–38. doi:10.1007/978-94-007-6169-8_37-2.
- [15] S. Pirhadi, F. Shiri, J.B. Ghasemi, Multivariate statistical analysis methods in QSAR, *RSC Adv.* 5 (2015) 104635–104665. doi:10.1039/C5RA10729F.
- [16] J.J. Naveja, F.I. Saldívar-González, N. Sánchez-Cruz, J.L. Medina-Franco, Cheminformatics Approaches to Study Drug Polypharmacology BT - Multi-Target Drug Design Using Chem-Bioinformatic Approaches, in: K. Roy (Ed.), Springer New York, New York, NY, 2019: pp. 3–25. doi:10.1007/7653_2018_6.
- [17] M. González-Medina, J.L. Medina-Franco, Chemical Diversity of Cyanobacterial Compounds: A Chemoinformatics Analysis, *ACS Omega*. 4 (2019) 6229–6237. doi:10.1021/acsomega.9b00532.
- [18] F.I. Saldívar-González, E. Lenci, A. Trabocchi, J.L. Medina-Franco, Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study, *RSC Adv.* 9 (2019) 27105–27116. doi:10.1039/C9RA04841C.
- [19] F.I. Saldívar-González, M. Valli, A.D. Andricopulo, V. da Silva

- Bolzani, J.L. Medina-Franco, Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization, *J. Chem. Inf. Model.* 59 (2019) 74–85. doi:10.1021/acs.jcim.8b00619.
- [20] F.D. Prieto-Martínez, A. Peña-Castillo, O. Méndez-Lucio, E. Fernández-de Gortari, J.L. Medina-Franco, Chapter One - Molecular Modeling and Chemoinformatics to Advance the Development of Modulators of Epigenetic Targets: A Focus on DNA Methyltransferases, in: C.Z.B.T.-A. in P.C. and S.B. Christov (Ed.), *Insights into Enzym. Mech. Funct. from Exp. Comput. Methods*, Academic Press, 2016: pp. 1–26. doi:<https://doi.org/10.1016/bs.apcsb.2016.05.001>.
- [21] F.D. Prieto-Martínez, E.F. Gortari, O. Méndez-Lucio, J.L. Medina-Franco, A chemical space odyssey of inhibitors of histone deacetylases and bromodomains, *RSC Adv.* 6 (2016) 56225–56239. doi:10.1039/C6RA07224K.
- [22] J.J. Naveja, U. Norinder, D. Mucs, E. López-López, J.L. Medina-Franco, Chemical space, diversity and activity landscape analysis of estrogen receptor binders, *RSC Adv.* 8 (2018) 38229–38237. doi:10.1039/C8RA07604A.
- [23] P. Aparoy, K. Kumar Reddy, P. Reddanna, Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors, *Curr. Med. Chem.* 19 (2012) 3763–3778.
- [24] D. Steinhilber, B. Hofmann, Recent advances in the search for novel 5-lipoxygenase inhibitors, *Basic Clin. Pharmacol. Toxicol.* 114 (2014) 70–77. doi:10.1111/bcpt.12114.
- [25] C. Pergola, O. Werz, 5-Lipoxygenase inhibitors: a review of recent developments and patents, *Expert Opin. Ther. Pat.* 20 (2010) 355–375. doi:10.1517/13543771003602012.
- [26] S. Sinha, M. Doble, S.L. Manju, 5-Lipoxygenase as a drug target: A review on trends in inhibitors structural design, SAR and mechanism based approach, *Bioorg. Med. Chem.* 27 (2019) 3745–3759. doi:<https://doi.org/10.1016/j.bmc.2019.06.040>.
- [27] Z.T. Gür, B. Çalışkan, E. Banoglu, Drug discovery approaches targeting 5-lipoxygenase-activating protein (FLAP) for inhibition of cellular leukotriene biosynthesis, *Eur. J. Med. Chem.* 153 (2018) 34–48. doi:<https://doi.org/10.1016/j.ejmech.2017.07.019>.

- [28] T. Sander, J. Freyss, M. Von Kor, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.* 55 (2015) 460–73. doi:10.1021/ci500588j.
- [29] H. van de Waterbeemd, Physicochemical concepts in drug design BT - Modern Methods of Drug Discovery, in: A. Hillisch, R. Hilgenfeld (Eds.), Birkhäuser Basel, Basel, 2003: pp. 243–257. doi:10.1007/978-3-0348-7997-2_12.
- [30] S.G. West, J.F. Finch, P.J. Curran, Structural equation models with nonnormal variables: Problems and remedies., in: *Struct. Equ. Model. Concepts, Issues, Appl.*, Sage Publications, Inc, Thousand Oaks, CA, US, 1995: pp. 56–75.
- [31] M. Ringnér, What is principal component analysis?, *Nat. Biotechnol.* 26 (2008) 303–304. doi:10.1038/nbt0308-303.
- [32] J. Braeken, M.A.L.M. van Assen, An empirical Kaiser criterion., *Psychol. Methods.* 22 (2017) 450–466. doi:10.1037/met0000074.
- [33] J.M. Blaney, E.J. Martin, Computational approaches for combinatorial library design and molecular diversity analysis, *Curr. Opin. Chem. Biol.* 1 (1997) 54–59. doi:https://doi.org/10.1016/S1367-5931(97)80108-1.
- [34] J. 1953- Willett, *Similarity and Clustering in Chemical Information Systems*, John Wiley & Sons, Inc., New York, NY, USA, 1987.
- [35] R.D. Brown, Y.C. Martin, Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection, *J. Chem. Inf. Comput. Sci.* 36 (1996) 572–584. doi:10.1021/ci9501047.
- [36] N. Nikolova, J. Jaworska, Approaches to Measure Chemical Similarity – a Review, *QSAR Comb. Sci.* 22 (2003) 1006–1026. doi:10.1002/qsar.200330831.
- [37] G.M. Maggiora, V. Shanmugasundaram, Molecular Similarity Measures, in: J. Bajorath (Ed.), *Chemoinformatics Concepts, Methods, Tools Drug Discov.*, Humana Press, Totowa, NJ, 2004: pp. 1–50. doi:10.1385/1-59259-802-1:001.
- [38] G.M. Maggiora, On Outliers and Activity Cliffs Why QSAR Often

- Disappoints, J. *Chem. Inf. Model.* 46 (2006) 1535. doi:10.1021/ci060117s.
- [39] J. Pérez-Villanueva, O. Méndez-Lucio, O. Soria-Arteche, T. Izquierdo, M. Concepción Lozada, W.A. Gloria-Greimel, J.L. Medina-Franco, Cyclic Systems Distribution Along Similarity Measures: Insights for an Application to Activity Landscape Modeling, *Mol. Inform.* 32 (2013) 179–190. doi:10.1002/minf.201200127.
- [40] M. Cruz-Montegudo, J.L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M.N.D.S. Cordeiro, F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?, *Drug Discov. Today*. 19 (2014) 1069–1080. doi:https://doi.org/10.1016/j.drudis.2014.02.003.

5

MODELING CoMFA BASED 3D-QSAR

5.1. Introduction

Chapter 4 shows the smooth regions of the SAR space that can be used for QSAR analysis. In this Chapter, we have tried to develop a few QSAR models by exploiting such SAR data. Three-dimensional quantitative structure-activity relationship (3D-QSAR) is a widely practiced ensemble technique used to explore quantitative relationships between the biological activities of compounds and their three-dimensional chemical structures. This method uses statistical techniques to predict the biological activity of known and unknown compounds and optimize new lead molecules [1]. Comparative Molecular Field Analysis (CoMFA) [2] is a promising new approach to 3D-QSAR. CoMFA models are constructed by using two field descriptors, such as steric and electrostatic fields. These fields provide all the information necessary to understand the relationship between the biological properties of a set of compounds with its 3D structural parameters via partial least square (PLS). The methodological overview of the CoMFA model is discussed in Chapter 2.

We have also discussed in Chapter 1 that the redox inhibitors are generally small lipophilic molecules, such as monocyclic and polycyclic aromatics and polyphenols [3], and they try to inhibit lipid peroxidation by scavenging peroxy free radicals and suppressing the formation of leukotrienes, thereby disrupting the inflammatory process [4]. The main drawback of redox inhibitors is that it facilitates the formation of methemoglobin due to non-specificity, but this can be prevented in some cases by utilizing natural antioxidants [5]. Natural antioxidants are generally not too redox-active and hence have no observable toxicities. Among different chemical classes of antioxidant (redox) 5-LOX inhibitors investigated, the flavones, chalcones, and quinones have been reported as an anti-inflammatory and anti-allergic agent. Redox potency of these molecules may directly link with hydrophobicity [6]. Research attempts have been made over the last few decades to develop molecular modeling and 3D QSAR analysis of 5-LOX redox inhibitors using alignment independent and dependent descriptors because the SAR and QSAR of 5-LOX redox inhibitors are challenging to develop and understand. So, in this Chapter, we have tried to develop statistically reliable predictive CoMFA models of flavones, chalcones and benzoquinones derivatives to understand the correlation between redox inhibitor's 3D structure and 5-LOX inhibitory potency.

5.2. CoMFA on 3', 4'-dihydroxyflavones as Rat 5-LOX Inhibitors

Flavonoids are a large group of secondary metabolic polyphenols which are widely distributed in dietary components like

fruits, vegetables, olive oil, tea, red wine, *etc.* They are well documented for their broad-spectrum pharmacological activities, including their potential role as anti-inflammatory agents. Several 3D-QSAR studies have examined the different structural features of flavonoid derivatives that contribute to biological activity [7,8]. Some natural flavonoids, as well as various synthetic derivatives, were identified as potent 5-LOX inhibitors. Also, some phenolic compounds were shown to inhibit both the cyclooxygenase and 5-LOX pathways [9,10]. Tokunaru Horie et al., suggest that the activity of the 3', 4'-dihydroxyflavones such as crisiliol was enhanced by modifying the oxygenated functions in the 2-phenyl ring with lipophilic alkyl groups [11]. This large and sterically crowded alkyl group might bring a hydrophobic nature to the parent flavones, thereby enhancing the activity. Thus, it is necessary to determine the quantitative influence of steric and electrostatic fields of 3', 4'-dihydroxyflavones on their 5-LOX inhibitory activity. In this scenario, this section proposes a CoMFA model based on the structure-activity relationship of 3', 4'-dihydroxyflavones studied by Tokunaru Horie et al.

5.2.1. Dataset of Flavone Derivatives

All 53 flavone derivatives that have been reported as inhibitors of 5-LOX (rat basophilic leukemia cells with arachidonic acid) were collected from literature [11]. The experimental half-maximal inhibitory concentration (IC₅₀) values of all compounds in micromolar (μM) were converted into pIC₅₀ by taking -Log (1/IC₅₀) and were used as the dependent variable. All the structures and associated inhibitory

activities are given in Table 1. A training set of 38 compounds (75%) for generating QSAR models and a test set of 13 compounds (25%) for validating the quality of the models were selected manually. For maintaining uniform distribution, molecules with low, moderate, and high activity were placed in both sets. Most active and least active molecules were retained in training set for better performance. Visual representation of the activity distribution of total, training, and test set are shown in Figure 5.1, and 2D structure of flavones core is displayed in Figure 5.2.

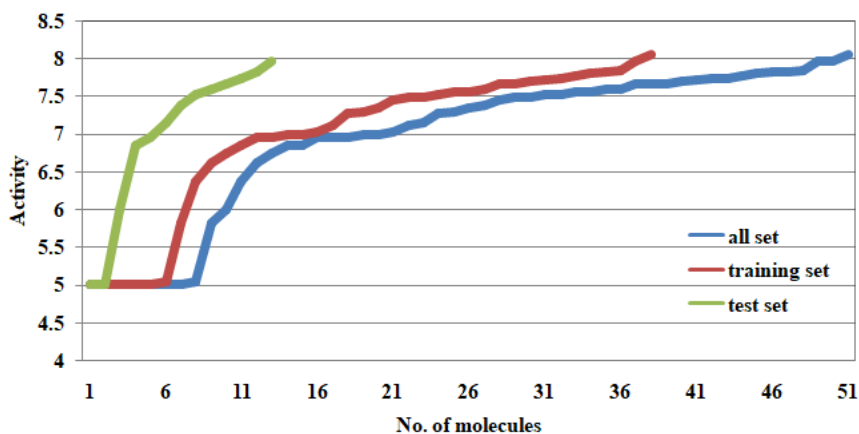


Fig. 5.1 Activity (pIC₅₀) distribution of the entire set, a training set, and a test set of flavone derivatives.

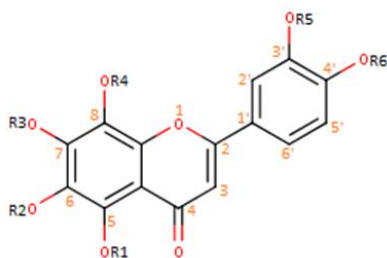


Fig. 5.2 The 2D chemical structure of the flavone core.

Table 5.1 Structural formulae of flavone derivatives and their IC₅₀ values

Compound	R1	R2	R3	R4	R5	R6	IC ₅₀ (μ M)
1	Me	Me	Me	-	H	H	240
2	Me	C ₄ H ₉	Me	-	H	H	52
3	Me	C ₅ H ₁₁	Me	-	H	H	35
4	Me	<i>i</i> - C ₅ H ₁₁	Me	-	H	H	26
5	Me	C ₆ H ₁₃	Me	-	H	H	14
6	Me	C ₈ H ₁₇	Me	-	H	H	22
7	Me	C ₁₀ H ₂₁	Me	-	H	H	17
8	Me	C ₁₂ H ₂₅	Me	-	H	H	30
9	Me	C ₁₄ H ₂₉	Me	-	H	H	70
10	Me	C ₁₆ H ₃₃	Me	-	H	H	1500
11	Me	C ₁₈ H ₃₇	Me	-	H	H	10000
12	C ₄ H ₉	Me	Me	-	H	H	45
13	C ₅ H ₁₁	Me	Me	-	H	H	32
14	<i>i</i> - C ₅ H ₁₁	Me	Me	-	H	H	16
15	C ₆ H ₁₃	Me	Me	-	H	H	15
16	C ₈ H ₁₇	Me	Me	-	H	H	19
17	C ₁₀ H ₂₁	Me	Me	-	H	H	18
18	C ₁₂ H ₂₅	Me	Me	-	H	H	18
19	C ₁₄ H ₂₉	Me	Me	-	H	H	28
20	C ₁₆ H ₃₃	Me	Me	-	H	H	140
21	C ₁₈ H ₃₇	Me	Me	-	H	H	9000
22	Me	Me	C ₁₂ H ₂₅	-	H	H	100
23	Me	-	Me	Me	H	H	430
24	C ₁₂ H ₂₅	-	Me	Me	H	H	22
25	Me	-	Me	C ₁₂ H ₂₅	H	H	33
26	Me	-	C ₁₂ H ₂₅	Me	H	H	110
27	Me	Me	Me	-	Ac	Ac	1000
28	Me	C ₄ H ₉	Me	-	Ac	Ac	100
29	Me	C ₅ H ₁₁	Me	-	Ac	Ac	28
30	Me	<i>i</i> - C ₅ H ₁₁	Me	-	Ac	Ac	Nd
31	Me	C ₆ H ₁₃	Me	-	Ac	Ac	50
32	Me	C ₈ H ₁₇	Me	-	Ac	Ac	42
33	Me	C ₁₀ H ₂₁	Me	-	Ac	Ac	30
34	Me	C ₁₂ H ₂₅	Me	-	Ac	Ac	75
35	Me	C ₁₄ H ₂₉	Me	-	Ac	Ac	10000
36	Me	C ₁₆ H ₃₃	Me	-	Ac	Ac	10000

37	Me	C ₁₈ H ₃₇	Me	-	Ac	Ac	10000
38	Me	Me	C ₁₂ H ₂₅	-	Ac	Ac	Nd
39	H	Me	Me	-	H	H	95
40	H	C ₄ H ₉	Me	-	H	H	15
41	H	C ₅ H ₁₁	Me	-	H	H	22
42	H	<i>i</i> - C ₅ H ₁₁	Me	-	H	H	9
43	H	C ₆ H ₁₃	Me	-	H	H	20
44	H	C ₈ H ₁₇	Me	-	H	H	11
45	H	C ₁₀ H ₂₁	Me	-	H	H	11
46	H	C ₁₂ H ₂₅	Me	-	H	H	26
47	H	C ₁₄ H ₂₉	Me	-	H	H	10000
48	H	C ₁₆ H ₃₃	Me	-	H	H	10000
49	H	C ₁₈ H ₃₇	Me	-	H	H	10000
50	H	Me	C ₁₂ H ₂₅	-	H	H	140
51	H	-	Me	Me	H	H	110
52	H	-	Me	C ₁₂ H ₂₅	H	H	180
53	H	-	C ₁₂ H ₂₅	Me	H	H	110

5.2.2. Molecular Modeling and Alignment of Flavones

All flavone derivatives that are in SMILES notation were downloaded from the ChEMBL database and converted to gjf format using Open Babel utility [12]. Gas-phase geometries were optimized using Density Functional Theory (DFT) [13] of Becke's three parameter hybrid exchange-correlation functional (B3LYP) [14,15] employing 6-31G (d,p) basis set using Gaussian 09 software package [16] and the lowest energy conformer was further converted to SDF file. The alignment procedure was executed by using all available molecules as possible templates. Hence, 53 alignments were produced, each obtained by superimposition on the corresponding template molecule. The alignment corresponding to the template of the most active compound was selected for further analysis and is shown in Figure 5. 3.

PyMOL For evaluation only.
Contact sales@deisici.com.

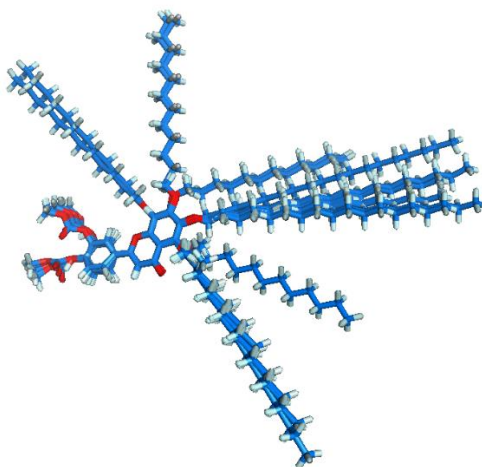


Fig. 5.3 Alignment of 53 flavone derivatives

5.2.3. Statistical Analysis of CoMFA Models of Flavones

Before going to build the CoMFA model, the training set was checked for spotting outliers and anomalies. This job was done by calculating the Tanimoto coefficient between compounds in the training set and then plotting it versus the experimental response (Figure 5.4.). For this purpose, the structural similarity was calculated by generating FP2 based 2D fingerprints for each compound. FP2 is a fingerprint-based on the path that indexes small fragments of the molecule based on linear segments of up to 7 atoms. It is a bit like the Daylight fingerprints. The resultant plot indicated that there were no outliers in the training set. All compounds show similarity in structure and activity.

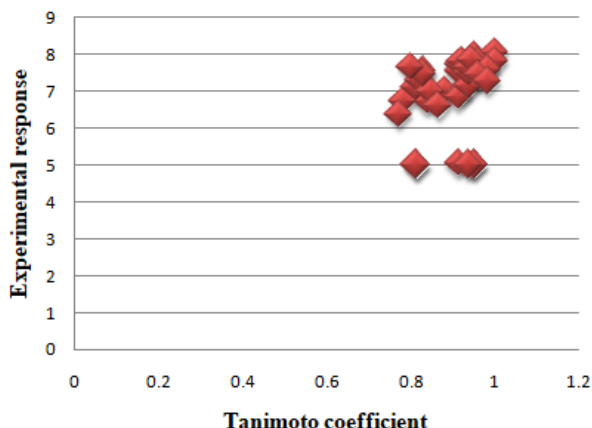


Fig. 5.4 Graph showing Tanimoto coefficients between the most active compound 42 and remaining compounds in the training set of flavone derivatives based on FP2 fingerprint vs. experimental responses (pIC₅₀).

Based on the effect of component numbers on the square correlation coefficient for LOO, LTO, and LMO cross-validation tests, the CoMFA model with five PLS components was constructed using this training set. Statistical quality parameters associated with the CoMFA model are listed in Table 5.2. The analysis of these parameters revealed that the best CoMFA model was obtained with a combination of steric and electrostatic fields. It also found that more than 91% of contributions from the steric field were observed for the creation of the CoMFA model, while only 8% contribution was observed from the electrostatic field. This result indicates that steric interactions are more relevant to bind flavone analogs to 5-LOX. The model shows a satisfactory cross-validated correlation coefficient Q^2 for LOO, LTO, and LMO as 0.6587, 0.6479, and 0.5547, respectively, indicating an excellent internal prediction of the model. The R^2 , SDEC, and F test values for the PLSR model were found to be 0.9320, 0.2460, and

87.7404 respectively and are reasonable. The predictive ability of that same model was once again assessed using a collection of 13 test compounds that are not included in the model generation.

The R^2_{pred} value and SDEP of test sets are 0.8259 and 0.4292, respectively, which points out that the CoMFA model is reliable for external predictions, and could be used in designing new inhibitors. The values of experimental and predicted activities, along with the residual values of the training set and test set molecules, are summarized in Table 5. 3. and 5. 4. respectively. The scatter plot of observed vs. predicted values of pIC_{50} of both the training and test set are shown in Figure 5.5. This data shows that the experimental and the predicted activities of inhibitors are very close to each other. Most of the molecules show residual values less than 0.4. These findings again indicate the excellent predictive power of the established model.

Table 5.2 Statistical quality parameters associated with the CoMFA model of flavone derivatives

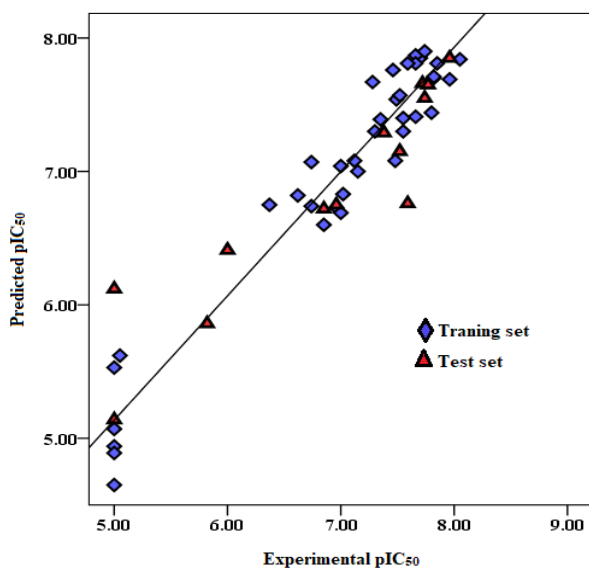
Statistical parameter	CoMFA
R^2	0.9320
$Q^2(\text{LOO})$	0.6587
$Q^2(\text{LTO})$	0.6479
$Q^2(\text{LMO})$	0.5547
R^2_{pred}	0.8259
F	87.740
SDEC	0.2460
SDEP	0.4292
Steric contribution	0.9136
Electrostatic contribution	0.0860
Component	5

Table 5.3 The experimental and predicted pIC₅₀ values of the training set of flavone derivatives

Compound	Experimental pIC ₅₀	Predicted pIC ₅₀	Residual
1	6.62	6.82	-0.20
2	7.28	7.67	-0.39
3	7.46	7.76	-0.30
4	7.59	7.81	-0.22
5	7.85	7.81	0.04
6	7.66	7.81	-0.15
8	7.52	7.57	-0.05
9	7.15	7.00	0.15
11	5.00	4.89	0.11
12	7.35	7.39	-0.04
13	7.49	7.54	-0.05
14	7.80	7.44	0.36
15	7.82	7.71	0.11
17	7.74	7.90	-0.16
19	7.55	7.30	0.25
20	6.85	6.60	0.25
21	5.05	5.62	-0.57
22	7.00	6.69	0.31
23	6.37	6.75	-0.38
24	7.66	7.41	0.25
25	7.48	7.08	0.40
28	7.00	7.04	-0.04
29	7.55	7.40	0.15
31	7.30	7.30	0.00
34	7.12	7.08	0.04
36	5.00	4.94	0.06
37	5.00	5.07	-0.07
39	7.02	6.83	0.19
40	7.82	7.70	0.12
41	7.66	7.87	-0.21
42	8.05	7.84	0.21
43	7.70	7.85	-0.15
45	7.96	7.69	0.27
47	5.00	5.53	-0.53
49	5.00	4.65	0.35
51	6.74	6.74	0.00
52	6.74	7.07	-0.33
53	6.96	6.73	0.23

Table 5.4 The experimental and predicted pIC_{50} values of the test set of flavone derivatives

Compound	Experimental pIC_{50}	Predicted pIC_{50}	Residual
7	7.77	7.65	0.12
10	5.82	5.86	-0.04
16	7.72	7.67	0.05
18	7.74	7.55	0.19
26	6.96	6.75	0.21
27	6.00	6.41	-0.41
32	7.38	7.29	0.09
33	7.52	7.15	0.39
35	5.00	6.12	-1.12
44	7.96	7.85	0.11
46	7.59	6.76	0.83
48	5.00	5.14	-0.14
50	6.85	6.72	0.13

**Fig. 5.5** Activity plots of observed vs. predicted pIC_{50} of training and test set of flavones.

The progressive Y-scrambling technique analyzed the stability of the CoMFA model. The results of the Y-scrambling test give a fitted Q^2 value of 0.137 for the model. In all cases, the obtained random models have much lower prediction accuracies than the model based on the real data, indicating no luck factor involved in the development of the CoMFA model.

5.2.4. Graphical Interpretation of the CoMFA Contour Maps of Flavones

The most significant advantage of CoMFA is that it generates 3D contour plots around the molecules. These contour maps help to identify important regions where any change in the steric and electrostatic field might affect the biological activity, and they also provide hints for the modification required to design new molecules with better activity. The CoMFA steric and electrostatic contour maps are shown in Figure. 5. 6. Green and yellow contours represent the steric fields. The green regions in the steric contour maps indicate the area where the bulky groups are favored for activity while the yellow contours represent the region where the bulky groups are not favored for the activity. Electrostatic contour maps are represented by the red and cyan contour. The cyan contour defines a region of space where positively charged substituent (electron-deficient group) increases activity, whereas the red contour defines a region of space where negatively charged substituent (electron-rich group) increases activity.

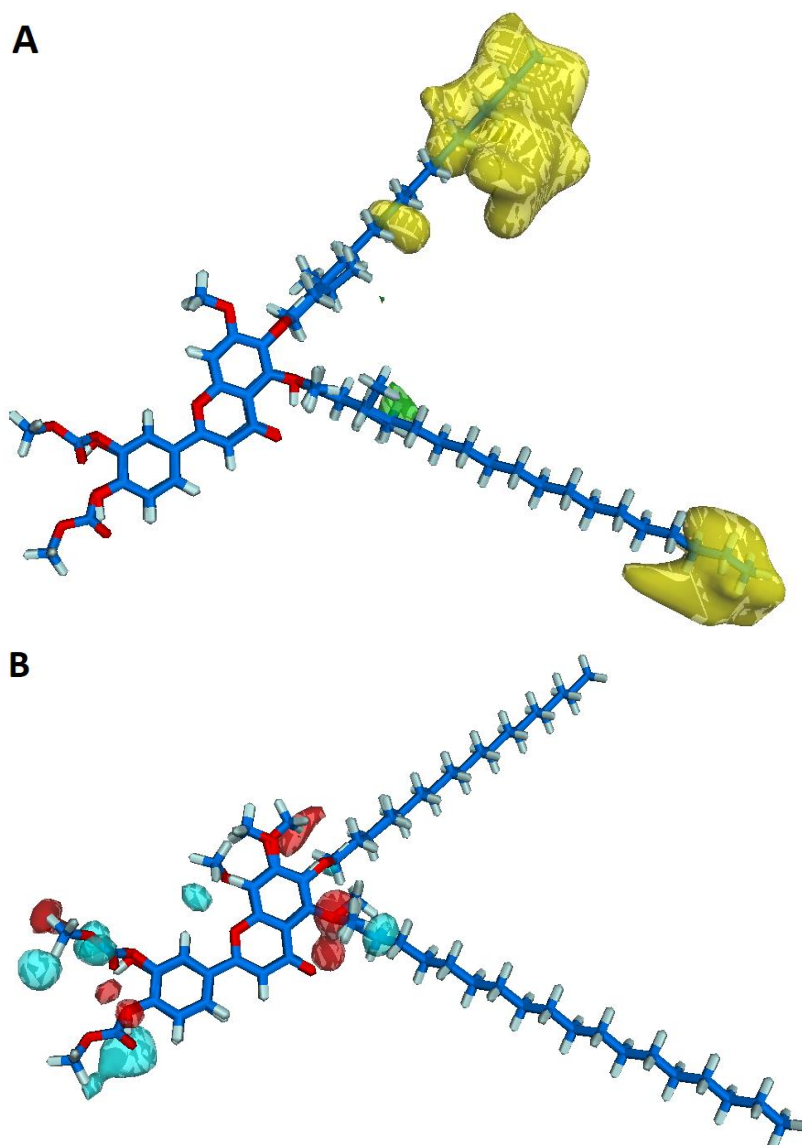


Fig. 5.6 PLS contours from 3D-QSAR models for 5-LOX inhibitors. (A) CoMFA steric contour maps, (B) CoMFA electrostatic contour maps.

These contour maps give us some general insight into the nature of the receptor-ligand binding region. A green plot was found

near the middle of the R1 group of the 5th carbon atom, and a tiny green portion was also seen near to the R2 group of the 6th carbon atom. Both of these indicate bulky groups in this region and are in favor of increasing the ligand's 5-LOX inhibitory activity. To justify this, we could say that the activities of the compounds 40 and 12 with isopentoxy substituent attached to the 5th and 6th carbon atom of flavones respectively are higher than those of the compounds 41 and 13 with pentoxy substituent. Thus, the presence of a bulky group in these regions is significant for a potentially active ligand. The broad yellow contours at the tail portions of the same alkyl residues (R1 and R2) shows that too long alkyl chain has an impeding influence on activity. These findings are confirmed by the lower activity of compounds like 10, 11, 21, 35, 36, 37, 47, 48, 49, *etc.*, which contain long alkyl chain at positions R1/R2. Hence the alkoxy substituent (OR1 and OR2) contain chain length greater than 6 or 5 at the 5th and 6th carbon atom of the parent flavones disfavors the inhibition of 5-LOX activity, i.e., the compounds with shorter (less than 5) or longer alkyl (greater than 10) groups at R1, and R2 positions are found to have lower activity.

In an electrostatic field, red and cyan contours are mostly distributed on the core of the flavone. So, the electronegative and electropositive substituents in these regions are likely to boost biological activity. The red contours surrounded the oxygen of alkoxy substituent of 5th, and 7th carbon residue suggests that the electron-donating substituents in this region are likely to enhance biological activity. It indicates -OR group is preferable at this position than -OH

group. Electrostatic contours in the 2-phenyl part of the flavones have mixed red and cyan shades, and it indicates both the electropositive and electronegative group promotes the activity. However, the experimental activity indicates that the -OH group is more favorable than the -OAc group. This is the reason why the compounds 1 to 10 has higher activity than the compounds 27 to 37. The cyan electrostatic contour near the R1 substituent of the 5th carbon indicates that the presence of the electron-withdrawing group is favorable at this position. This observation has been confirmed by the higher activity of compounds 39 - 46 with -OH at the said position than compounds 27 - 36 with -OCH₃ at the same position. The reason is that both -OH and -OCH₃ are electron-donating, but -OH is less electron donating than that of -OCH₃. The same result can be seen in R2 and R4 substituents of the flavone.

5.2.5. Docking Analyses of Flavones

Molecular docking was used to clarify the binding mode between flavon derivatives and the binding site of 5-LOX and to examine the stability and rationality of the CoMFA model. It provides straightforward knowledge for further structural optimization. We have developed a theoretical 3D model of rat 5-LOX by homology modeling and explained it in Chapter 3. This homologous model of rat 5-LOX built from the crystal structure of human 5-LOX crystal was used for the docking study. Autodock Vina [17] software was used to carry out molecular docking analysis. A grid box of $42 \times 35 \times 35 \text{ \AA}^0$ dimension with a spacing 1 \AA^0 was built around the protein-ligand complex and was centered at 3.7, 25, 3.5 for x, y, and z, respectively. The binding region selected for the docking studies containing a list of

The 2D binding interaction of the most active (compound 42) and the least active molecule (compound 49) to the rat 5-LOX model binding pocket is shown in Figure 5.7A and B, respectively. The binding affinity of compound 42 is much higher than the binding affinity of compound 49. Compound 42 forms a strong hydrogen bond with Tyr 365 and can be seen in the yellow dotted line. The -OH group of a 2-phenyl part in the polar head of compound 49 forms a hydrogen bond with His 601. Figure 5.7C displays the image of the binding pocket on the surface of the protein molecule, which reveals that the molecule was well embedded in the active site pocket. The compound 42 occupied this deep groove and made strong van der Waal interactions with the neighboring residues. Trp 600, Ala 604, Tyr 182, Phe 178, Leu 608, Ile 674, Ile 407, Asn 408, His 373, Leu 369, His 388, Ala 411, Leu 415, Gln 364, Phe 422, Asn 426 and His 601 were found to be the most crucial residues of the deep groove, most of them are hydrophobic amino acids, suggesting that hydrophobic interactions play a significant role in modulating the 5-LOX inhibitor. It can be observed that all these interacting residues lie within the range of $<1 \text{ \AA}$. The docking study further depicts that the tail portion of the compound 42 (alkyl side chain of the 6th carbon atom) interacts with less crowded amino acids Ala 411 Asn 408 and Ile 407 of the target protein. This observation is compatible with CoMFA green contours found around R1 and R2 group of the 5th and 6th carbon atom, bulky groups in these regions were favorable may be due to its interaction with less crowded amino acids of the target protein. The yellow contours in CoMFA suggested that more extended alkyl groups at R1 and R2 positions

reduce the activity. Docking result also agrees with this observation, i.e., a compound with a long alkyl chain (compound 49) cannot occupy the active site of the protein much effectively as compared to compound 42 (Figure 5.7D). This observation indicates the steric contours complemented very well with the docking results. The polar head, -OH substituted phenyl group of flavones, is located in the small polar hydrophilic binding cleft lined with His 601, Asn 426, and Asp 458, which were observed to form indirect nonbonding interactions with the substrate. CoMFA electrostatic contours further support this result. Red and blue shades are observed mainly in this region, indicating the possibility of polar interaction with the target protein. The phenylalanine residues, Phe 422, and the tryptophan residue Trp 600 are also found to be located in the same pocket. The overall result shows that molecular docking interaction coincides very well with the CoMFA contour.

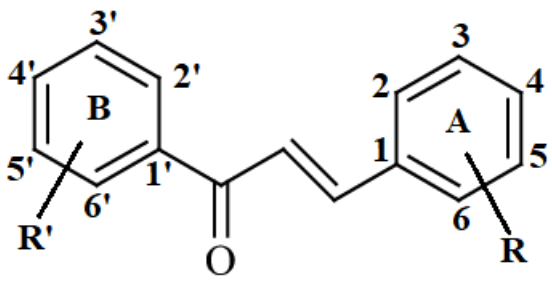
5.3. CoMFA on 3, 4-dihydroxychalcones as Rat 5-LOX Inhibitors

Chalcones, 1, 3-diphenyl-2-propen-1-one, are open-chain flavonoids consist of two phenolic rings (A and B rings) connected by a three-carbon bridge. Chalcone is a privileged scaffold in medicinal chemistry, which is extensively used as an effective template for drug discovery. On account of their rapid and efficient metabolism after systemic administration, the chalcones have been identified as promising nontoxic topical anti-inflammatory agents. To understand and predict the modes of action of chalcones and to gain an insight into the essential structural and physicochemical requirements for the

design of novel chalcones based 5-LOX inhibitors, we must discover the complex relationships hidden in experimental data. CoMFA provides an excellent platform for this purpose. In this study, we have used CoMFA descriptors to gain insight into the steric, the electrostatic properties of these molecules, and their influence on the activity. We are also given particular attention to data noise reduction techniques since the predictive power of the QSAR model is mostly depends on variable reduction.

5.3.1. Dataset of Chalcone Derivatives

All 53 chalcone derivatives and their biological activity data used in this study were collected from the literature [21]. The experimental IC₅₀ values of all compounds in μM (micromole) were converted into pIC₅₀ by taking $-\text{Log}(1/\text{IC}_{50})$ and were used as the dependent variable. All the structures and associated inhibitory activities were listed in Table 5.5. The dataset covered a wide range of pIC₅₀ values, spanning from 4.00 up to 8.62 log units. The dataset was divided into two subsets: a training set of 36 compounds (75%) for generating QSAR models and a test set of 12 compounds (25%) for validating the quality of the models. Five compounds (1, 2, 3, 4, and 18) with activity value higher than 100000nM were removed. For maintaining uniform distribution, molecules with low, moderate, and high activity were placed in both sets. Most active and least active molecules were retained in training set for better performance.

Table 5.5 Structural formulae of compounds and their IC₅₀ values


Compound	R'	R	IC ₅₀ (nM)
1	2'-OH	3-H, 4-H	230000
2	4'-OH	3-H, 4-H	400000
3	2',4'-OH	3-H, 4-H	140000
4	2',4',6'-OH	3-H, 4-H	42000
5	2'-OH	3-H, 4-OH	35000
6	2',4'-OH	3-H, 4-OH	100000
7	2',4',6'-OH	3-H, 4-OH	43
8	-	3,4-OH	23
9	2'-OH	3,4-OH	4.2
10	3'-OH	3,4-OH	4
11	4'-OH	3,4-OH	4.6
12	2',4'-OH	3,4-OH	140
13	2',4',6'-OH	3,4-OH	22
14	2-thienyl	3,4-OH	210
15	3-pyridyl	3,4-OH	17000
16	2'-OH	3-OCH ₃ , 4-OH	8900
17	4'-Cl	3-OCH ₃ , 4-OH	12000
18	4'-OCH ₃	3-OCH ₃ , 4-OH	-
19	2'-OH	3-OH, 4-OCH ₃	92
20	2'-Cl	3,4-OH	8.5
21	4'-Cl	3,4-OH	23
22	4'-NO ₂	3,4-OH	58
23	2'-CF ₃	3,4-OH	27
24	3'-CH ₃	3,4-OH	76
25	4'-CH ₃	3,4-OH	27
26	2'-OCH ₃	3,4-OH	6.5
27	4'-OCH ₃	3,4-OH	20
29	3'-N(CH ₃) ₂	3,4-OH	9.8
30	4'-N(CH ₃) ₂	3,4-OH	4.7
31	4'-OCH(CH ₃) ₂	3,4-OH	41

32	2'-OH, 4'-OCH ₃	3,4-OH	15
33	2'-OH, 5'-OCH ₃	3,4-OH	41
34	4'-OH, 3'-OCH ₃	3,4-OH	9
35	2'-CH ₃ , 4'-CH ₃	3,4-OH	17
36	2'-OCH ₃ , 4'-OCH ₃	3,4-OH	10
37	2'-OCH ₃ , 5'-OCH ₃	3,4-OH	7.8
38	2'-OCH ₃ , 6'-OCH ₃	3,4-OH	370
39	3'-OCH ₃ , 4'-OCH ₃	3,4-OH	18
40	2'-CH ₃ ,4'-CH ₃ 6'-CH ₃	3,4-OH	400
41	3'-OCH ₃ ,4'-OCH ₃ 5'-OCH ₃	3,4-OH	16
42	2'-OCH ₃ , 5'-OCH ₃	3,4-OH	64
43	2'-OH, 5'-OH	3,4-OH	39
44	2'-OH, 5'-OC ₂ H ₅	3,4-OH	5.3
45	2'-OH, 5'-CH(CH ₃) ₂	3,4-OH	4
46	2'-OH, 5'-OCH(CH ₃) ₂	3,4-OH	11
47	2'-OH, 5'-OC ₄ H ₉	3,4-OH	1000
48	2'-CH ₃ , 5'-CH ₃	3,4-OH	16
49	2'-OCH ₃ , 5'-CH ₃	3,4-OH	24
50	2'-OCH ₃ , 5'-OC ₂ H ₅	3,4-OH	3.8
51	2'-OCH ₃ , 5'-OCH(CH ₃) ₂	3,4-OH	14
52	2'-OC ₂ H ₅ , 5'-OCH ₃	3,4-OH	27
53	2'-OC ₂ H ₅ , 5'- OC ₂ H ₅	3,4-OH	2.4

5.3.2. Molecular Modeling and Alignment of Chalcones

Molecular modeling and geometry optimization of chalcones are performed in the same way as with flavone derivatives. The alignment procedure was executed by using all available molecules as possible templates. Fifty-three alignments were produced, each of them was obtained by superimposition on the corresponding template molecule. The alignment corresponding to the highest cumulative O3A score was selected for further analysis. Figure 5.8 shows the best alignment in which compound 3 was selected as a template.

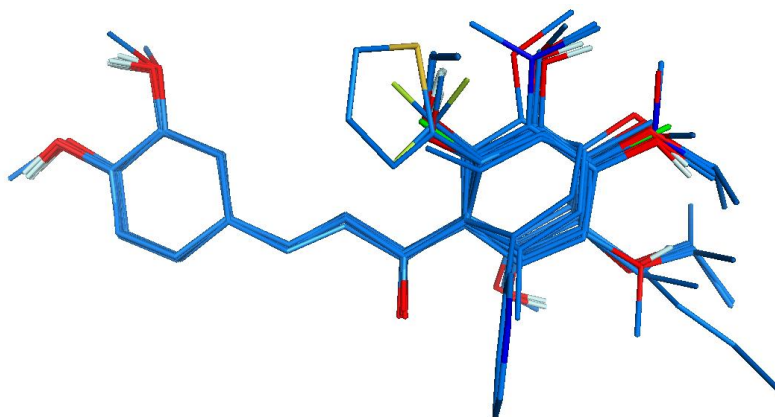


Fig. 5.8 Alignment of 53 chalcone derivatives.

5.3.3. Statistical Analysis of CoMFA Model of Chalcones

The result of the Tanimoto similarity analysis (Figure 5.9.) shows that the training set did not contain outliers and anomalies and is perfect for building the CoMFA model.

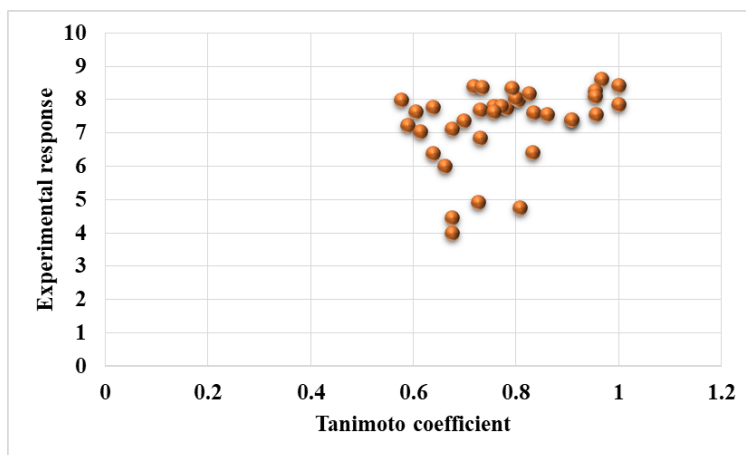


Fig. 5.9 Graph showing Tanimoto coefficients between reference compound 51 and remaining compounds in the training set of chalcones based on FP2 fingerprint vs. experimental responses (pIC_{50}).

Using this training set CoMFA model with five PLS components was built. In this study, we achieved a comparable result when applying both SRD/FFD and UVE/IVE for the variable reduction procedure. So, statistical quality parameters associated with CoMFA models based on descriptors obtained by applying both SRD/FFD and UVE/IVE procedures for noise reduction in the input data are given in Table 5.6. The analysis of these parameters revealed that the best CoMFA model was obtained with a combination of steric and electrostatic fields. The almost equal contribution was observed from the steric and electrostatic field, indicating that both these two interactions are essential to the binding of chalcones analogs with 5-LOX. According to statistical criteria given in Tables 5.6, even though both variable reductions methods are providing good result, it is confirmed that most statistically significant results were provided one with UVE/IVE procedures rather than SRD/FFD procedure. CoMFA-UVE model gave good cross-validated correlation coefficient Q^2 for LOO, LTO, and LMO as 0.7365, 0.7298, and 0.6877, respectively, indicating an excellent internal predictive power of the model. The Q^2 for values for CoMFA-FFD such as LOO, LTO, and LMO as 0.7365, 0.7298, and 0.6877 respectively were also reasonable. Even though the R^2 (0.9731), SDEC (0.1873) and F test value (216.84) for CoMFA-UVE model were found to better than that of R^2 (0.9660), SDEC (0.2106) and F test value (170.29) for CoMFA-FFD model, both datasets showed the comparable external predictivity (R^2_{pred} for CoMFA-FFD = 0.8089 and CoMFA-UVE = 0.8084). This observation indicates, both SRD/FFD and UVE/IVE could be very much suitable procedures for removing sufficient information from the input data

matrix; however, CoMFA-UVE could be very useful for designing new inhibitors.

Table 5.6 Statistical quality parameters associated with CoMFA models of chalcones

Statistical parameter	CoMFA -FFD	CoMFA -UVE
R ²	0.9660	0.9731
Q ² (LOO)	0.6936	0.7365
Q ² (LTO)	0.6846	0.7298
Q ² (LMO)	0.6406	0.6877
R ² _{pred}	0.8089	0.8084
F	170.29	216.84
SDEC	0.2106	0.1873
SDEP	0.5317	0.5324
Steric contribution	0.5384	0.5234
Electrostatic contribution	0.4616	0.4766
Component	5.0000	5.0000

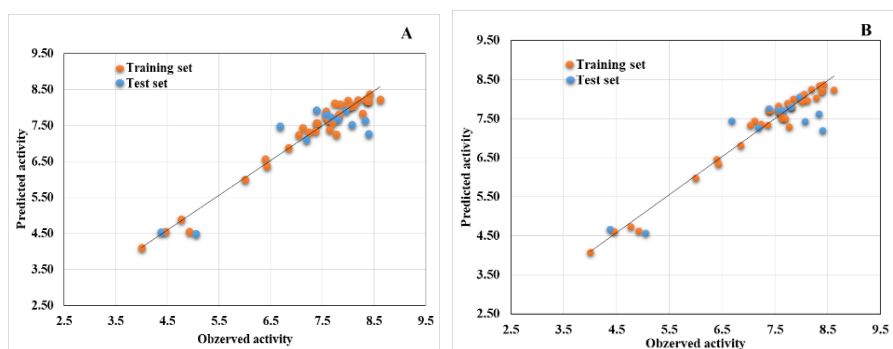
The values of experimental and predicted activities, along with the residual values of the training set and test set molecules, are summarized in Tables 5.7 and 5.8, respectively. The scatter plot of observed vs. predicted values of pIC₅₀ of both the training and test set are shown in Figure 5.10. These data show that the experimental and the predicted activities of inhibitors are very close to each other. Most of the molecules show residual values less than 0.4. This result again indicates the excellent predictive power of the established models. The validity of the CoMFA models was also analyzed by progressive Y-scrambling. The results of the Y-scrambling test are shown that fitted Q² for the CoMFA -FFD and CoMFA -UVE models is around 0.387 and 0.291, respectively. In all cases, the obtained random models have much lower prediction accuracies than the model based on the real data, indicating no perceptible chance of correlation in the CoMFA model.

Table 5.7 The experimental and predicted pIC₅₀ values of training set by SRD/FFD and UVE/IVE CoMFA models of chalcones

Compound	Experimental pIC ₅₀	CoMFA -FFD		CoMFA -UVE	
		Predicted pIC ₅₀	Residual	Predicted pIC ₅₀	Residual
6	4.46	4.55	-0.09	4.62	-0.16
7	4.00	4.10	-0.10	4.09	-0.09
8	7.37	7.34	0.03	7.34	0.03
9	7.64	7.38	0.26	7.50	0.14
10	8.38	8.19	0.19	8.19	0.19
12	8.34	8.23	0.11	8.35	-0.01
13	6.85	6.88	-0.03	6.81	0.04
16	4.77	4.89	-0.12	4.74	0.03
18	4.92	4.55	0.37	4.64	0.28
20	7.04	7.24	-0.20	7.33	-0.29
22	7.64	7.67	-0.03	7.56	0.08
23	7.24	7.31	-0.07	7.37	-0.13
25	7.12	7.44	-0.32	7.44	-0.32
26	7.57	7.75	-0.18	7.69	-0.12
27	8.19	8.22	-0.03	8.25	-0.06
28	7.70	7.58	0.12	7.50	0.20
29	8.01	8.03	-0.02	7.94	0.07
32	7.82	7.81	0.01	7.80	0.02
33	7.39	7.57	-0.18	7.71	-0.32
34	8.05	8.03	0.02	8.13	-0.08
35	7.77	7.25	0.52	7.29	0.48
36	8.00	8.20	-0.20	8.07	-0.07
37	8.11	8.06	0.05	7.97	0.14
38	6.43	6.38	0.05	6.33	0.10
39	7.74	8.13	-0.39	7.91	-0.17
40	6.40	6.57	-0.17	6.47	-0.07
41	7.80	7.75	0.05	7.83	-0.03
43	7.41	7.57	-0.16	7.71	-0.30
44	8.28	7.85	0.43	8.04	0.24
45	8.40	8.18	0.22	8.24	0.16
47	6.00	6.00	0.00	5.98	0.02
49	7.62	7.64	-0.02	7.62	0.00
50	8.42	8.38	0.04	8.37	0.05
51	7.85	8.09	-0.24	8.01	-0.16
52	7.57	7.89	-0.32	7.83	-0.26
53	8.62	8.22	0.40	8.25	0.37

Table 5.8 The experimental and predicted pIC_{50} values of test set by SRD/FFD and UVE/IVE CoMFA models of chalcones

Compound	Experimental pIC_{50}	CoMFA -FFD		CoMFA -UVE	
		Predicted	Residual	Predicted	Residual
11	8.40	7.27	1.13	7.21	1.19
14	7.66	7.73	-0.07	7.70	-0.04
15	6.68	7.47	-0.79	7.45	-0.77
17	5.05	4.50	0.55	4.58	0.47
21	8.07	7.53	0.54	7.44	0.63
24	7.57	7.81	-0.24	7.72	-0.15
30	8.33	7.64	0.69	7.63	0.70
31	7.39	7.93	-0.54	7.77	-0.38
42	7.19	7.11	0.08	7.27	-0.08
46	7.96	7.92	0.04	8.05	-0.09
48	7.80	7.68	0.12	7.78	0.02

**Fig. 5.10** Activity plots of observed vs. predicted pIC_{50} of training and test set of chalcones resulting from PLS (PC = 5) of SRD/FFD (A) and UVE/IVE (B) adjusted datasets.

5.3.3. Graphical Interpretation of the CoMFA Contour Maps of Chalcones

The contour maps are used to identify regions in MIFs of the molecules included in the training set where any change in the steric and electrostatic field might affect the biological activity, and they also provide hints for the modification required to design new molecules with better activity. The CoMFA steric and electrostatic contour maps are shown in Figure 5. 11. Green and yellow contours represent the steric fields while the red and cyan contours represent electrostatic contour maps as same as that of contours of flavones.

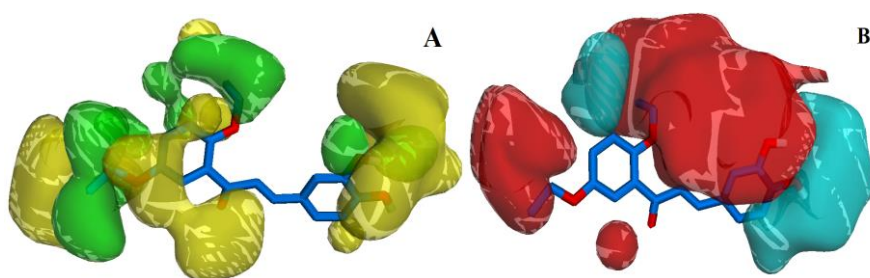


Fig. 5.11 PLS contours from 3D-QSAR models for 5-LOX chalcone inhibitors. (A) CoMFA steric contour maps, (B) CoMFA electrostatic contour maps.

These contour maps provide us with an overall idea about the nature of the receptor-ligand binding region. A large green contour is located around the end of the $-OC_2H_5$ group on the 2nd position of B ring indicate bulky groups in this region are favorable to increase the activity of the ligand. To justify this, we could say that the activities of the compound 53 with $-OC_2H_5$ substituent attached to the 2nd carbon atom of chalcone is higher than those of the compounds 42-52 with

either -OH or -OCH₃ substituent. However, a small yellow contour just behind the above-mentioned green contour indicates that too bulky a group has an impeding influence on activity. This result is confirmed by the lower activity of compounds 14, 20, and 23, which contain 2-thienyl, -Cl and -CF₃ group respectively on the 2nd position. Similar observation also found in the 5th position of the B ring, where a green plot was found around the middle of the -OC₂H₅ group, and a large yellow contour just behind this yellow contour suggests that groups with moderate steric tolerance are required at this position to increase the activity. This may be the reason why compounds 53, 50, 51, 45 and 46 with a moderate bulky group like -OC₂H₅, -OC₂H₅, -CH(CH₃)₂, -CH(CH₃)₂, -OCH(CH₃)₂ respectively are more potent than molecules with either smaller substituents (like -OH, -CH₃, OCH₃) such as compounds 33, 37, 41, 42, 43, 48, 49 and 52 or larger substituent (-OC₄H₉) like compound 47. A long yellow contour started from -C=O group of carbon bridge was found spread to the plane of the substituted B ring of compound indicating that steric crowdedness in these regions is disfavoring the inhibition of 5-LOX activity. Also, a large yellow contour surrounded the 3-OH and 4-OH group of ring A, indicating a less bulky group in these regions is very important for a potentially active ligand. This observation leads to the conclusion that OH is better substituent in the 3rd and 4th positions than OCH₃. However, a green contour was found overlapping -OH substitution at 3rd position indicating the possibility of the addition of bulky groups may increase high binding affinity.

In an electrostatic field, red and cyan contours are distributed on the entire surface of the chalcone. So, electronegative and

electropositive substituents have a significant role in boosting the biological activity of chalcones. The red contours surrounded the substituent of the 5th and 6th position of the B ring suggest that the electron-donating substituents in this region are likely to enhance biological activity. This is in agreement experimental result, compounds (53, 50, 51, 45 and 46) with electron-rich alkoxy substituent at 5th position having a greater binding affinity towards rat 5-LOX than the compound with -OH substituent at 5th position and compounds 38 with -OCH₃ at 6th position having a greater binding affinity towards rat 5-LOX than compound 7 with -OH substituent at 6th position. The electron-donating substituent favorable large red contour spread on the A ring, 3C Bridge, and 2'-position of the B ring indicated that electron-rich groups around these areas increased the 5-LOX activity of chalcones. Compounds 8-53 with an electronegative substituent (OH) at 3rd, 4th positions showed higher activity than compounds 1-7 (no substituent on the same position). Similar explanations are for the importance of electronegative group at 2', here compounds (20, 23) with electron-withdrawing groups like -Cl and CF₃ having IC₅₀ value (>50 nM) was higher than the IC₅₀ value (=2.4 nM) of 53, implying the importance of electron-donating group at 2' position of B ring. The cyan electrostatic contour near the 3rd and 4th positions of B and A ring respectively indicated that the presence of an electron-withdrawing group is favorable at this position.

5.3.4. Docking Analyses of Chalcones

Molecular docking provides the nature of binding mode between chalcone derivatives and the amino acids at the active site of 5-LOX. Furthermore, it is a way to examine the stability and

rationality of the CoMFA model. The 2D images of binding interaction of most active (compound 53), and least active (compound 2) ligand with rat 5-LOX binding pocket are shown in Figure 5.12 A and B, respectively.

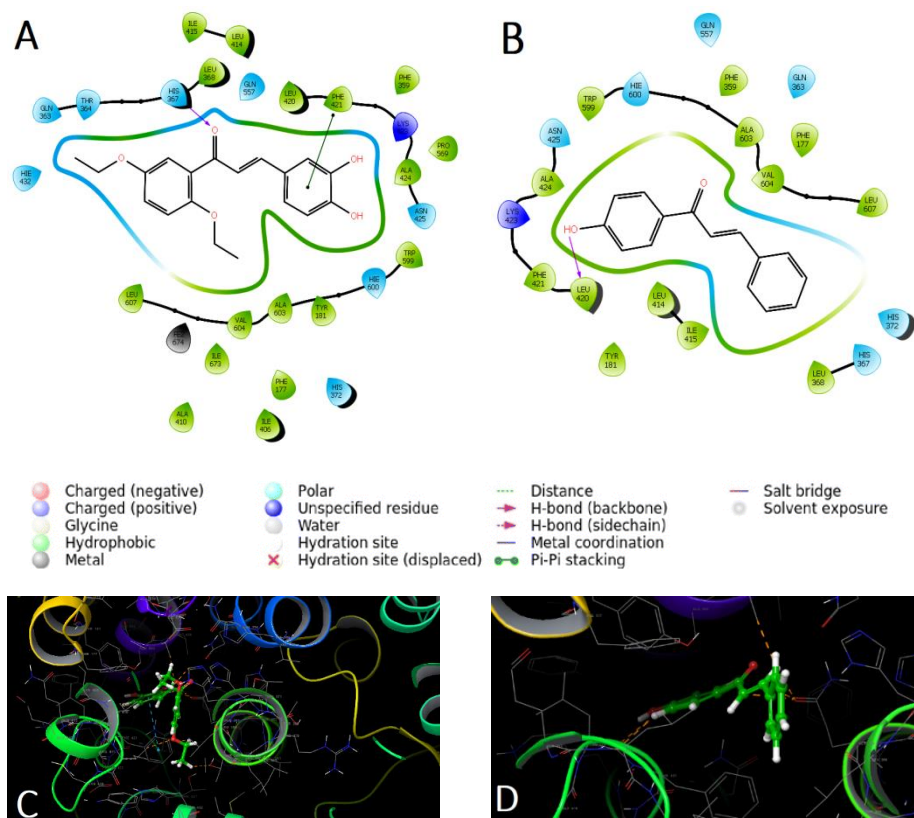


Fig. 5.12 Binding interaction flavones with 5-LOX model: (A) and (B) are respectively are the 2D Image of the binding interaction of the most active compound 53 and least active compound 2 with the amino acid at the active site of 5-LOX model. Figure (C) and (D) are respectively are the 3D Image of the binding interaction of the most active compound 53 and the least active compound 2 with the amino acid at the active site of the 5-LOX model.

The nature of binding interaction between compound 53 and 5-LOX model reveals that compound 53's inhibitory mechanism is almost similar to a good 5-LOX inhibitor's typical inhibitory mechanism. Figure 5.12C and D respectively display the 3D image of the most active and least active molecule at the binding pocket of the protein molecule. It can be seen from Figure 5.12 C that the most active molecule (compound 53) was well embedded in the active site pocket by forming strong van der Waal interactions with the neighboring residues. The interaction of most active molecules 53 with amino acids at the active site supports the CoMFA steric and electrostatic contours. Polar amino acids are found to be located at the position of the chalcone, where the red contour has been observed. For instance, the C = O group forms a strong H-bond interaction with His 367, a polar amino acid. According to the CoMFA result, a red color contour is present near this group. Besides, polar amino acids such as His 432, Gln 363, Thr 364 lined up near the 5th position of the 'Ring B,' which endorses the CoMFA result where the red contour is seen. Likewise, nonpolar hydrophobic amino acids are situated near the position of chalcone where cyan color contours observed. The same result is getting for steric contour. Overall, the docking result supports the CoMFA model.

5.4. CoMFA of Benzoquinone Derivatives as Human 5-LOX Inhibitors

From the literature, it is observed that 5-LOX activity of redox inhibitors was enhanced by the presence of extended hydrophobic

alkyl groups. In our previous study, we reported the quantitative influence of long hydrophobic alkyl groups of 3', 4'-dihydroxyflavones derivative over the 5-LOX potency using CoMFA methodology [22]. In this section, we have tried to formulate QSAR models to investigate the interaction of a series of benzoquinone derivatives containing various lipophilic and bulky alkyl substituents reported by Rosanna Filosa *et al.*, [23] with the binding site of 5-LOX and predict their inhibitory activities.

5.4.1 Dataset of Benzoquinone Derivatives

The dataset used in this study consisted of a series of benzoquinone derivative that has been reported as 5-LOX inhibitors in a cell-free assay using purified human recombinant 5-LOX enzyme by Rosanna Filosa *et al.* [23]. The 2D structure of the benzoquinone core is displayed in Figure 5.13. The experimental IC_{50} values of all compounds in μM (micromole) were converted into pIC_{50} by taking $-\text{Log}(1/IC_{50})$. These pIC_{50} values of each compound are then be used as the dependent variable. A total of 48 benzoquinone derivatives were divided into a training set of 30 compounds for generating QSAR models and a test set of 11 compounds for validating the quality of the models. Seven compounds have IC_{50} value were higher than $10\mu M$ were removed. The compounds in the test set were manually selected from the original pool of structures based on Y-response (dependent variable): This approach is based on the activity (Y-response) sampling. For maintaining uniform distribution, molecules with low, moderate, and high activity were placed in both sets. Most active and

least active molecules were retained in training set for better performance. All the structures and associated inhibitory activities are listed in Table 5.9.

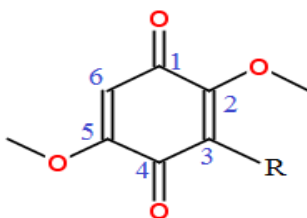


Fig. 5.13 The 2D chemical structure of the benzoquinone core.

Table 5.9 Structural formulae of compounds and their IC₅₀ values

Compound	C1	C2	C3	C4	C5	C6	IC ₅₀ (μm)
17a	=O	-OCH ₃	n-butyl	=O	-OCH ₃	H	3.3 ± 0.9
18a	=O	-OCH ₃	n-hexyl	=O	-OCH ₃	H	2.6 ± 1.2
19a	=O	-OCH ₃	n-octyl	=O	-OCH ₃	H	1.2 ± 0.3
20a	=O	-OCH ₃	n-decyl	=O	-OCH ₃	H	1.8 ± 0.7
21a	=O	-OCH ₃	n-undecyl	=O	-OCH ₃	H	0.93 ± 0.13
22a	=O	-OCH ₃	n-dodecyl	=O	-OCH ₃	H	0.61 ± 0.08
23a	=O	-OCH ₃	n-tridecyl	=O	-OCH ₃	H	0.26 ± 0.01
24a	=O	-OCH ₃	n-tetradecyl	=O	-OCH ₃	H	0.21 ± 0.03
25a	=O	-OCH ₃	n-pentadecyl	=O	-OCH ₃	H	0.27 ± 0.08
26a	=O	-OCH ₃	n-hexadecyl	=O	-OCH ₃	H	1.6 ± 0.2
27a	=O	-OCH ₃	geranyl	=O	-OCH ₃	H	3.3 ± 0.7
28a	=O	-OCH ₃	farnesyl	=O	-OCH ₃	H	1.7 ± 0.7
17b	=O	-OH	n-butyl	=O	-OCH ₃	H	>10
18b	=O	-OH	n-hexyl	=O	-OCH ₃	H	>10
19b	=O	-OH	n-octyl	=O	-OCH ₃	H	3.0 ± 0.4
20b	=O	-OH	n-decyl	=O	-OCH ₃	H	4.3 ± 0.3
21b	=O	-OH	n-undecyl	=O	-OCH ₃	H	3.8 ± 0.5
22b	=O	-OH	n-dodecyl	=O	-OCH ₃	H	0.74 ± 0.08
23b	=O	-OH	n-tridecyl	=O	-OCH ₃	H	0.92 ± 0.47
24b	=O	-OH	n-tetradecyl	=O	-OCH ₃	H	0.42 ± 0.01
25b	=O	-OH	n-pentadecyl	=O	-OCH ₃	H	0.27 ± 0.10
26b	=O	-OH	n-hexadecyl	=O	-OCH ₃	H	>10
27b	=O	-OH	geranyl	=O	-OCH ₃	H	>10
28b	=O	-OH	farnesyl	=O	-OCH ₃	H	5.6 ± 0.1
17c	=O	=O	n-butyl	-OCH ₃	-OCH ₃	H	>10
18c	=O	=O	n-hexyl	-OCH ₃	-OCH ₃	H	2.6 ± 0.3
19c	=O	=O	n-octyl	-OCH ₃	-OCH ₃	H	0.33 ± 0.05
20c	=O	=O	n-decyl	-OCH ₃	-OCH ₃	H	0.13 ± 0.01

21c	=O	=O	n-undecyl	-OCH ₃	-OCH ₃	H	0.09 ± 0.04
22c	=O	=O	n-dodecyl	-OCH ₃	-OCH ₃	H	0.13 ± 0.12
23c	=O	=O	n-tridecyl	-OCH ₃	-OCH ₃	H	0.08 ± 0.01
24c	=O	=O	n-tetradecyl	-OCH ₃	-OCH ₃	H	0.04 ± 0.02
25c	=O	=O	n-pentadecyl	-OCH ₃	-OCH ₃	H	0.62±0.14
26c	=O	=O	n-hexadecyl	-OCH ₃	-OCH ₃	H	n.d.
27c	=O	=O	geranyl	-OCH ₃	-OCH ₃	H	0.6 ± 0.08
28c	=O	=O	farnesyl	-OCH ₃	-OCH ₃	H	0.3 ± 0.07
17d	=O	-OH	n-butyl	=O	-OH	H	>10
18d	=O	-OH	n-hexyl	=O	-OH	H	4.0 ± 1.1
19d	=O	-OH	n-octyl	=O	-OH	H	0.38 ± 0.04
20d	=O	-OH	n-decyl	=O	-OH	H	0.18 ± 0.01
2	=O	-OH	n-undecyl	=O	-OH	H	0.06 ± 0.001
22d	=O	-OH	n-dodecyl	=O	-OH	H	0.17 ± 0.03
23d	=O	-OH	n-tridecyl	=O	-OH	H	0.22 ± 0.09
24d	=O	-OH	n-tetradecyl	=O	-OH	H	0.17 ± 0.08
25d	=O	-OH	n-pentadecyl	=O	-OH	H	0.23 ± 0.08
26d	=O	-OH	n-hexadecyl	=O	-OH	H	0.19 ± 0.04
27d	=O	-OH	geranyl	=O	-OH	H	1.8 ± 0.2
28d	=O	-OH	farnesyl	=O	-OH	H	2.5 ± 1.4

5.4.2. Molecular Modeling and Alignment of Benzoquinones

Molecular modeling and geometry optimization of benzoquinones are performed in the same way as flavone and chalcone derivatives. That is, DFT with B3LYP/ 6-31G (d,p) method is used for the optimization process. The resultant Gaussian output file is then converted to the SDF file. Then the alignment procedure was executed by using all available molecules as possible templates. Hence, 48 alignments were produced. For each alignment, the O3A score is computed, which measures the quality of the superimposition. The alignment corresponding to the highest cumulative O3A score was selected for further analysis. Figure 5.14 shows the best alignment in which compound 26A was selected as the template.

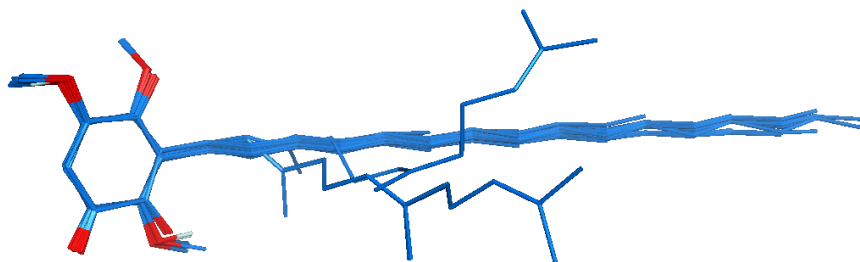


Fig. 5.14 Alignment of 48 benzoquinone derivatives.

5.4.3. Statistical Analysis of CoMFA Models of Benzoquinones

The CoMFA model with five PLS components was built using the training set of 30 benzoquinone derivatives, and then the external test set, including 11 compounds were used to assess the reliability and applicability of the built model. Statistical quality parameters associated with CoMFA models based on FFD procedures for noise reduction in the input data are listed in Table 5.10. The analysis of these parameters revealed that the best CoMFA model was obtained with a combination of steric and electrostatic fields. However, more than 70% contribution was observed from the steric field, indicating that steric interaction is essential to the binding of benzoquinone analogs with 5-LOX. CoMFA model gave good cross-validated correlation coefficient (Q^2) for LOO, LTO, and LMO as 0.5976, 0.5851, and 0.5361, respectively, indicating an excellent internal predictive power of the established model. The non-cross-validated PLS analysis with the five components resulted in traditional R^2 value of 0.8489, an F value of 26.97, and an SDEP value of 0.2203 for the

CoMFA model and was found to be a reasonable value. The values of experimental and predicted activities, along with the residual values of the training set and test set molecules, are summarized in Tables 5. 11 and 5.12, respectively. This data indicates that the inhibitors ' experimental and predicted activities are very similar to each other. Most molecules have residual values of less than 0.4. The scatter plot of observed vs. predicted values of pIC₅₀ of both the training and test set of CoMFA models is shown in Figure 5.15. This visual representation again indicates the excellent predictive power of the established model, which points out that the CoMFA model is reliable, and could be used in designing new inhibitors.

Table 5.10 Statistical data of optimal CoMFA model

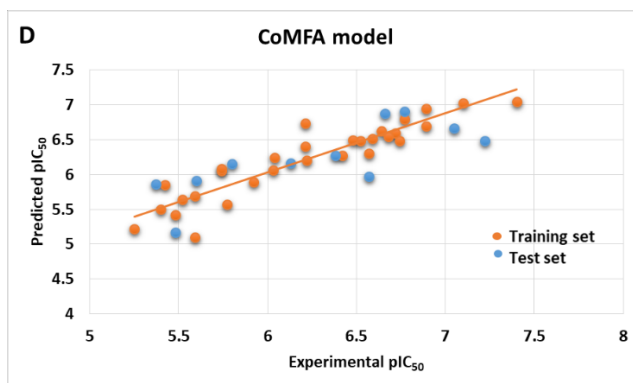
Statistical parameter	CoMFA
R ²	0.8489
Q ² (LOO)	0.5976
Q ² (LTO)	0.5851
Q ² (LMO)	0.5361
R ² _{pred}	0.6000
F	26.9743
SDEC	0.2203
SDEP	0.3651
Steric contribution	0.7715
Electrostatic contribution	0.2285

Table 5.11 The experimental and predicted pIC₅₀ values of the training set values of benzoquinones

Compound	Experimental pIC ₅₀	CoMFA	
		Predicted	Residual
28c	6.52	6.4813	0.0387
28b	5.25	5.222	0.028
28a	5.77	5.5692	0.2008
27d	5.74	6.0507	-0.3107
27c	6.22	6.209	0.011
27a	5.48	5.4236	0.0564
26d	6.72	6.5969	0.1231
25d	6.64	6.6262	0.0138
25c	6.21	6.7325	-0.5225
25a	6.57	6.3	0.27
24c	7.4	7.0461	0.3539
24a	6.68	6.5485	0.1315
23c	7.1	7.0308	0.0692
23b	6.04	6.2495	-0.2095
23a	6.59	6.5113	0.0787
22d	6.77	6.7947	-0.0247
22c	6.89	6.9464	-0.0564
22a	6.21	6.4069	-0.1969
21b	5.42	5.8567	-0.4367
21a	6.03	6.0686	-0.0386
20d	6.74	6.4875	0.2525
20c	6.89	6.6923	0.1977
20a	5.74	6.0824	-0.3424
19d	6.42	6.2737	0.1463
19c	6.48	6.4968	-0.0168
19b	5.52	5.6407	-0.1207
19a	5.92	5.8941	0.0259
18d	5.4	5.5057	-0.1057
18c	5.59	5.6966	-0.1066
18a	5.59	5.0994	0.4906

Table 5.12 The experimental and predicted pIC_{50} values of the test set values of benzoquinones

Compound	Experimental pIC_{50}	CoMFA	
		Predicted	Residual
28d	5.6	5.9155	-0.3155
26a	5.8	6.1547	-0.3547
25b	6.57	5.9767	0.5933
24d	6.77	6.9075	-0.1375
24b	6.38	6.2739	0.1061
23d	6.66	6.8809	-0.2209
22b	6.13	6.1596	-0.0296
21c	7.05	6.6664	0.3836
20b	5.37	5.8612	-0.4912
17a	5.48	5.1668	0.3132
2	7.22	6.4851	0.7349

**Fig. 5.15** Activity plots of observed vs. predicted pIC_{50} of training and test set for 5-LOX by the CoMFA model of benzoquinones.

The stability and validity of the CoMFA model were also analyzed by the progressive Y-scrambling technique. The results of the Y-scrambling test give a fitted Q^2 value of 0.281 for the model. In all cases, the obtained random models have much lower prediction accuracies than the model based on the real data, indicating no luck factor involved in the development of the CoMFA model.

5.4.4. Graphical Interpretation of the CoMFA Contour Maps of Benzoquinone

The CoMFA steric and electrostatic contour maps of benzoquinone derivatives are shown in Figure 5.16. Green and yellow contours represent the steric fields. Briefly, the green region in the steric contour maps indicates an area where the bulky groups are favored for activity while the yellow contours represent regions where the bulky groups are not favored for the activity. The red and cyan contours represent electrostatic contour maps. The cyan contour defines a region of space where positively charged substituent increases activity, whereas the red contour defines a region of space where negatively charged substituent increases activity. In the present study, the percentage contribution of the steric field and electrostatic field to the PLS model is 77.15 and 22.85%, respectively.

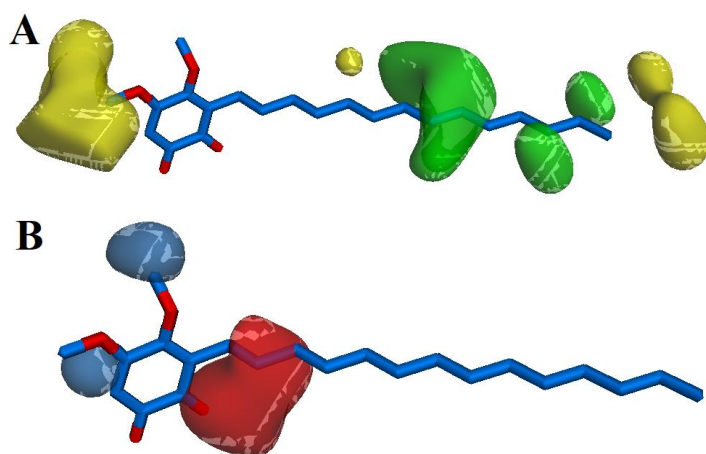


Fig. 5.16 PLS contours from 3D-QSAR models for 5-LOX benzoquinone inhibitors. (A) CoMFA steric contour maps, (B) CoMFA electrostatic contour maps.

These contour maps give us some general insight into the nature of the receptor-ligand binding region. A three-green plot was found around the middle of the 'n-alkyl' residue in position 3 indicate large groups in this region (C10-, C11, C12-, C13-, C14-, C15- and C16-) is favorable to increase the activity of the ligand. To justify this, we could say that the compounds 20d, 22d, 24d, and 26d with C10-, C12-, C14-, and C16-n-alkyl chains, respectively inhibited 5-LO with low IC₅₀ values between 0.17 and 0.19 μM than those of the compounds 17d, 18d and 19d with C4-, C6- and C8- alkyl substituent respectively. The extended, bulky alkyl group in these regions is significant for a potentially active ligand. This observation is in agreement with general findings of parent literature, showing that the potency of (poly) phenol-based 5-LO inhibitors is often enhanced due to increasing lipophilicity. This sterically crowded alkyl group may bring a hydrophobic nature to the parent benzoquinone, thereby enhances the activity. The small yellow contours at the tail portions of the alkyl residue in position 3 show that too extended alkyl chains (longer than -C16) might have a negative influence on its activity. A large yellow contour at the 5th position indicates that the -OH group is preferable at this position as compared to -OCH₃. This result is confirmed by the lower activity of compounds in 'b series' like 18b, 26b, and 27b (Table 5.9), which are the derivative of methylated hydroxyl at the 5th position of benzoquinone.

In an electrostatic field, red and blue contours are mostly distributed near to the core of the benzoquinone. So, the electronegative and electropositive substituents in these regions are likely to boost biological activity. The red contour surrounded the 2nd,

and 3rd positions suggest that the electron-rich substituents in this region are likely to enhance biological activity. It indicates alkoxy (–OCH₃) groups are preferably in the 2nd position compared to –OH group because –OCH₃ group is more electron-rich, and it allows the electrons to be donated easily. This observation is further confirmed by the higher activity of “a-series” with –OCH₃ at the position mentioned above than “b-series” and comparable to the activity of the unmethylated “d-series” with –OH at the same position. Even the n-butyl-derivative 17a was active (IC₅₀ = 3.3 mM), in contrast to the unmethylated analogs 17b and 17d. The blue electrostatic contour near the 4th and 5th position indicates the presence of an electron-deficient group is favorable at this position. Higher activities of “d-series” and “c-series” with –OH group at the 5th position and –OCH₃ at 4th position respectively support this result.

5.4.5. Molecular Docking Analysis of Benzoquinones

Interaction of benzoquinone derivative with 5-LOX was observed to get the view of ligand conformational change when undergoes docking. Since experimental activity is calculated by using human 5-LOX protein, the docking studies also were done using a human 5-LOX crystal structure with PDB ID 3O8Y. Ligand active site of 5-LOX was identified and described in Chapter 3. The optimized dimension of the grid box is 20 × 20 × 25 Å cube at -8.374, 66.379, -1.009 for x, y, and z, respectively.

Using the optimized grid box and through the molecular docking process, the interaction between protein 5-LOX and benzoquinone derivatives was deduced in the form of binding affinity value. The binding mode between the natural derivative of benzoquinone 'Embelin' (compound 2) and 5-LOX (Figure 5.17)

reveals that the inhibitory mechanism of compounds is almost similar to the typical inhibitory mechanism of a good inhibitor for 5-LOX, which should have a polar head and a hydrophobic body. The polar OH- and O=C groups at the benzoquinone head portion interact with the polar amino acids of 5-LOX by forming hydrogen bonds with His 600, Gln 363, and Leu 420 residues and can be seen in green dotted line. This observation is compatible with CoMFA electrostatic contours found around the benzoquinone head portion indicating these regions are favorable may be due to its interaction with polar amino acids of the target protein. This compound also forms hydrophobic interactions with the protein through its non-polar long alkyl part with residues like Leu 368, Ile 415, Phe 421, Phe 359, Leu 414, Leu 607, Phe 177. These findings again support the CoMFA result, which has shown the importance of large, bulky alkyl group at position 3 for a potentially active ligand.

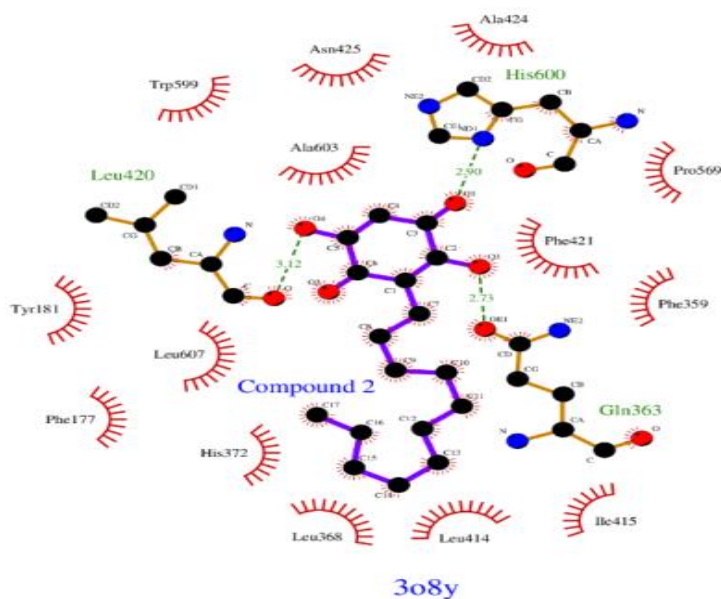


Fig. 5.17 2D view of the binding interaction of a benzoquinone derivative Embelin (Compound 2) with 5-LOX.

5.5. Conclusion

In this study, ligand-based CoMFA QSAR models with five PLSR components were developed to predict the 5-LOX inhibitory potency of three class of redox inhibitors of 5-LOX such as 3', 4'-dihydroxyflavones, 3, 4-dihydroxychalcones and benzoquinones. The developed CoMFA models were found to be statistically significant with respect to high Q^2 and R^2 values and were robust and had high internal predictive power. Moreover, the good R^2_{pred} value for an external test set confirms the excellent external predictive ability of the established CoMFA models. The contour maps extracted for each class of compound give an idea of the critical regions where any change in the steric and electrostatic field around the aligned molecules could affect 5-LOX inhibitory activity. They also provided the necessary hints of modification for the design of new molecules with better activity. Molecular docking analysis has also been carried out to examine the stability and rationality of the CoMFA models. The most and least active molecules of all class of compounds were docked to the 5-LOX active site, and the lowest energy binding pose was then used to characterize binding residues. The comparison of the interactions of the most active molecule and the contour maps of the CoMFA model provides a better understanding of the 5-LOX inhibitor interactions. In conclusion, docking results coincide well with the CoMFA result. Based on molecular docking results and extracted contour map, we could design novel inhibitors with respect to the most active compound in the dataset.

References

- [1] J. Verma, V.M. Khedkar, E.C. Coutinho, 3D-QSAR in Drug Design - A Review, *Curr Top Med Chem.* 10 (2010) 95–115. doi:10.2174/156802610790232260.
- [2] H. Kubinyi, Comparative Molecular Field Analysis (CoMFA), *Handb. Chemoinformatics.* (2003) 1555–1574. doi:10.1002/9783527618279.ch44d.
- [3] R.N. Young, Inhibitors of 5-lipoxygenase: A therapeutic potential yet to be fully realized?, *Eur. J. Med. Chem.* 34 (1999) 671–685. doi:10.1016/S0223-5234(99)00225-1.
- [4] O. Werz, D. Steinhilber, Development of 5-lipoxygenase inhibitors - Lessons from cellular enzyme regulation, *Biochem. Pharmacol.* 70 (2005) 327–333. doi:10.1016/j.bcp.2005.04.018.
- [5] S. Asgary, G.H. Naderi, N. Askari, Protective effect of flavonoids against red blood cell hemolysis by free radicals, *Exp. Clin. Cardiol.* 10 (2005) 88–90.
- [6] D.G. Batt, 1 5-Lipoxygenase Inhibitors and their Anti-inflammatory Activities, in: G.P. Ellis, D.K.B.T.-P. in M.C. Luscombe (Eds.), *Prog. Med. Chem.*, Elsevier, 1992: pp. 1–63. doi:https://doi.org/10.1016/S0079-6468(08)70004-3.
- [7] S. Shityakov, J. Broscheit, N. Roewer, C. Forster, Three-dimensional quantitative structure-activity relationship and docking studies on a series of anthocyanin derivatives as cytochrome P450 3A4 inhibitors, *Mol. Immunol.* 56 (2013) 305. doi:10.1016/j.molimm.2013.05.183.
- [8] W. Samee, P. Nunthanavanit, J. Ungwitayatorn, 3D-QSAR investigation of synthetic antioxidant chromone derivatives by molecular field analysis, *Int. J. Mol. Sci.* 9 (2008) 235–246. doi:10.3390/ijms9030235.
- [9] A.M. Ferrandiz ML, Anti-inflammatory activity and inhibition of arachidonic acid metabolism by flavonoids, *Agents Actions.* 32 (1991) 283–8.
- [10] A.M. Ferrandiz ML, Nair AG, Inhibition of sheep platelet arachidonate metabolism by flavonoids from Spanish and Indian

- medicinal herbs., *Pharmazie*. 45 (1990) 206–8.
- [11] T. Horie, M. Tsukayama, H. Kourai, C. Yokoyama, M. Furukawa, T. Yoshimoto, S. Yamamoto, S. Watanabe-Kohno, K. Ohata, Syntheses of 3',4'-dihydroxy-5,6,7- and 5,7,8-trioxygenated 3',4'-dihydroxy flavones having alkoxy groups and their inhibitory activities against arachidonate 5-lipoxygenase, *J. Med. Chem.* 29 (1986) 2256–2262. doi:10.1021/jm00161a021.
- [12] N.M.O. Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, *Open Babel: An open chemical toolbox*, (2011) 1–14.
- [13] C.J. Cramer, *Essentials of Computational Chemistry Theories and Models*, 2nd ed., John Wiley and Sons Ltd, 2004.
- [14] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* 98 (1993) 5648–5652.
- [15] R.G. Lee, C., Yang, W. and Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B.* 37 (1998) 785–789.
- [16] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J. V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, *Gaussian 09 Revision A.02*, (2009).
- [17] O. Trott, A. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading., *J. Comput. Chem.* 31 (2010) 455–461. doi:10.1002/jcc.21334.

- [18] G. Morris, R. Huey, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J. Comput. Chem.* 30 (2009) 2785–2791. doi:10.1002/jcc.21256.AutoDock4.
- [19] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera - A visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612. doi:10.1002/jcc.20084.
- [20] W.L. Delano, The PyMOL Molecular Graphics System, (2002).
- [21] S. Sogawa, Y. Nihro, H. Ueda, A. Izumi, T. Miki, H. Matsumoto, T. Satoh, 3,4-Dihydroxychalcones as potent 5-lipoxygenase and cyclooxygenase inhibitors, *J. Med. Chem.* 36 (1993) 3904–3909. doi:10.1021/jm00076a019.
- [22] T.K.S. Ahamed, K. Muraleedharan, A ligand-based comparative molecular field analysis (CoMFA) and homology model based molecular docking studies on 3', 4'-dihydroxyflavones as rat 5-lipoxygenase inhibitors: Design of new inhibitors, *Comput. Biol. Chem.* 71 (2017) 188–200. doi:10.1016/j.compbiolchem.2017.08.010.
- [23] R. Filosa, A. Peduto, A.M. Schaible, V. Krauth, C. Weinigel, D. Barz, C. Petronzi, F. Bruno, F. Roviezzo, G. Spaziano, B.D. Agostino, M. De Rosa, O. Werz, Novel series of benzoquinones with high potency against 5-lipoxygenase in human polymorphonuclear leukocytes, *Eur. J. Med. Chem.* 94 (2015) 132–139. doi:10.1016/j.ejmech.2015.02.042.

6

MODELING MACHINE LEARNING BASED QSAR

6.1. Introduction

Like CoMFA models developed in the previous chapters, several other QSAR models [1,2] for predicting 5-LOX inhibition activity have been reported in recent years to formulate an excellent predictive model consisting of common chemical characteristics. Most of these linear QSAR models were derived from a relatively small experimental dataset based on a particular type of compounds and are very limited for a complex biological system. So, with the increase in the amount and complexity of available chemical and biological data of 5-LOX inhibitors, the development of new QSAR models by enclosing all the diverse structural scaffold (chemical space of 5-LOX) and corresponding biological data is becoming more and more important for understanding and predicting the unique nature of interactions between inhibitors and 5-LOX protein. QSAR classification models are best for this purpose because bioactivity data for each class of inhibitors are obtained from different laboratories, have been generated at different experimental and assay conditions.

Also, the bioactivity results are in the various unit like IC₅₀, Ki, EC₅₀. So, it is good that the response variable (bioactivity) to be categorical. In order to analyze these vast and complex data, linear methods are not enough, so we used non-linear machine learning algorithms.

In the past ten years, Machine learning (ML) methods, mainly developed in the computer science community, have been gradually applied to cheminformatics disciplines such as SAR and QSAR to conduct more sophisticated analysis to create better models [3,4]. The QSAR models developed using non-linear ML techniques have been praised as being effective in modeling the real world more effectively than most linear models and having better predictive power. Non-linear ML technique's versatility and predictivity help them to discover more complex non-linear relationships in experimental data. Artificial Neural Network (ANN), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest (RF) or Decision Trees (DT), *etc.*, are some of the ML techniques that are commonly used. To the best of our knowledge, there is no QSAR classification model designed for 5-LOX inhibitors utilizing ML methods based on a large and diverse dataset.

In this scenario, this study developed several non-linear QSAR classification models to identify and predict 5-LOX inhibitors using a large updated and structurally diverse dataset, comprising 1605 compounds (786 inhibitors and 819 non-inhibitors) with the help of ML and data mining methods. Besides, this study evaluated the predictive performance of QSAR models using different validation

criteria such as Y-scrambling, five Cross-Validations (CV), and an external test set prediction. Moreover, the best QSAR model is used in the screening of the 'e-Drug3D' compound database.

6.2. Dataset

The dataset used in this study is retrieved from the ChEMBL database [5], which consists of a diverse set of molecules that have been tested for 5-LOX inhibition activity. This database contains 3170 compounds collected from over 100 literature and has IC_{50} , which ranges from 0.5 to 227000 nM. The purification of the dataset was carried out using the following criteria: initially, the compounds with undefined activity were removed and then, compounds containing noncovalent, mixtures, or containing salt were excluded and finally duplicate, and overlapping compounds were removed. When a compound appeared in multiple datasets or different types of literature showing different activity in each source, the lowest activity value is considered as the final activity of this compound. After all the pre-processing steps, the total number of molecules is reduced to 1605. These remaining molecules have either been classified as active compounds (786 compounds with $IC_{50} > 500\text{nM}$) or inactive compounds (819 compounds with $IC_{50} > 500\text{nM}$). All the compounds assembled, consisting of both active and inactive molecules, were randomly divided into a training set (1284 molecules) and test set (321 molecules) with a ratio of 80:20. Figure 6.1 shows a detailed workflow of *in silico* methods used in this study.

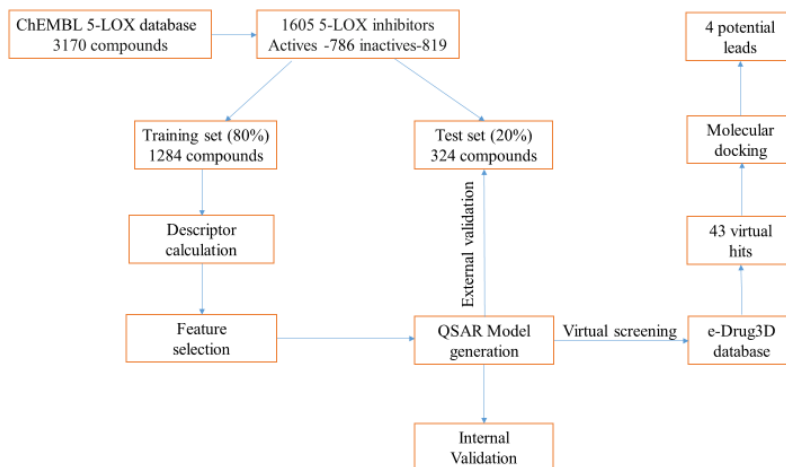


Fig. 6.1. The workflow of various *in silico* method used in this study.

6.3. Molecular Modeling and Descriptor Calculation

Chemical structures of active and inactive molecules obtained in SMILES format have been converted to 3D structural formats such as SDF (structural data file) format with the help of Open Babel [6]. The semi-empirical PM6 method implemented in the Gaussian 09 program was used to optimize the geometry of all these compounds. It was challenging to perform descriptor calculation as a single file due to the large size of the dataset containing all the compounds. In order to overcome these difficulties, the dataset in SDF files was split into smaller files using the SplitSDFfiles Perl Script available in the Mayachem tools [7] then calculated the descriptors for each set. For the development of effective and robust non-linear binary QSAR models with superior performance, descriptors like topological, constitutional, electronic, geometrical, and spatial are calculated for each of the compounds utilizing three different software such as E-DRAGON, PowerMV, and OCHEM.

1. E-DRAGON descriptors

E-DRAGON [8] is an online version of Dragon software, contains scripts for calculating thousands of molecular descriptors like topological descriptors, walk and path counts, connectivity indices, information indices, or 2D-autocorrelations, RDF, GETAWAY, functional groups, WHIM, Randic, 3D-Morse, *etc.* For each compound, a total of 1666 E-DRAGON descriptors were computed. These 1666 descriptors encompassed different categories, which are tabulated in Table 6.1.

Table 6.1 E-DRAGON descriptor categories used for this study

Cat. No.	Block description	Descriptor count
1	Constitutional descriptors	43
2	Topological indices	119
3	Walk and path counts	47
4	Connectivity Indices	33
5	Information indices	47
6	2D autocorrelations	96
7	Burden eigenvalues	64
8	Edge adjacency indices	107
9	Geometrical descriptors	74
10	RDF descriptors	150
11	3D-MoRSE descriptors	160
12	WHIM descriptors	99
13	GETAWAY descriptors	197
14	Randic molecular profiles	41
15	Functional group counts	154
16	Atom-centered fragments	120
17	Charge descriptors	14
18	Molecular properties	31
19	Topological charge indices	21
20	Eigenvalue-based indices	44

2. OCHEM descriptors

The Online Chemical Modeling Environment (OCHEM) [9] is a web-based platform aimed at automating and simplifying the typical QSAR modeling steps. It also provides a platform for the calculation of several different descriptor packages. A total of 6237 descriptors from seven categories are illustrated in Table 6. 2.

Table 6.2 OCHEM descriptor categories used for this study

Cat. No.	descriptor packages	Descriptor count
1	GSFragment	1138
2	Inductive descriptors	54
3	MERSY descriptors	42
4	RDKit descriptors	223
5	Spectrophore	144
6	Structural Alert	2318
7	Toxicity Alert	2318

3. PowerMV descriptors

PowerMV software is commonly used to compute Pharmacophore Fingerprint descriptors and Weighted Burden Number descriptors [10]. Pharmacophore Fingerprint descriptors are constructed using bioisosteric principles (Two atoms or groups that are expected to have roughly the same biological effect are called bioisosteres). The categories of descriptors included in the PowerMV database is given in Table 6.3.

Table 6.3 PowerMV descriptor categories used for this study

Cat. No.	descriptor packages	Descriptor count
1	pharmacophore fingerprints	147
2	weighted burden numbers	24
3	property descriptors	8

Both online (E-DRAGON, OCHEM) and offline (PowerMV) software generates different types and numbers of descriptors for the same compound. Selected descriptors from each program can be considered as separate databases. So now we have three databases. Again, a new database called a combined database, which covers an almost entire range of descriptor space, is constructed by combining all these three single databases. Then these four databases (E-DRAGON, OCHEM, PowerMV, and Combined) were cleaned up by implementing a filter tool available in the data mining program WEKA, eliminating the constant and near-constant variables.

6.4. Chemical Space Characterization

A growing interest in the effort to develop QSAR models from structurally diverse datasets has recently been observed because the chemical diversity of the dataset is highly recommended to build robust and efficient predictive QSAR models. The higher the diversity of the compounds in the dataset, the higher the model's applicability domain. Visual representation of chemical space occupied by training and test set of 5-LOX inhibitors would provide an idea about how diverse the dataset we have taken. This study used Principal

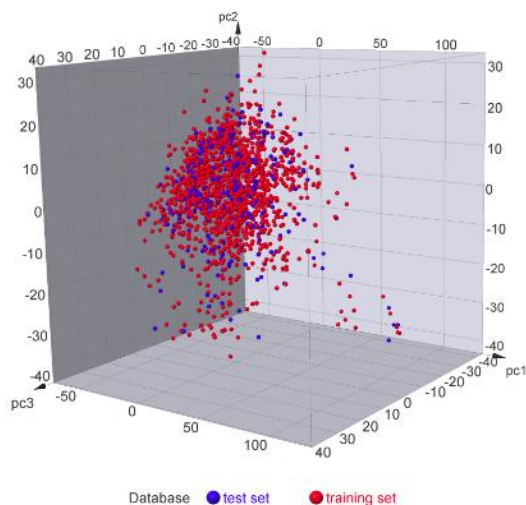
Component Analysis (PCA) and Self-Organizing Map (SOM) implemented in DataWarrior [11] to analyze and perceive dataset. These two common chemical space visualization methods reduce the multi-dimensional space into a graph of two or three dimensions. Since the model generation involved three different descriptor database containing complex descriptors, the chemical space was also generated using the same descriptors.

We built the three PCA models separately with the E-DRAGON, OCHEM, and PowerMV descriptors set. Figures 6. 2A, B, and C depict the scatter plot of the PC1, PC2, and PC3 space of E-DRAGON, OCHEM, and PowerMV database, respectively. All three descriptor databases show the maximum diversity points in the compound space because most of these descriptors represent the two- and three-dimensional diverse structural and physicochemical aspects of compounds considered in this study. The distribution of the training (red-colored marker) and test set (blue colored marker) compounds over the space of the principal components in all three cases indicates the diversity and representative ability of both subsets. We can see from the PCA map of all three datasets that the entire test set compound fell within the applicability domain of the training set PCs. This result concludes that both training and test sets of 5-LOX inhibitor datasets were shared similar chemical space.

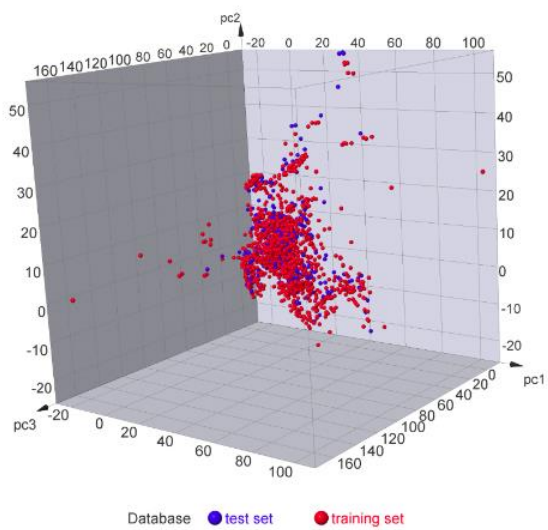
In contrast to PCA, SOM is a non-linear multi-dimensional mapping tool that can be used to represent low-dimensional, topology-preserving projections of high-dimensional data. Figure 6. 2D shows,

SOM of the chemical space of 5-LOX inhibitors utilizing the E-DRAGON descriptors as similarity criteria. Similar compounds wind up in landscape influenced color as topographical neighbors. The view's background colors envision neighborhood similarity of colors inspired by adjacent neurons in the landscape. Blue to green colors imply valleys of similar neurons, while yellow to orange areas of adjacent neurons reveal ridges of more abrupt changes in the chemical space. Such a way very similar compounds congregate in the same valleys, while yellow ridges separate slightly different clusters. The majority of the training (red-colored marker) and test set (blue colored marker) compounds have clear topographical neighbors, and each valley of similar training set compounds offer one or more test set as close neighbors, this indicates test compounds fell within the applicability domain (AD).

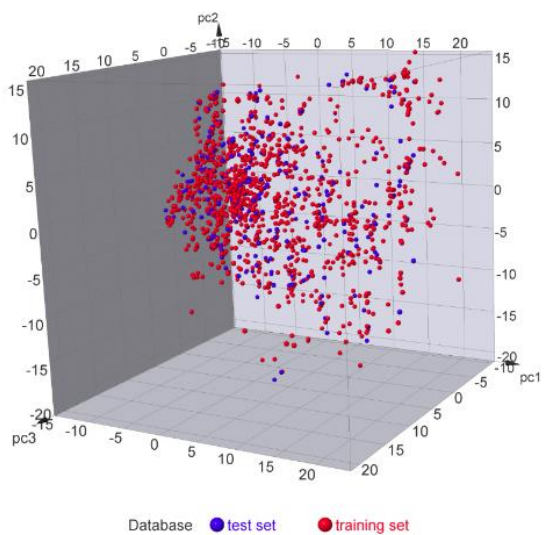
A



B



C



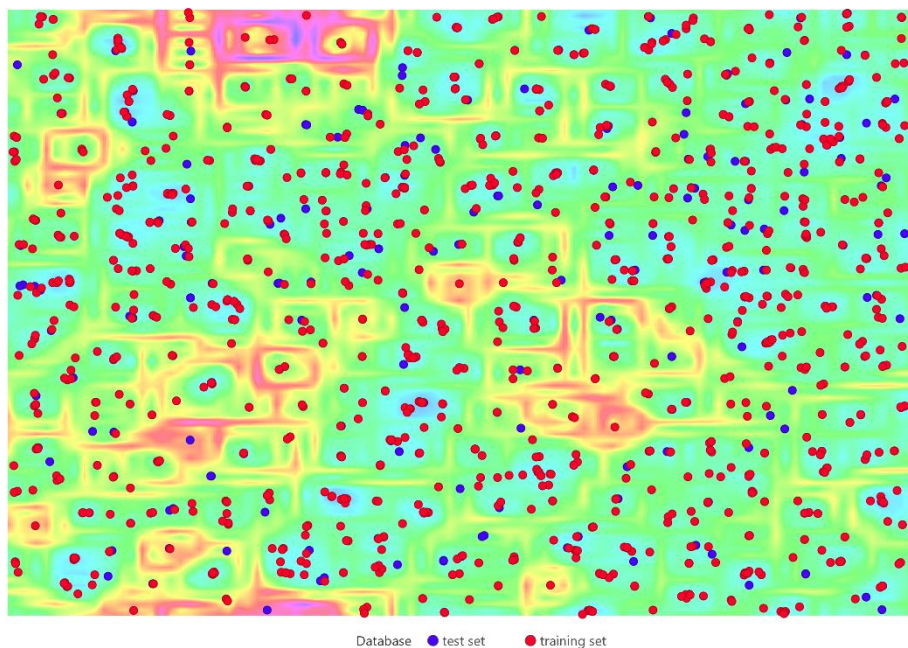


Fig. 6.2 PCA plot for 1605 5-LOX inhibitors developed using A) E-DRAGON descriptors, B) OCHEM descriptors, and C) PowerMV descriptors. D) Scaffold grouping on a SOM for the training and test set compounds using E-DRAGON descriptors. A red and blue circle represents training and test set compounds.

A scatter graph represents the property space of the dataset used in this study with a molecular weight (MW) along with the y-axis and partition coefficient (LogP) along the x-axis, as shown in Figure 6.3. This diagram shows that the compounds in the training set and test set shared similar property space, and this space is restricted for compounds having log P value greater than 12 and MW greater than 1200 atomic units.

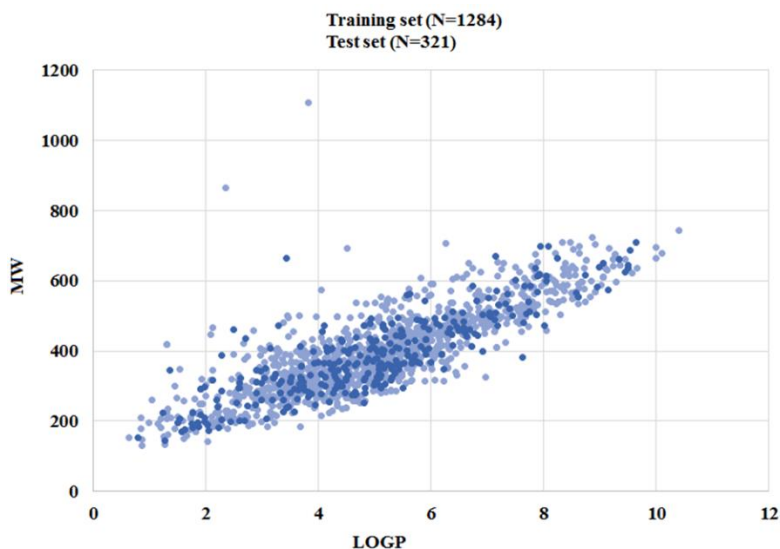


Fig. 6.3 Chemical space was defined by molecular weight (MW) and LogP (N= no of chemicals), where light blue indicates training set and dark blue indicate test set.

6.5. Structure-Activity Landscape Analysis

Although the above-mentioned unsupervised approaches such as PCA and SOM for displaying the chemical space provide some confidence for pursuing modeling activities, further evaluations have been carried out in the dataset to understand changes in biological properties as a result of minimal changes in structural patterns of compound sets. For all pairs of similar molecules, the Structure-Activity Landscape Index (SALI) is calculated. It provides a metric of how much activity is gained or lost due to a relatively small structural change [12]. SALI networks can be used to measure the ability of a QSAR model to encode one or more SAR trends. SALI value between two molecules is defined as the ratio between the difference in

biological activity (ΔpIC_{50}) to the dissimilarity (1–similarity) of the pair. High SALI valued region indicates that small structural changes yield substantial changes in an activity. The presence of more of these regions in the landscape indicates there will be hardly any possibility of developing a strong QSAR. However, a small SALI valued region indicates an exciting starting point for the development of QSAR. That is, for the development of QSAR models, the dataset should contain maximum pairs of molecules that have a similar chemical structure with small SALI values.

In this study, the SkeletonShperes descriptor used as a similarity criterion. SkeletonShperes descriptor is a byte vector with a resolution of 1024 bins, and it includes additional consideration of stereochemistry, counting of duplicate fragments, encoding heteroatom depleted skeletons. Compound activities and SALI values for each of the molecule pair being compared are represented in a scatter plot (Figure 6.4A). Big circles represent pairs of compounds exhibiting the highest SALI values, and small circles indicate pairs exhibiting the minimum SALI values. Markers (here, circles) have colored from red to blue depending on the increase in delta activities (ΔpIC_{50}) of the pairs of compounds. Around 2967 pairs were identified based on a similarity cut off threshold 95%. Among these, only fewer pairs can be attributed to the training set, and this provides another inference for the compound diversity within the training dataset. Biological property space analysis shows that the overall landscape is pretty smooth (high molecular similarity and high activity similarity), with ~ 2579 pairs exhibiting $\Delta\text{pIC}_{50} < 1$ log unit. However, there exist some rugged

landscapes due to the presence of the activity cliff. The activity cliff is already explained in Chapter 4 has been defined as a pair of structurally similar compounds with a tremendous difference in bioactivity [13]. SALI values were analyzed to generate a graph showing the activity cliffs.

Structure-activity similarity (SAS) analysis was conducted to identify the pairs of 5-LOX inhibitors that display activity cliffs using DataWarrior. The image in Figure 6.4B shows a similarity map of all 5-LOX inhibitors, which encodes pIC_{50} of the compounds represented by marker color (green to dark blue). Similar compounds are connected with a line. By looking at the image we can easily recognize clusters of similar compounds with similar activity (clusters of similar colored markers attached with a line), locate activity cliffs (green markers connected to red and blue markers) and locate training and test set compounds (circle and square markers respectively) in the chemical space. The analysis shows that clusters of 5-LOX inhibitors that displayed activity cliffs are rare and each cluster of near neighbors in the training set that surrounds one or more test set based on similarity and the SALI index. This observation indicates that this smooth region of the structure-activity landscape formed by training and test set compounds that share similar chemical and biological space can be used as modeling space for building efficient and predictive QSAR models that can ease the process of lead optimization efforts for 5-LOX protein target.

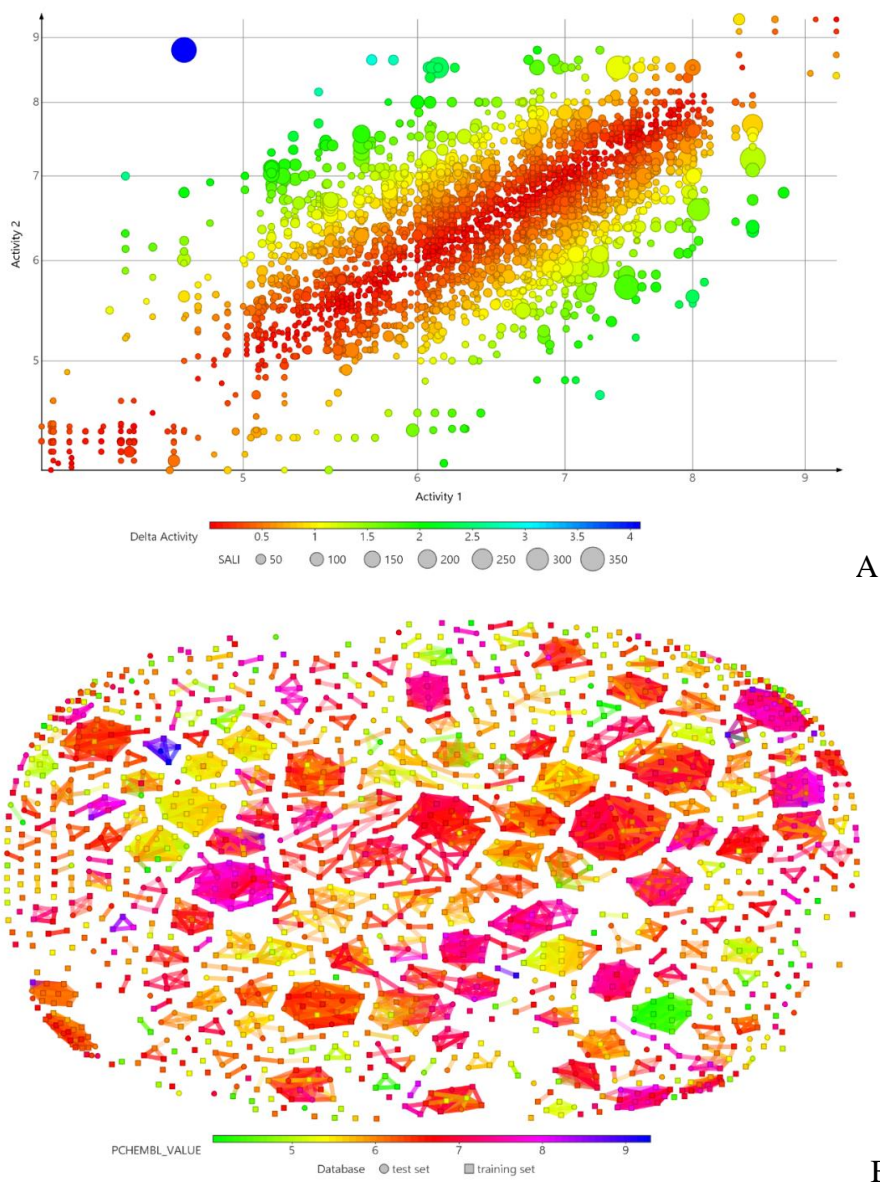


Fig. 6.4 A) SALI plot of compound pairs (training and test) with >95% similarity. B) Activity cliffs (marker size) for the training and test set grouped based on their neighborhood similarity relationships.

6.6. Predictive Modeling Using Machine Learning Techniques

Machine learning (ML) techniques show excellent performance in constructing QSAR models using the dataset in which the structure-activity relationship is often complex and non-linear [14]. Four types of ML algorithms, such as SVM, kNN, logistic regression, and decision tree, were used in this study to build QSAR classification models from an input dataset of molecular descriptors and activity labels and are termed as 'Classifiers.' The necessary details of these methods are given in Chapter 2. In order to maximize the algorithm performance, we have re-trained the algorithm by finely tuning the parameter values. So, in SVM, this study used both the Radial basis function (RBF) kernel and Polynomial kernel, and the parameters C and γ for RBF kernel were tuned on the training set by 5-fold cross-validation. The Sequential Minimal Optimization (SMO) of John Platt [15] implemented in WEKA is used in this study to train the support vector classifier. SMO normalizes all attributes, replaces all missing values, and transforms nominal attributes into binary ones. For kNN, this study used 1,3,5,7, and 9 numbers of k values and the appropriate values for each descriptor set were searched. The nearness was measured by the Euclidian distance metrics, and kNN prediction accuracies are estimated through five-fold cross-validation. The J48 is an open-source Java implementation of the C4.5 algorithm introduced in WEKA, which is used in this study to build decision trees from a training set based on the criterion of normalized information. So, these four ML techniques with assigned parameters and various learning algorithms are used in this study to construct the QSAR. However,

before building the QSAR model using these techniques, we need to remove noisy and irrelevant descriptors. This strategy will be explained in the following section 6.6.1.

6.6.1. Feature Selection

The main objective of this study was to compare the accuracy of the predictive performance of 5-LOX inhibitors QSAR classification models built by various ML methods and different descriptor combinations. However, not all of the calculated molecular descriptors are needed for representing features between 5-LOX inhibitors and non-inhibitors. Noisy, redundant, or irrelevant descriptors should be removed without much loss of information, thereby reducing the risk of overfitting. Feature Selection methods have been used to select suitable descriptors without loss of information from a vast number of raw descriptors that contain little information or are correlated with other descriptors. This study selectively chose two types of filter method that has been implemented in Weka to perform feature selection such as CfsSubsetEval (CFS) module in combination with the BestFirst search method and InfoGainAttributeEval (IG) module in combination with the Ranker search method. These two methods have been discussed in detail in Chapter 2.

CFS and IG feature selection methods separately extracted one set of the descriptor from each of four databases (E-DRAGON, PowerMV, OCHEM, and Combined) and to form a total of 8 training sets. Each of these training sets contains the same compounds but with

different descriptor subsets so treated as a different dataset. With these training sets, several classifiers were trained, and the best single classifier with best descriptor combinations is chosen by comparing the predictive performance of each model. Based on this, it is also possible to identify the best filter method.

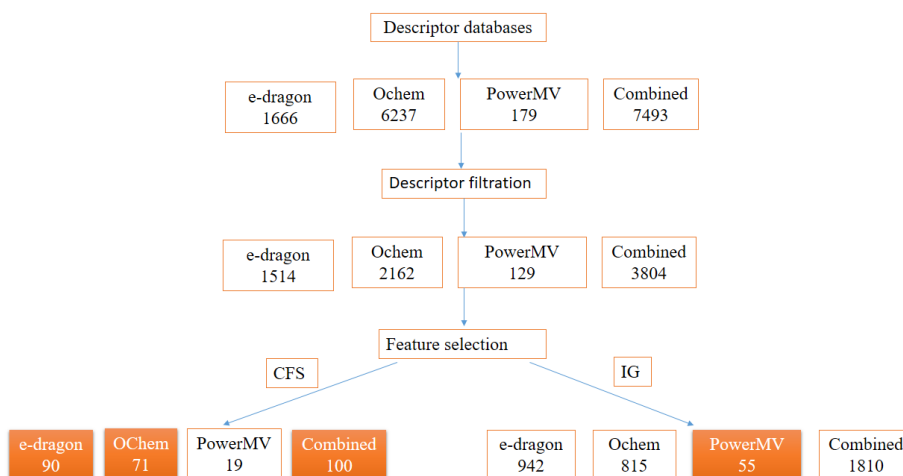


Fig. 6.5 Flowchart showing descriptor databases and the number of descriptors that remained after each step of the descriptor reduction process. Orange colored square indicate optimal descriptors set that have been used to construct ML models.

Before building classification, models based on the filtered descriptor set, an estimate of the predictive performance of models generated by databases containing the full set of descriptors must be obtained. This process allows performing a comparative evaluation of the performance of the filtered set and the original descriptor set that is not filtered. For this purpose, different classification algorithm has been applied to the original databases of E-DRAGON, OCHEM,

PowerMV, and Combined containing 1666, 6237, 179 and 7493 molecular descriptors, respectively. Simple classifiers were generated by using Weka software such as SVM, kNN, Logistic Regression, and J48 (Decision Tree). The sizes of training and external test sets were 1284 and 321 compounds, respectively. Nonetheless, the number of molecular descriptors for each compound was much higher. The number of descriptors was reduced to avoid this fundamental problem and ranked them according to their variability.

Figure 6.5 shows a flow chart that describes different descriptor reduction pathways used in this study. Of the approximately 1666 (E-DRAGON), 6237 (OCHEM) and 179 (PowerMV) and 7493 (Combined) descriptors initially calculated, a RemoveUseless filter method available in Weka was applied to eliminate descriptors that exhibited a low variance throughout the dataset. This process left 1514, 2162, 129, and 3804 molecular descriptors of E-DRAGON, OCHEM, PowerMV, and Combined databases, respectively. These sets were further undergone feature selection to extract the most relevant descriptors. The number of descriptors that have a higher information gain value was selected, and descriptors that have a lower score were removed using the information gain (IG) method. As a result, E-DRAGON descriptors decreased from 1514 to 942. The CFS filter method measured the correlation between nominal features in the descriptor set. So the most relevant attribute set was filtered to produce the most promising subset. CFS method reduces the number of E-DRAGON descriptors to 90. Likewise, IG and CFS filter method reduced OCHEM descriptors from 2162 to 815 and 71, respectively,

and PowerMV descriptors were reduced from 129 to 55 and 19, respectively. Feature selection of combined database descriptors is also made, a total of 1810 descriptors is selected through the IG method, and 100 highly relevant descriptors were selected by the CFS method. The performance of feature selections method is assessed by constructing classification models of each descriptor datasets built with the help of ML algorithms such as kNN, LOGISTIC, SVM, and J48. The performance of each model was evaluated by 5-fold cross-validation is given in Table 6.4, and the best models were selected based on the values of classification accuracy (CA) and area under the ROC curve (AUC). The detailed description of the above mentioned parameters is provided in Chapter 2. A good model of binary classification always yields high sensitivity, specificity, CA, and AUC values.

Table 6.4 Performance of classification models in each step of the feature selection estimated by five-fold cross-validation

MODEL	SEN	SPE	CA	AUC
DRAGON-ALL-KNN	72.8	71.7	72.2	72.2
DRAGON-IG-KNN	79.8	71.1	75.2	81.7
DRAGON-CFS-KNN	81.1	72.7	76.7	82.7
DRAGON-ALL-J48	65.6	70.8	68.3	67.7
DRAGON--IG-J48	66.4	69.9	68.2	69.3
DRAGON--CFS-J48	73.1	68.0	70.4	72.1
DRAGON-ALL-LOGISTIC	59.4	64.4	62.0	67.2
DRAGON--IG-LOGISTIC	67.5	65.9	66.7	71.4
DRAGON--CFS-LOGISTIC	71.5	75.7	73.7	71.4
DRAGON-ALL-SVM	72.8	71.7	72.1	72.2
DRAGON--IG-SVM	73.1	74.1	72.2	83.7
DRAGON--CFS-SVM	67.5	76.6	72.3	86.4
OCHEM-ALL-KNN	79.1	72.0	75.4	83.7

Modeling Machine Learning Based QSAR

OCHEM-IG-KNN	77.7	74.8	76.2	83.9
OCHEM-CFS-KNN	80.6	73.3	76.8	84.2
OCHEM-ALL-J48	69.2	70.1	70.1	71.0
OCHEM-IG-J48	70.8	72.0	71.4	72.2
OCHEM-CFS-J48	70.1	74.1	72.2	72.2
OCHEM-ALL-LOGISTIC	69.2	66.6	67.8	72.9
OCHEM-IG-LOGISTIC	70.6	68.4	69.5	82.2
OCHEM-CFS-LOGISTIC	72.9	75.3	74.1	82.2
OCHEM-IG-SVM	73.9	76.2	75	85.7
OCHEM-CFS-SVM	71.3	78.8	75.2	85.5
POWERMV-ALL-KNN	76.2	69.9	72.9	80.3
POWERMV-IG-KNN	79.8	73.8	76.6	82.4
POWERMV-CFS-KNN	80.1	68.1	73.8	81.7
POWERMV-ALL-J48	67.2	72.1	69.7	73.5
POWERMV-IG-J48	73.4	73.5	73.4	80.0
POWERMV-CFS-J48	68.2	72.0	70.2	76.0
POWERMV-ALL-LOGISTIC	61.7	74.5	68.2	74.0
POWERMV-IG-LOGISTIC	69.7	73.9	71.9	72.6
POWERMV-CFS-LOGISTIC	61.7	74.5	68.4	74.1
POWERMV-IG-SVM	63.5	77.6	70.9	85.2
POWERMV-CFS-SVM	59.7	74.4	67.4	85.2
COMB-ALL-KNN	82.5	71.1	76.5	84.0
COMB-IG-KNN	80.8	72.9	76.6	84.1
COMP-CFS-KNN	82.4	74.2	78.1	84.2
COMB-ALL-J48	69.6	70.5	70.1	70.0
COMB-IG-J48	69.7	72.6	71.2	71.7
COMB-CFS-J48	69.5	75.0	72.4	72.8
COMB-ALL-LOGISTIC	63.0	59.8	61.2	64.9
COMB-IG-LOGISTIC	68.2	63.6	65.8	75.5
COMB-CFS-LOGISTIC	72.3	76.5	74.5	82.9
COMB-ALL-SVM	74.6	74.1	74.3	74.3
COMB-IG-SVM	76.5	74.5	75.5	85.6
COMB-CFS-SVM	71.8	79.4	75.8	87.3

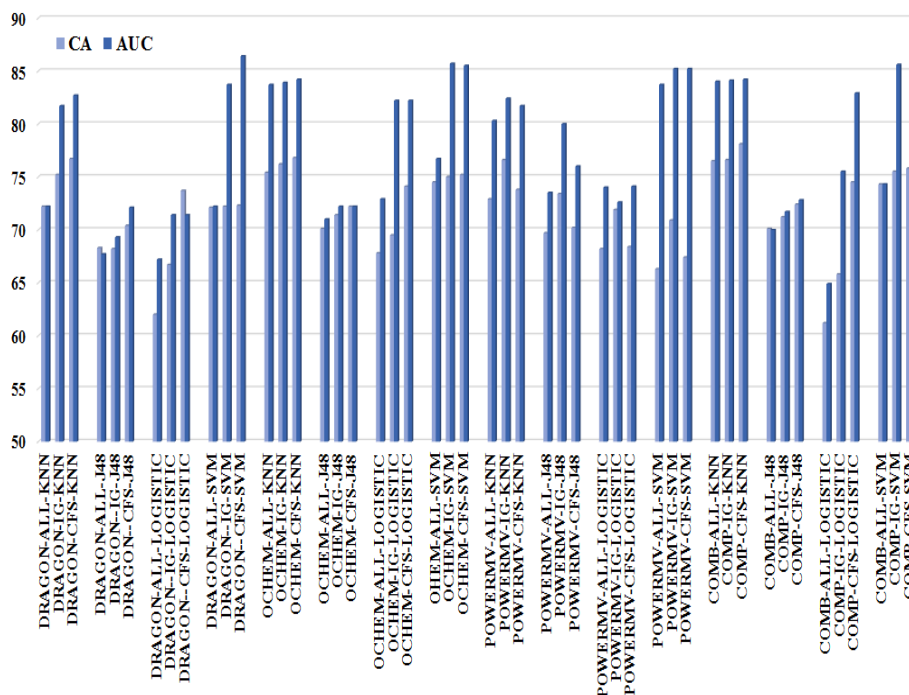


Fig. 6.6 Performance of classification models in each step of the feature selection estimated by five-fold cross-validation

Figure 6.6 displays the improvement in the predictive accuracy of each model by the feature selection method. In general, the classification models generated after feature selection gives much better statistical result than the classification models obtained using all the molecular descriptors. This result indicates the performance of all the classifiers for predicting 5-LOX inhibitory activity was improved, employing the feature selection methods. E-DRAGON, OCHEM, and Combined descriptors-based models that used the CFS method for feature selection had the CA and AUC values higher than models that used the IG method, in contrast, PowerMV descriptor-based models that used CFS method had CA, and AUC values are much lower than

that of the IG method. This result leads to the conclusion that the CFS method is more accurate in removing redundant descriptors than IG for all descriptor databases except for PowerMV. For removing unwanted descriptors from the PowerMV set, the IG method is better than the CFS method. After feature selection, the optimum size of the descriptor sets is in the range of 50-100. The descriptors of each database obtained after the feature selection are given in Table 6.5.

Table 6.5 Descriptors set obtained from each database after feature selection

COMBINED-CFS
MSD, D/Dr07, D/Dr10, SRW07, CIC3, MATS3v, MATS4v, MATS5v, MATS6v, MATS2e, MATS3e, MATS5e, MATS6e, MATS7e, MATS8p, GATS2m, GATS4v, GATS3p, EEig01d, EEig04d, ESpm10x, ESpm14d, BELm1, BEHv1, BEHp1, JGI2, JGT, SEigZ, DISPv, DISPp, RDF110m, RDF155m, Mor14u, Mor02m, Mor09v, Mor02p, Mor15p, P1u, E2m, E2e, E1s, HATS7u, HATS8u, H0m, HATS3m, R5m+, R8v+, R7e+, c8A, c5AC-6N, SmallestNegHardness, SmallestNegSoftness, MostPosSigmaMolI, SYMC3X, SYMC4X, PEOE_VSA10, PEOE_VSA13, SMR_VSA1, SMR_VSA4, ESTATE_VSA7, LOGP, SpectrophoresPartial_8, SpectrophoresLipophilicity_3, SpectrophoresLipophilicity_4, SpectrophoresLipophilicity_6, SpectrophoresLipophilicity_9, SpectrophoresElectrophilicity_1, SpectrophoresElectrophilicity_5, Alert892, Alert919, Alert1076, Alert1960, Alert2021, Alert2132, Alert893, Alert920, NEG_06_ARC, POS_04_ARC, ARC_03_ARC, WBN_GC_L_0.25, WBN_GC_L_0.50, WBN_GC_H_0.50, R4p, Hy, c6AD, c6ABC, c6ABD, p1p2-4O, p5-3O, p1p1p2-6O, p3-1F, c5AC-6C, p3-2N, c5AB-4O, WBN_GC_H_0.75, WBN_GC_L_1.00, WBN_EN_L_0.25, WBN_EN_L_0.50, WBN_EN_H_0.50
DRAGON-CFS
nR08, nR10, nR11, MSD, PJI2, D/Dr07, D/Dr10, D/Dr11, T(N..N), T(O..O), T(F..F), SRW07, piPC10, CIC3, MATS3v, MATS4v, MATS5v, MATS6v, MATS2e, MATS3e, MATS5e, MATS6e, MATS7e, MATS8p, GATS2m, GATS4v, GATS3p, EEig01d, EEig04d, ESpm10x, ESpm14d, BELm1, BEHv1, BELe1, BEHp1, JGI2, JGT, SEigZ, J3D, DISPv, DISPp, G(N..O), G(O..O), RDF110m, RDF155m, Mor14u, Mor02m, Mor09v, Mor02p, Mor15p, P1u, E2m, E2e, E1s, HATS7u, HATS8u, H0m, HATS3m, HATS7e,

R4u, R5m+, R8v+, R7e+, R4p, nRCNO, nOHs, nArOR, nPyrazoles, N-074, Hy
OCHEM-CFS
c6AD, c6ABC, c6ABD, c10, p3-2C, p5-2C, p1p2-4O, p5-3O, p1p1p2-6O, p1p4-5O, p3-1F, c5AC-6C, p3-2N, c6-1N, p1c7, c5-1S, c6A-2S, c5AB-4O, c8A, c5AC-6N, SumPosHardness, SmallestNegHardness, SmallestNegSoftness, MostPosSigmaMolI, SmallestRsIMol, SYMC3X, SYMC4X, SYMS3, MAXPARTIALCHARGE, PEOE_VSA10, PEOE_VSA13, PEOE_VSA3, SMR_VSA1, SMR_VSA4, SLOGP_VSA2, SLOGP_VSA3, SLOGP_VSA8, TPSA, ESTATE_VSA7, ESTATE_VSA9, NUMAROMATICCARBOCYCLES, MOLLOGP, FR_HDRZINE, LOGP, SpectrophoresPartial_8, SpectrophoresPartial_9, SpectrophoresLipophilicity_3, SpectrophoresLipophilicity_4, SpectrophoresLipophilicity_6, SpectrophoresLipophilicity_9, SpectrophoresShape_8, SpectrophoresElectrophilicity_1, SpectrophoresElectrophilicity_2, SpectrophoresElectrophilicity_3, SpectrophoresElectrophilicity_5, SpectrophoresElectrophilicity_6, Alert153, Alert374, Alert498, Alert731, Alert892, Alert908, Alert916, Alert919, Alert1076, Alert1960, Alert2021, Alert2132, Alert2166, Alert893, Alert920
POWERMV-IG
ARC_01_ARC, ARC_02_ARC, ARC_03_ARC, ARC_03_HYP, ARC_04_ARC, ARC_06_ARC, ARC_07_HYP, BadGroup, HBA_02_HYP, HBA_05_ARC, HBD_03_ARC, HBD_04_ARC, HBD_05_HBD, HBD_06_HBA, HBD_07_ARC, HBD_07_HBA, HBD_07_HBD, MW, NEG_05_HYP, NEG_06_ARC, NEG_06_HYP, NEG_07_ARC, NumHBA, NumHBD, NumRot, POS_04_ARC, POS_05_HBA, POS_06_HBA, PSA, XLogP, WBN_EN_H_0.25, WBN_EN_H_0.50, WBN_EN_H_0.75, WBN_EN_H_1.00, WBN_EN_L_0.25, WBN_EN_L_0.50, WBN_EN_L_0.75, WBN_EN_L_1.00, WBN_GC_H_0.25, WBN_GC_H_0.50, WBN_GC_H_0.75, WBN_GC_H_1.00, WBN_GC_L_0.25, WBN_GC_L_0.50, WBN_GC_L_0.75, WBN_GC_L_1.00, WBN_LP_H_0.25, WBN_LP_H_0.50, WBN_LP_H_0.75, WBN_LP_H_1.00, WBN_LP_L_0.25, WBN_LP_L_0.50, WBN_LP_L_0.75, WBN_LP_L_1.00,

6.6.2. Performance Evaluation of Classifiers

Descriptor sets that obtained from the feature selection method were used for building classification models. A total of 52 individual classifiers derived by thirteen ML techniques based on four descriptor sets filtered from four different databases were developed. ML algorithms used were a Logistic regression, SVM polynomial with complexity parameter $C=0.1, 1$ and 10 , and RBF with $C=1$ and $\gamma=0.01, 1$ and 10 , kNN with $k=1, 3, 5, 7$ and 9 (Euclidean distance as metric distance) and J48. The detailed evaluation of the results from these models was given in Table 6.6-6.9

Table 6.6 Comparison of classification models of the E-DRAGON database built with different MLTs estimated by a five-fold cross-validation

Model	TP	FN	FP	TN	SEN	SPE	CA	AUC
DRAGON-CFS-kNN(k=1)	447	166	186	485	72.9	72.3	72.6	71.4
DRAGON-CFS-kNN (k=3)	485	128	184	487	79.1	72.6	75.7	80.3
DRAGON-CFS-kNN (k=5)	490	123	179	492	79.9	73.3	76.5	82.3
DRAGON-CFS-kNN (k=7)	497	116	183	488	81.1	72.7	76.6	82.7
DRAGON-CFS-kNN (k=9)	491	122	196	475	80.1	70.8	75.2	82.5
DRAGON-CFS-J48	448	165	215	456	73.1	68.0	70.4	70.3
DRAGON-CFS-LOGISTIC	438	175	163	508	71.5	75.7	73.7	80.9
DRAGON-CFS-SVM (Polykernel, C=0.1)	380	233	139	532	62.0	79.3	71.0	70.6
DRAGON-CFS-SVM (Polykernel, C=1)	414	199	157	514	67.5	76.6	72.3	72.1
DRAGON-CFS-SVM (Polykernel, C=10)	425	188	153	518	69.3	77.2	73.4	73.3
DRAGON-CFS-SVM (RBFkernel, C=1 $\gamma=0.01$)	310	303	108	563	50.6	83.9	68.0	67.2
DRAGON-CFS-SVM (RBFkernel, C=1 $\gamma=1$)	453	160	137	534	73.9	79.6	76.9	76.7
DRAGON-CFS-SVM (RBFkernel, C=1 $\gamma=10$)	311	302	73	598	50.7	89.1	70.8	69.9

Table 6.7 Comparison of classification models of OCHEM database built with different MLT estimated by five-fold cross-validation

Model	TP	FN	FP	TN	SEN	SPE	CA	AUC
OCHEM-CFS-kNN (k=1)	457	156	189	482	74.6	71.8	73.1	72.2
OCHEM-CFS-kNN (k=3)	475	138	189	482	77.5	71.8	74.5	80.5
OCHEM-CFS-kNN (k=5)	483	130	173	498	78.8	74.2	76.4	83.5
OCHEM-CFS-kNN (k=7)	494	119	179	492	80.6	73.3	76.8	84.2
OCHEM-CFS-kNN (k=9)	487	126	183	488	79.4	72.7	75.9	84.2
OCHEM-CFS-j48	434	179	188	483	70.8	72.0	71.4	71.2
OCHEM-CFS-LOGISTIC	447	166	166	505	72.9	75.3	74.1	82.2
OCHEM-CFS-SVM (Polykernal, C=0.1)	354	259	102	569	57.7	84.8	71.9	71.3
OCHEM-CFS-SVM (Polykernal, C=1)	437	176	142	529	71.3	78.8	75.2	75.1
OCHEM-CFS-SVM (Polykernal, C=10)	454	159	170	501	74.1	74.7	74.4	74.4
OCHEM-CFS-SVM (RBFkernal, C=1 $\gamma= 0.01$)	312	301	87	584	50.9	87.0	69.8	40.9
OCHEM-CFS-SVM (RBFkernal, C=1 $\gamma= 1$)	447	166	134	537	72.9	80.0	76.6	76.5
OCHEM-CFS-SVM (RBFkernal, C=1 $\gamma= 10$)	307	306	81	590	50.1	87.9	69.9	69.0

Table 6.8 Comparison of classification models of PowerMV database built with different MLT estimated by five-fold cross-validation

Model	TP	FN	FP	TN	SEN	SPE	CA	AUC
POWERMV-IG-kNN (k=1)	468	145	172	499	76.3	74.4	75.3	75.5
POWERMV-IG-kNN (k=3)	483	130	171	500	78.8	74.5	76.6	81.5
POWERMV-IG-kNN (k=5)	489	124	176	495	79.8	73.8	76.6	82.4
POWERMV-IG-kNN (k=7)	480	133	183	488	78.3	72.7	75.4	82.3
POWERMV-IG-kNN (k=9)	477	136	186	485	77.8	72.3	74.9	82.1
POWERMV-IG-j48	450	163	178	493	73.4	73.5	73.4	72.6
POWERMV-IG-LOGISTIC	427	186	175	496	69.7	73.9	71.9	80.0
POWERMV-IG-SVM (Polykernal, C=0.1)	374	239	144	527	61.0	78.5	70.2	69.8
POWERMV-IG-SVM (Polykernal, C=1)	389	224	150	521	63.5	77.6	70.9	70.6
POWERMV-IG-SVM (Polykernal, C=1)	403	210	176	495	65.7	73.8	69.9	69.8
POWERMV-IG-SVM (RBFkernal, C=1 $\gamma= 0.01$)	374	239	151	520	61.0	77.5	69.6	69.3
POWERMV-IG-SVM (RBFkernal, C=1 $\gamma= 1$)	459	154	135	536	74.9	79.9	77.5	77.4
POWERMV-IG-SVM (RBFkernal, C=1 $\gamma= 10$)	382	231	88	583	62.3	86.9	75.2	74.6

Table 6.9 Comparison of classification models of Combined database built with different MLT estimated by five-fold cross-validation

Model	TP	FN	FP	TN	SEN	SPE	CA	AUC
COMB-CFS-kNN (k=1)	465	148	192	479	75.9	71.4	73.5	72.8
COMB-CFS-kNN (k=3)	487	126	183	488	79.4	72.7	75.9	81.3
COMB-CFS-kNN (k=5)	502	111	175	496	81.9	73.9	77.7	83.0
COMB-CFS-kNN (k=7)	505	108	173	498	82.4	74.2	78.1	84.1
COMB-CFS-kNN (k=9)	487	126	186	485	79.4	72.3	75.7	83.5
COMB-CFS-J48	426	187	168	503	69.5	75.0	72.4	44.5
COMB-CFS-LOGISTIC	443	170	158	513	72.3	76.5	74.5	82.9
COMB-CFS-SVM (Polykernal, C=0.1)	369	244	108	563	60.2	83.9	72.6	72.1
COMB-CFS-SVM (Polykernal, C=1)	440	173	138	533	71.8	79.4	75.8	75.6
COMB-CFS-SVM (Polykernal, C=10)	443	170	149	522	72.3	77.8	75.2	75.0
COMB-CFS-SVM (RBFkernal, C=1 $\gamma=0.01$)	331	282	103	568	54	84.6	70.0	69.3
COMB-CFS-SVM (RBFkernal, C=1 $\gamma=0.1$)	393	220	106	565	64.1	84.2	74.6	74.2
COMB-CFS-SVM (RBFkernal, C=1 $\gamma=1$)	461	152	133	538	75.2	80.2	77.8	77.7
COMB-CFS-SVM (RBFkernal, C=1 $\gamma=10$)	160	453	38	633	26.1	94.3	61.8	60.2

All models had the CA and AUC values higher than 60%. Algorithms like SVM and kNN were tuned to get a better result by varying parameters like C, γ , and k, respectively, made by the minimization of the misclassification rate of the 5-fold cross-validated training data. SVM generated a better result by using RBF kernel than a poly kernel with complexity parameter C=1, and $\gamma=1$ for all the databases and kNN produces a better result at k = 7 for all databases except PowerMV. In the case of PoweMV descriptors, kNN produces a better result at k = 5. The J48 and LOGISTIC models have poor performance as compared to SVM and kNN models.

An excellent binary classification model always results with high values of Sensitivity, Specificity, Accuracy, and Area under ROC. If anyone of the sensitivity and specificity is high, then accuracy will bias towards that highest value. From these tables, it can be identified

that some of the SVM models show a significant difference in sensitivity and specificity values. Models like them were not considered for further study. In some case, classification models which have AUC and CA values were not consistent, for example, DRAGON-CFS-SVM (RBF kernel, $C=1$ $\gamma=1$) model had higher CA, but lower AUC on the contrary COMB-CFS-kNN ($k=7$) model had lower CA and higher AUC. This is because the dataset we have taken is a slightly imbalanced one. In that case, it seems that kNN ($K=7$) classifier was more focused on sensitivity while the SVM (RBF kernel, $C=1$ $\gamma=1$) focused on specificity. Generally, sensitivity and specificity values contribute to the overall accuracy by different weighted, so it influences classification accuracy. However, here the differences between them were small (not significant) so that the models had a good predictivity to identify active and inactive compounds. Similar observations can be found in all other classifiers. In this case, AUC can provide more information than the overall accuracy so that it can be used as a better performance indicator.

The best model based on the DRAGON descriptor set was obtained with kNN ($k=7$), *i.e.*, DRAGON-CFS-kNN ($k=7$) and had CA and AUC of 76.6 and 82.7% respectively. Although the accuracy of this model became lower than DRAGON-CFS-SVM (RBF kernel, $C=1$, $\gamma=1$) model, it exhibits a very good value of AUC (82.7%). The least predictive classification model based on DRAGON descriptors was developed with SVM kernel Radial Basis Function with $C=1$ and $\gamma=0.01$ ($CA=68.0\%$ and $AUC=67.2\%$). Similarly, best models with OCHEM is one with kNN ($k=7$) ($CA=76.8\%$ and

AUC = 84.2%) and least predictive classifiers developed with SVM kernel Radial Basis Function with $C = 1$ and $\gamma = 0.01$ (CA = 69.8% and AUC 63.6%). The combined database also generated the best performing model with kNN ($k = 7$) while the least performing model with SVM (RBF kernel, $C=1$ $\gamma= 10$) having CA and AUC values is 61.8 and 60.2% respectively. PowerMV descriptors generated models were quite different from the model generated by other databases. The best model with PowerMV descriptors was obtained with kNN ($k = 5$) have CA and AUC of 76.6 and 82.4% respectively, and the bad model was obtained with SVM and kernel Radial Basis Function with $C = 1$ and $\gamma = 0.01$ (CA = 69.6% and AUC = 69.3%).

Out of 52 classification models generated, the 16 optimized models, *i.e.*, models with comparatively high CA and AUC, were further validated by the test sets. The detailed results of the test sets evaluation of 16 models were given in Table 6.10 (where kNN and SVM with the optimized parameter are taken). Except for models DRAGON-CFS-J48 and COMP-CFS-J48, all other models exhibit good predictive performance for test sets. The performance of POWERMV-IG-kNN ($k = 5$), OCHEM-CFS-kNN ($k = 7$), DRAGON-CFS-kNN ($k = 7$) and COMP-CFS-kNN ($k = 7$) models demonstrated slightly superior to that of other models. Figure 6.7 display the histogram showing CA of training and test set of four best-performing classifiers of each descriptor database estimated by five-fold cross-validation.

Table 6.10 Comparison of prediction accuracies of different classification models by the external test set

Model	TP	FN	FP	TN	SEN	SPE	CA	AUC
DRAGON-CFS-kNN (k=7)	128	45	39	109	74.0	73.6	73.8	82.3
DRAGON-CFS-J48	115	58	48	100	66.5	67.6	67.0	68.0
DRAGON-CFS-LOGISTIC	116	57	41	107	67.1	72.3	69.5	78.3
DRAGON-CFS-SVM (RBFkernel, C=1 $\gamma=1$)	117	56	27	121	67.6	81.8	74.1	74.7
OCHEM-CFS-kNN (k=7)	128	45	38	110	74.0	74.3	74.1	81.4
OCHEM-CFS-j48	119	54	30	118	68.8	79.7	73.8	73.6
OCHEM-CFS-LOGISTIC	111	62	29	119	64.2	80.4	71.7	79.0
OCHEM-CFS-SVM (RBFkernel, C=1 $\gamma=1$)	116	57	30	118	67.1	79.7	72.9	73.4
POWERMV-IG-kNN (k=5)	128	45	26	122	74.0	82.4	77.9	83.3
POWERMV-IG-j48	128	45	30	118	74.0	79.7	76.6	79.1
POWERMV-IG-LOGISTIC	114	59	31	117	65.9	79.1	72.0	80.8
POWERMV-IG-SVM (RBFkernel, C=1 $\gamma=1$)	112	61	24	124	64.7	83.8	73.5	74.7
COMB-CFS-kNN (k=7)	135	38	40	108	78.0	73.0	75.7	80.8
COMB-CFS-J48	112	61	36	112	64.7	75.7	69.8	70.0
COMB-CFS-LOGISTIC	117	56	34	114	67.6	77.0	72.0	81.5
COMB-CFS-SVM (RBFkernel, C=1 $\gamma=1$)	119	54	27	121	68.8	81.8	74.8	75.3

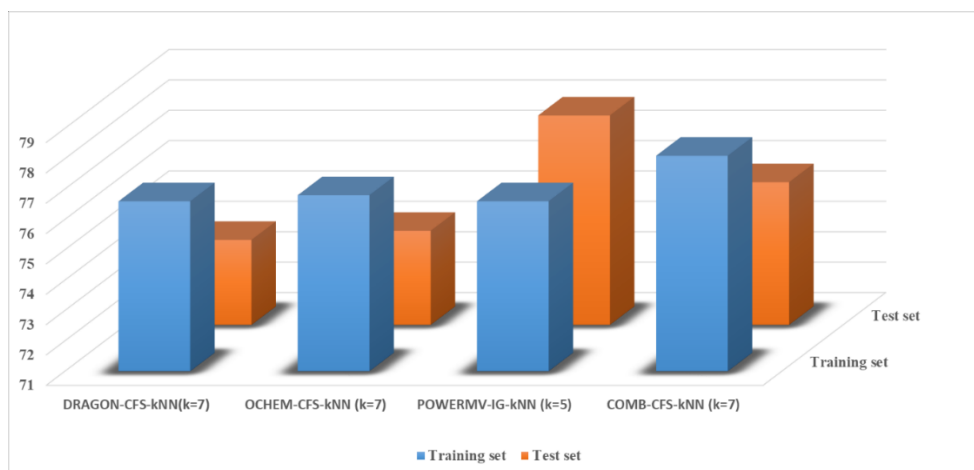


Fig. 6.7 CA of training and test set of best-performing classifiers of each descriptor database estimated by five-fold cross-validation.

In brief, among E-DRAGON, OCHEM, and Combined databases-based classifiers, the best model obtained is one with kNN

($k = 7$) having higher CA and AUC for both the training and test set. That is, kNN ($k = 7$) outperforms all other models except for PowerMV based models (Figure 6.7). Among PowerMV based models, kNN ($k = 5$) shows excellent performance. By comparing the performance of four databases, most performed training set COMB-CFS-kNN ($k = 7$) failed to predict the external test set more accurately on the other hand PowerMV database (POWERMV-IG-kNN ($k = 5$)) yielded the best results for both training and test set, this means descriptors from PowerMV has some significant influence on predicting 5-LOX activity. So model POWERMV-IG-kNN ($k = 5$) is the best model that shows excellent performance in internal and external validation. Our findings would seem to show that the best performing model could very profitable to be used for uncovering 5-LOX active compounds through virtual screening.

6.6.3. Model Validation Through Y-Scrambling

Y-Scrambling (Y-randomisation) [16] method is used to exclude the probability that the performance of our ML models might have happened by chance. The Y-vector (the label: active, inactive) of the 1284 compounds in the training set were reordered in N times to generate N number of the permuted training set. After that, attempts were made to run each of these permuted training sets (y-scrambled sets) using all the classifiers. A total of 10 randomization runs was performed for each database. The results of the Y-scrambling test for kNN classifiers are given as an example in Table 6.11. The average CA to the 5-LOX inhibitory potency of the Y-scrambled models is

only in the range of 47-52%. All other validation parameters of Y-scrambled models is given in Table 6.11, which are also far away from being of good value. That is, in all cases, the obtained random models have much lower prediction accuracy than the model based on the real data, genuineness of the kNN model that is not by chance. The same result is seen for the Y-scrambled models generated using other classifiers used in this study.

Table 6.11 Prediction accuracies of Y-scrambled models estimated by five-fold cross-validation

kNN model	No. of descriptors	Av. TP	Av. FN	Av. FP	Av. TN	Av. SEN	Av. SPE	Av. CA	Av. AUC
kNN-COMB-CFS (k=7)	99	263	350	281	390	42.9	58.1	50.8	0.49
kNN-COMB-IG (k=7)	1810	255	358	295	376	41.6	56.0	49.1	0.48
kNN-DRAGON-CFS (k=7)	70	291	322	324	347	47.5	51.7	49.7	0.50
kNN-DRAGON-IG (k=7)	942	261	352	269	402	42.6	59.9	51.6	0.50
kNN-OCHEM-CFS (k=7)	71	282	331	315	356	46.0	53.1	49.6	0.51
kNN-OCHEM-IG (k=7)	815	263	350	288	383	42.9	57.1	50.3	0.51
kNN-POWEMV-CFS (k=5)	18	261	352	324	347	42.6	51.7	47.4	0.47
kNN-POWERMV-IG (k=5)	54	286	327	318	353	46.7	52.6	49.8	0.50

6.7. Virtual Screening of e-Drug3D Database

6.7.1. Virtual Screening (VS)

The soul of virtual screening in this study is to find out potential leads with different scaffolds from massive molecular databases using the best performing QSAR model. Considering their pharmacological and toxicological profiles that are previously established, approved drugs are a very attractive and lucrative starting point for VS. Drug database screening is also used to identify compounds with polypharmacological properties (a drug that acts on

multiple targets). Polypharmacology is key to the rational design of the next generation of less toxic therapeutic agents. Therefore, we used the ' e-Drug3D ' [17] drug database for virtual screening. This drug database aims to provide free and ready-to-screen virtual collections of FDA-approved drugs and their commercially available substructures (fragments). This large and diverse collection of compounds has emerged as a significant natural input for various cheminformatic and virtual screening applications. The database e-Drug3D mirrors the current content of the US Pharmacopoeia of small drugs (molecular weight ≤ 2000), which contains 1822 molecular structures approved between 1939 and 2016 with a molecular weight of ≤ 2000 .

The PowerMV-IG-kNN (k=5) model was used to virtually screen the compounds from the e-Drug3D database. All compounds from the e-Drug3D database were pre-processed, as explained in section 7.2. Finally obtained 1460 unique compounds, which were then energy minimized by optimization process and submitted to the virtual screening. We found 43 potential hit candidates for 5-LOX inhibitors as a result of virtual screening, including zileuton (which is potent 5-LOX inhibitors). The name and SMILES of these 43 virtual hits are listed in Table 6.12.

Table 6.12 43 potential 5-LOX inhibitors identified through virtual screening

No	Screened Drug	Smiles
1	Moxifloxacin	<chem>[C@H]12[C@H](NCCC1)CN(C2)c1c(c2c(cc1F)c(=O)c(cn2C1CC1)C(=O)[O-])OC</chem>
2	Cabozantinib	<chem>c12c(cc(c(c1)OC)OC)c(ccn2)Oc1ccc(cc1)NC(=O)C1(CC1)C(=O)Nc1ccc(cc1)F</chem>
3	Elvitegravir	<chem>c1c(c(c(cc1)Cl)F)Cc1cc2c(cc1OC)n(cc(c2=O)C(=O)[O-])][C@@H](C(C)C)CO</chem>
4	Bosutinib	<chem>c12c(cc(c(c1)OCCCN1CCN(CC1)C)OC)c(c(cn2)C#N)Nc1c(cc(c(c1)OC)Cl)Cl</chem>
5	Bedaquiline	<chem>c1(c(cc2c(n1)ccc(c2)Br)[C@H]([C@](c1c2ccccc2ccc1)(CCN(C)C)O)c1ccccc1)OC</chem>
6	Canagliflozin	<chem>O1[C@H]([C@@H]([C@H]([C@@H]([C@H]1CO)O)O)O)c1ccc(cc1)C)Cc1ccc(s1)c1ccc(cc1)F</chem>
7	Eslicarbazepine Acetate	<chem>c12c(cccc1N(c1c([C@H](C2)OC(=O)C)cccc1)C(=O)N</chem>
8	Simeprevir	<chem>[C@H]12/C=C\C\CCCCN(C(=O)[C@H]3[C@H](C(=O)N[C@@]1(C)2)C(=O)NS(=O)(=O)C1CC1)[C@H](C3)Oc1cc(nc2c(c(ccc12)OC)C)c1nc(cs1)C(C)C)C</chem>
9	Apremilast	<chem>c12c(c(ccc1)NC(=O)C)C(=O)N(C2=O)[C@H](CS(=O)(=O)C)c1ccc(c(c1)OCC)OC</chem>
10	Belinostat	<chem>c1c(cccc1)NS(=O)(=O)c1cc(ccc1)/C=C/C(=O)NO</chem>
11	Ledipasvir	<chem>C1C[C@@H]2[C@H](N([C@H]1C2)C(=O)[C@H](C(C)C)NC(=O)OC)c1[nH]c2c(n1)ccc(c2)c1ccc2c(c1)C(c1c2ccc(c1)c1cnc([nH]1)[C@@H]1CC2(CN1C(=O)[C@@H](NC(=O)OC)C(C)C)CC2)(F)F</chem>
12	Dasabuvir	<chem>c12c(cc(cc1)NS(=O)(=O)C)ccc(c2)c1cc(cc(c1OC)C(C)(C)C)n1c(=O)[nH]c(=O)cc1</chem>
13	Brexipiprazole	<chem>c1cc2c(s1)cccc2N1CCN(CC1)CCCCOc1cc2c(cc1)ccc(=O)[nH]2</chem>
14	Eluxadoline	<chem>c1(c2ccccc2)[nH]c(nc1)[C@@H](N(C(=O)[C@H](Cc1c(cc(c1C)C(=O)N)C)N)Cc1cc(c(cc1)OC)C(=O)[O-])C</chem>
15	Grazoprevir	<chem>c12c(nc3c(n1)ccc(c3)OC)O[C@H]1CN(C(=O)[C@@H](NC(=O)O)[C@H]3[C@H](CCCC2)C3)C(C)(C)C)[C@@H](C1)C(=O)N[C@]1([C@@H](C1)C=C)C(=O)NS(=O)(=O)C1CC1</chem>
16	Velpatasvir	<chem>c1c(cccc1)[C@H](C(=O)N1C[C@H](C[C@H]1c1[nH]c(cn1)c1ccc2c(c1)COc1c2cc2c(c1)c1c(cc2)nc([nH]1)[C@@H]1CC[C@H](N1C(=O)[C@H](C(C)C)NC(=O)OC)C)COC)NC(=O)OC</chem>
17	Adapalene	<chem>O(c1c([C@]23C[C@@H]4C[C@H](C2)C[C@H](C3)C4)cc(cc1)c1cc2c(cc1)cc(cc2)C(=O)[O-])C</chem>
18	Asenapine	<chem>c12c(ccc(c1)Cl)Oc1c([C@@H]3[C@@H]2CN(C3)C)cccc1</chem>
19	Bosentan	<chem>S(=O)(=O)(Nc1nc(nc(OCCO)c1Oc1c(OC)cccc1)c1ncccn1)c1ccc(C(C)(C)C)cc1</chem>
20	Clobazam	<chem>Clc1cc2c(cc1)N(C)C(=O)CC(=O)N2c1ccccc1</chem>

Modeling Machine Learning Based QSAR

21	Cromolyn	<chem>c1(=O)c2c(oc(c1)C(=O)[O-])cccc2OCC(COc1c2c(=O)cc(oc2ccc1)C(=O)[O-])O</chem>
22	Flavoxate	<chem>O(CCN1CCCCC1)C(=O)c1c2oc(c(=O)c2ccc1)C)c1cccc1</chem>
23	Furosemide	<chem>Clc1c(S(=O)(=O)N)cc(c(NCc2occc2)c1)C(=O)[O-]</chem>
24	Gatifloxacin	<chem>Fc1c(N2C[C@H](NCC2)C)c(OC)c2n(C3CC3)cc(c(=O)c2c1)C(=O)[O-]</chem>
25	Halofantrine	<chem>Clc1c2cc([C@H](O)CCN(CCCC)CCCC)c3c(c2cc(Cl)c1)cc(c3)C(F)(F)F</chem>
26	Irinotecan	<chem>c1(ccc2c(c1)c(c1c(n2)c2n(C1)c(=O)c1c(c2)[C@](C(=O)OC1)(CC)O)CC)OC(=O)N1CC[C@H](CC1)N1CCCCC1</chem>
27	Lapatinib	<chem>Clc1cc(Nc2nnc3c2cc(c2oc(CNCCS(=O)(=O)C)cc2)cc3)ccc1OCc1cc(F)ccc1</chem>
28	Levofloxacin	<chem>Fc1c(N2CCN(CC2)C)c2OC[C@H](n3c2c(c1)c(=O)c(c3)C(=O)[O-])C</chem>
29	Masoprocol	<chem>Oc1cc(C[C@H])([C@H](C2cc(O)c(O)cc2)C)C)ccc1O</chem>
30	Mefloquine	<chem>FC(F)(F)c1nc2c(c([C@H](O)[C@H]3NCCCC3)c1)cccc2C(F)(F)F</chem>
31	Mesoridazine	<chem>S1c2c(N(CC[C@H]3N(CCCC3)C)c3c1cccc3)cc([S@](=O)C)cc2</chem>
32	Miconazole	<chem>c1(c(cc(cc1)Cl)Cl)[C@H](Cn1cncc1)OCc1c(cc(cc1)Cl)Cl</chem>
33	Nedocromil	<chem>o1c2c(c3n(CC)c(cc(=O)c3cc2c(=O)cc1C(=O)[O-])C(=O)[O-])CCC</chem>
34	Novobiocin	<chem>c1(c(c2c(oc1=O)c(c(O[C@H]1[C@H]([C@H]([C@H](C(O1)(C)C)OC)OC(=O)N)O)cc2)C)O)NC(=O)c1cc(c(cc1)O)C=C(C)C</chem>
35	Ofloxacin	<chem>Fc1c(N2CCN(CC2)C)c2OC[C@H](n3c2c(c1)c(=O)c(c3)C(=O)[O-])C</chem>
36	Propofol	<chem>Oc1c(C(C)C)cccc1C(C)C</chem>
37	Quetiapine	<chem>S1c2c(C(=Nc3c1cccc3)N1CCN(CC1)CCOCCO)cccc2</chem>
38	Sitagliptin	<chem>FC(F)(F)c1n2CCN(Cc2nn1)C(=O)C[C@H](N)Cc1c(F)cc(F)c(F)c1</chem>
39	Sulfoxone	<chem>S(=O)(=O)(c1ccc(NC[S@](=O)O)cc1)c1ccc(NC[S@](=O)O)cc1</chem>
40	Tazarotene	<chem>S1CCC(c2c1ccc(c2)C#Cc1ncc(cc1)C(=O)OCC)(C)C</chem>
41	Valrubicin	<chem>FC(F)(F)C(=O)N[C@H]1C[C@H](O[C@H]2C[C@@](O)(C)c3c2c(O)c2c(c3O)C(=O)c3c(C2=O)c(OC)ccc3)C(=O)COC(=O)CCCCO[C@H]([C@H]1O)C</chem>
42	Vilazodone	<chem>c12c(cc(cc1)N1CCN(CC1)CCCCc1c3c(ccc(c3)C#N)[nH]c1)cc(o2)C(=O)N</chem>
43	Zileuton	<chem>s1c([C@H](N(O)C(=O)N)C)cc2c1cccc2</chem>

6.7.2. Molecular Docking Analysis

Docking simulations of the top 43 compounds were carried out to reduce the number of potential hits. The interaction between protein 5-LOX and virtual hits was investigated using this process in order to find compounds with the best docking score. For this, a stable human 5-LOX crystal structure with a PDB ID 3O8Y at 2.4-angstrom resolution was used. Autodock Vina [18] software was used to carry out molecular docking analysis. The protein was prepared with AutoDock Tools (ADT) by removing water molecules and by adding polar hydrogens, appropriate charge, *etc.* Because there was no co-crystal ligand for 5-LOX protein, prediction of the size and spatial orientation of the ligand-binding sites was performed and explained in Chapter 3. After, molecular docking, Protein-ligand complexes were visualized and analyzed using three different molecular modeling software Autodock tool 1.5.6 [19], Chimera [20], and PyMol [21].

Out of 43 compounds, eight compounds having binding affinity values greater than -4 kcal/mol were identified as potential lead compounds for 5-LOX inhibition. Chemical structure and binding affinity value of these eight virtual hits and zileuton are shown in Figure 6.8.

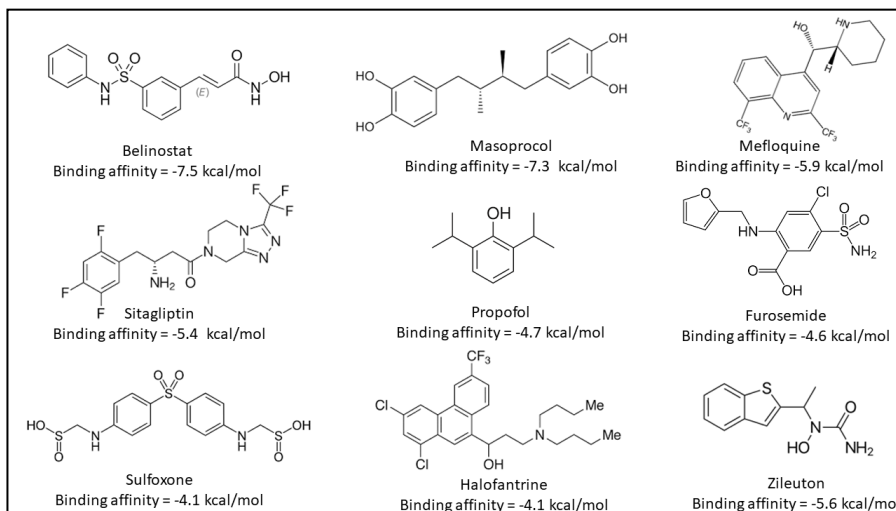


Fig. 6.8 Chemical structures of top eight virtual hits with best 5-LOX binding affinity value along with established 5-LOX inhibitor zileuton.

It is Evident from the docking score that compounds belinostat, masoprocol, mefloquine, and sitagliptin shows higher binding affinity in comparison to all the screened compounds towards 5-LOX. Masoprocol is a meso form of a well-known 5-LOX inhibitor Nordihydroguaiaretic acid (NDGA). Belinostat and masoprocol showed almost 1.3 folds higher affinity than Zileuton (Binding affinity = -5.6 kcal/mol) while mefloquine and Sitagliptin display similar binding affinity to the reference inhibitor zileuton. From the docking result, it can be concluded that compounds belinostat, Masoprocol, Mefloquine, and Sitagliptin can form high-affinity candidates against 5-LOX. These top 4 compounds are selected to investigate the nature of the molecular level interaction of the ligand with 5-LOX protein. 2D Binding interaction of belinostat, Masoprocol, Mefloquine, and Sitagliptin with 5-LOX binding pocket are shown in Figures 6.9 A, B,

C, and D, respectively, based on the ligand's conformation corresponding to the lowest binding free energies calculated by Vina and their 3D views, are shown in Figure 6.10 A, B, C, and D, respectively.

The binding mode between belinostat and 5-LOX reveals that the inhibitory mechanism of compounds is almost similar to the typical inhibitory mechanism of a suitable inhibitor for 5-LOX, which should have a polar head and tail and a hydrophobic body. The -NH group of an O=C-NH-OH part in the polar head of belinostat interact with the hydrophilic portion of 5-LOX by forming hydrogen bonds with His 372 and Asn 407 amino acid residue and -NH group at the tail portion also interact with 5-LOX through the formation of a hydrogen bond with Gln 363. This compound form hydrophobic interactions with the protein through its non-polar styrene part with residues Leu 368, Ile 415, Phe 421 Leu 414, Leu 420, and Leu 607.

Similarly, masoprocol also situated in the hydrophobic channel of 5-LOX created by Leu 368, Phe 421, Leu 607, Ala 410, and Phe 177 running by the catalytic iron. Masoprocol is a highly symmetrical compound with a 1, 2-dihydroxy benzene on either side. These -OH groups at the head portion interact with a polar amino acid, like His 372 and His 367, and -OH group at the tail portion interact with an amino acid, like Trp 599, and form a hydrogen bond with His 600. Even though both belinostat and masoprocol displayed an excellent fit into the hydrophobic binding pocket, belinostat shows a higher binding affinity than masoprocol this may be because belinostat stabilizes

position close to the catalytic center by forming a hydrogen bond with His 372 which is one of the amino acids coordinating the catalytic iron.

Mefloquine, a small compound without a hydrophobic body, is also situated in the same hydrophobic channel where its quinoline ring lines up relatively well, and its –OH group form hydrogen bond with one of amino acid coordinating the catalytic iron, His 367. The formation of this H-bond stabilized position close to the catalytic center could explain why mefloquine is having a higher binding affinity than zileuton.

The fourth one, 'Sitagliptin,' is a compound with a lot of electronegative substituents throughout the body, is also occupied in the hydrophobic channel by forming hydrogen bonds with Asn 425 and Ala 424. However, its binding affinity is weak as compared to Belinostat, Masoprocol, and Mefloquine.

In conclusion, the first two ligands, such as Belinostat and Masoprocol, occupied an entire portion of the active site cavity and preventing substrate access to the iron atom; therefore has a good binding affinity. In the case of mefloquine, it leaves a large opening around the iron atom by occupying at the end of the cavity; this reduces its affinity value. The fact 'hydrophobic interactions with 5-LOX determine the inhibitor's binding affinity' is well applied for sitagliptin, this compound cannot occupy in the binding cavity of the protein much effectively as compared to others because of its polar nature and so have a lower binding affinity as compared to others.

Nevertheless, compared to other virtual hits, these four compounds show better affinity value and could act as good leads against 5-LOX.

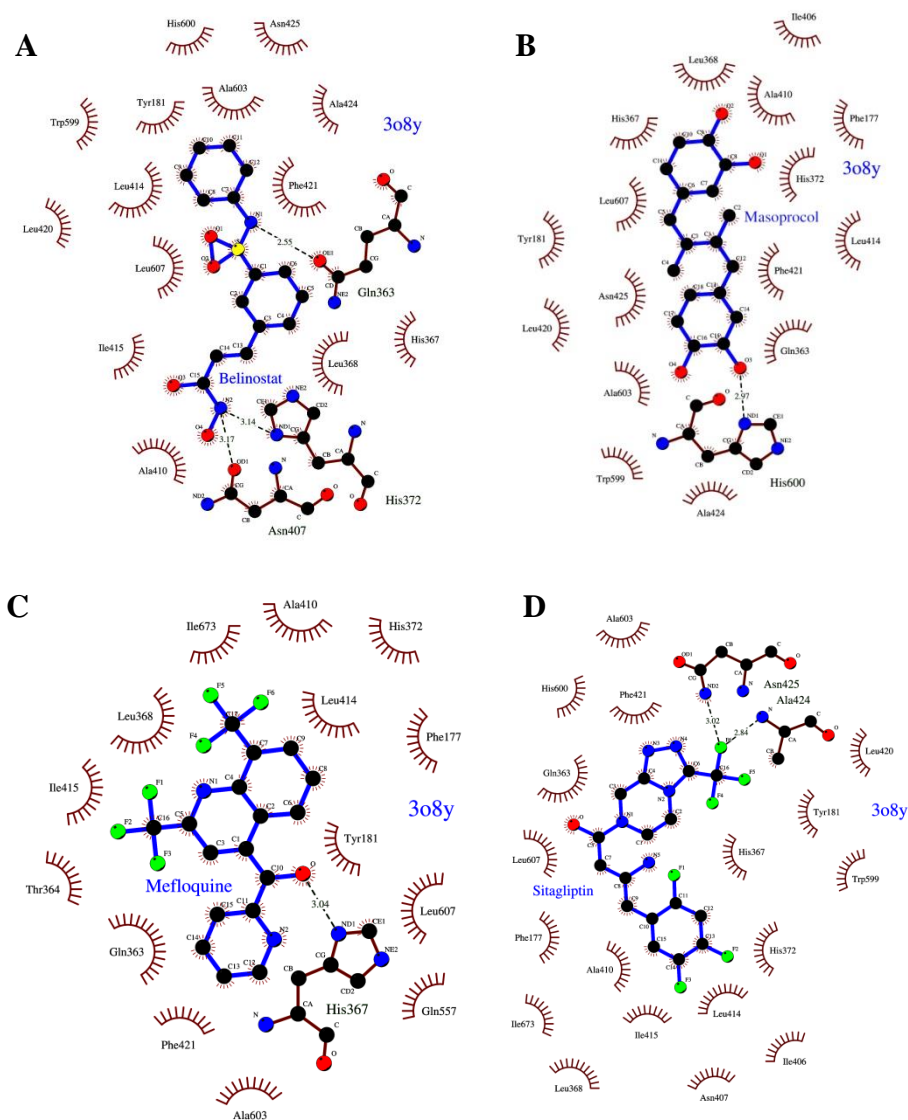


Fig. 6.9 2D view of the binding interaction of four virtual hits with 5-LOX. A - belinostat, B - masoprocol, C - mefloquine, and D - sitagliptin. H-bond can be seen in the dotted line.

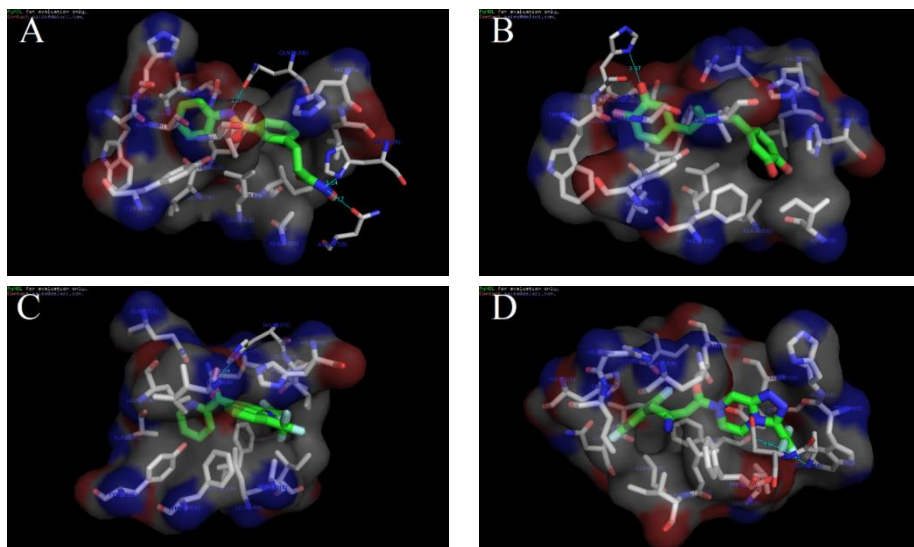


Fig. 6.10 3D view of Binding interaction of four virtual hits with 5-LOX. A - belinostat, B - masoprocol, C - mefloquine, and D - sitagliptin.

6.8. Conclusion

In this Chapter, QSAR classification models were developed for predicting the 5-LOX inhibitors and non-inhibitors by using four sets of most relevant descriptors extracted from four descriptor datasets such as DRAGON, OCHEM, PowerMV, and Combined. Two efficient feature selection methods, such as CFS and IG, were used to remove noisy descriptors. The CFS approach is more effective for eliminating redundant descriptors than using IG for all databases, except PowerMV. The best classification models for OCHEM, E-DRAGON, and Combined dataset were obtained using kNN ($k = 7$) ML algorithm from CFS selected descriptors. While the model obtained with IG selected descriptors and kNN ($k = 5$) algorithm outperforms all other

models that are based on PowerMV. Among the 52 ML model constructed, PowerMV-IG-kNN (k = 5) model gave better predictive results. The best model proposed here achieved an overall accuracy of 76.6% for the training set using a 5-fold CV procedure and an overall accuracy of 77.9% for the test set.

Furthermore, the best performing model, PowerMV-IG-kNN (k = 5), has used to virtually screen the compounds from the e-Drug3D database. As a result, 43 potential hits were identified. Also, the kNN method has been identified as the best tool for the screening of large compound databases. Furthermore, molecular docking-based virtual screenings were also performed to rank these 43 hits and identified four hits such as belinostat, masoprocol, mefloquine, and sitagliptin with high potential activity against 5-LOX protein. Among them, compounds belinostat and masoprocol occupied in the entire portion of the active site cavity, thereby preventing substrate access to the iron atom and therefore showing higher binding affinity than zileuton. The remaining two, mefloquine and sitagliptin, have shown a comparable binding affinity to zileuton. Thus, we successfully identified four potential lead compounds as 5-LOX inhibitors using a combination of different *in silico* techniques, which can be further evaluated by biological studies.

Reference

- [1] G. Eren, A. MacChiarulo, E. Banoglu, From molecular docking to 3D-quantitative structure-activity relationships (3D-QSAR): Insights into the binding mode of 5-Lipoxygenase inhibitors, *Mol. Inform.* 31 (2012) 123–134. doi:10.1002/minf.201100101.
- [2] M.A. Babu, N. Shakya, P. Prathipati, S.G. Kaskhedikar, A.K. Saxena, Development of 3D-QSAR Models for 5-Lipoxygenase Antagonists: Chalcones, 10 (2002) 4035–4041.
- [3] R.D. King, J.D. Hirst, M.J.E. Sternberg, New approaches to QSAR: Neural networks and machine learning, *Perspect. Drug Discov. Des.* 1 (1993) 279–290. doi:10.1007/BF02174529.
- [4] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, *Drug Discov. Today.* 23 (2018) 1538–1546. doi:https://doi.org/10.1016/j.drudis.2018.05.010.
- [5] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.* 40 (2012) D1100–D1107. doi:10.1093/nar/gkr777.
- [6] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An Open chemical toolbox, *J. Cheminform.* 3 (2011) 1–14. doi:10.1186/1758-2946-3-33.
- [7] M. Sud, MayaChemTools: An Open Source Package for Computational Drug Discovery, (2016). doi:10.1021/acs.jcim.6b00505.
- [8] I. V Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E. V Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V. V Prokopenko, Virtual Computational Chemistry Laboratory – Design and Description, *J. Comput. Aided. Mol. Des.* 19 (2005) 453–463. doi:10.1007/s10822-005-8694-y.
- [9] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V Prokopenko, V.Y. Tanchuk,

- R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I.I. Baskin, V.A. Palyulin, E. V Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I. V Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, *J. Comput. Aided. Mol. Des.* 25 (2011) 533–554. doi:10.1007/s10822-011-9440-2.
- [10] K. Liu, J. Feng, S.S. Young, PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation, *J. Chem. Inf. Model.* 45 (2005) 515–522. doi:10.1021/ci049847v.
- [11] T. Sander, J. Freyss, M. Von Kor, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *J. Chem. Inf. Model.* 55 (2015) 460–73. doi:10.1021/ci500588j.
- [12] R. Guha, Exploring Structure-Activity Data Using the Landscape Paradigm, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2 (2012) 10.1002/wcms.1087. doi:10.1002/wcms.1087.
- [13] R. Guha, J.H. Van Drie, Structure–Activity Landscape Index: Identifying and Quantifying Activity Cliffs, *J. Chem. Inf. Model.* 48 (2008) 646–658. doi:10.1021/ci7004093.
- [14] A. Lavecchia, Machine-learning approaches in drug discovery: Methods and applications, *Drug Discov. Today.* 20 (2015) 318–331. doi:10.1016/j.drudis.2014.10.012.
- [15] J. Platt, Fast Training of Support Vector Machines using Sequential Minimal Optimization, *Adv. Kernel Methods.* (1998) 41–64.
- [16] C. Rücker, G. Rücker, M. Meringer, γ -Randomization and Its Variants in QSPR/QSAR, *J. Chem. Inf. Model.* 47 (2007) 2345–2357. doi:10.1021/ci700157b.
- [17] E. Pihan, L. Colliandre, J.-F. Guichou, D. Douguet, e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-

- based drug design, *Bioinformatics*. 28 (2012) 1540–1541. doi:10.1093/bioinformatics/bts186.
- [18] O. Trott, A. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading., *J. Comput. Chem.* 31 (2010) 455–461. doi:10.1002/jcc.21334.
- [19] G. Morris, R. Huey, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J. Comput. Chem.* 30 (2009) 2785–2791. doi:10.1002/jcc.21256.AutoDock4.
- [20] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF Chimera - A visualization system for exploratory research and analysis, *J. Comput. Chem.* 25 (2004) 1605–1612. doi:10.1002/jcc.20084.
- [21] W.L. Delano, The PyMOL Molecular Graphics System, (2002).

7

DOCKING-BASED VIRTUAL SCREENING OF SMALL MOLECULE INHIBITORS

7.1. Introduction

Up to this Chapter, we have explained the various regression and classification based predictive models of 5-LOX inhibitors developed by using smooth SAR in the structure-activity landscape of 5-LOX inhibitor. But scientists are always fascinated to identify novel 5-LOX inhibitors. The continued expansion of virtual chemical space that opens up the possibility of vast chemical scaffoldings, among them, one or few compounds may have therapeutic potential against 5-LOX protein. Mining or reducing the chemical space to find out potential hits by the computational method is now the preliminary procedure for lead identification in rational drug discovery. From several computational techniques, Virtual Screening (VS) of a series of compounds has become an initial approach to reduce the virtual space of chemicals up to a manageable level [1,2]. Therefore, only a limited number of compounds have to be synthesized and experimentally tested for their biological activity. Also, it is well known that the primary purpose of VS procedures is to enrich the subsets of molecules that are active while discarding compounds that are to be inactive by

scoring [3]. Although experimental methods are robust, they are also costly, time-consuming, and difficult, particularly for screening large compound databases. One of the main VS methods, 'Structure-Based Virtual Screening (SBVS),' assists in finding out the protein interactions with a ligand at the atomic level. It allows us to identify the behavior of molecules in the binding pocket of the target protein [4]. The most popular methodology for performing a virtual screening includes the flexible docking algorithm, in which the ligand is placed in the receptor by conformational sampling techniques, and a type of measurement function is used to achieve a prediction of binding free energy [5].

Several publications have appeared in recent years documenting varieties of compounds that have high potency towards 5-LOX protein [6–8]. The different class of 5-LOX inhibitors (redox, non-redox, iron chelators, and FLAP inhibitors) provides different structural scaffold. Therefore, virtual screening of the large and chemically diverse databases without accounting the similarity toward any of the representative compounds from each class can uncover the possibility of the presence of a more potent 5-LOX inhibitor with a novel scaffold. Besides, large compound libraries have not been screened yet for discovering 5-LOX inhibitors. From these two perspectives, structure-based virtual screening studies, especially molecular docking based virtual studies, may present the best way to screen large databases for the identification of 5-LOX inhibitors with reduced cost. In the absence of experimental information of 5-LOX co-crystallized ligand and its biological confirmation, docking programs are necessary for ligand positioning. And selecting a reliable program

is crucial for such optimization to be successful. Nowadays, because of a range of docking programs that are accessible to the science community, a thorough knowledge of each docking program's benefits and constraints is essential to perform more sensible docking and docking-based virtual screening.

In this Chapter, we have conducted a comparative assessment of four commonly used docking programs in support of our attempts in virtual screening of novel 5-LOX inhibitors. We have selected the docking programs, Glide, LeDock, DOCK 6, and AutoDock Vina for this purpose. Additionally, a consensus model that combines all of the four docking programs is also developed. The assessment was based on the scoring reliability of four individual and one consensus docking scoring functions to recognize the known active inhibitors seeded in a random library of "drug-like" compounds. This method helped to identify a precise program for screening potential 5-LOX inhibitors. And the best scoring docking algorithm can then be used to carry out virtual screening of the ZINC 15 database to identify potential 5-LOX inhibitors that could be the next lead.

7.2. Screening Database

ZINC 15 is a free database of commercially available compounds for virtual screening [9]. It contains over 230 million purchasable compounds in ready-to-dock, 3D formats. For this study, we have downloaded a subset of 2.7 million in stock ZINC 15 lead like molecules with zero charges in pdbqt format, which can be directly used for docking with autodock vina. The lead-like molecules are selected based on criteria properties that are: molecular weight

between 250 and 350 g/mol, predicted partition constant (xLogP) ≤ 3.5 , and the number of rotatable bonds (RBs) ≤ 7 . ZINC 15 Tranche of all chemicals and a subset of chemicals selected for this study is respectively shown in Figure 7.1A and B. For docking with other software, compounds in pqbqt format are converted to Mol2 format using Open Babel utility [10].

A		Molecular Weight (up to, Daltons)											Totals, by LogP
		200	250	300	325	350	375	400	425	450	500	>500	
LogP (up to)	-1	31,520	267,142	844,459	1,877,402	3,668,610	822,692	255,392	56,712	42,782	23,759	5,540	7,886,010
	0	154,788	1,320,357	4,219,562	8,471,167	16,942,899	3,679,181	1,651,513	467,063	380,132	209,867	4,053	37,500,582
	1	414,273	3,915,351	13,647,007	25,235,841	51,379,682	12,665,806	6,959,986	2,568,795	2,189,894	1,206,882	8,718	120,192,235
	2	555,197	6,231,388	26,283,313	46,471,275	88,330,050	28,983,131	18,747,494	8,741,886	7,748,815	4,550,854	22,897	236,666,300
	2.5	214,735	3,036,363	15,286,260	26,740,405	55,786,220	20,475,678	15,102,617	8,256,989	7,745,534	4,762,980	24,012	157,431,793
	3	125,492	2,358,111	13,704,358	24,309,245	51,334,377	22,333,418	18,228,180	11,103,726	10,986,050	6,968,806	39,378	161,491,141
	3.5	57,244	1,528,693	10,472,005	18,997,194	40,816,610	21,458,732	19,668,020	13,319,769	13,704,019	9,066,863	64,156	149,153,305
	4	18,215	735,658	6,331,494	10,051,185	15,490,949	15,830,775	18,510,747	13,888,572	15,182,916	10,473,721	95,178	106,609,408
	4.5	2,275	222,783	2,971,590	6,100,024	10,414,291	11,243,146	14,716,486	12,430,134	14,587,078	10,654,985	131,612	83,474,404
	5	94	34,151	885,819	2,762,724	5,683,659	6,605,991	10,092,547	9,229,001	11,854,934	9,234,778	160,182	56,543,880
	>5	28	890	45,622	178,848	555,480	1,230,946	2,067,330	2,641,238	2,957,334	2,471,283	817,747	12,966,746
Totals, by Weight		1,573,861	19,650,885	94,691,489	171,195,310	340,402,827	145,329,496	126,000,312	82,703,885	87,379,488	59,624,778	1,373,473	1130M Substances 1.7K Tranches

B		Molecular Weight (up to, Daltons)											Totals, by LogP
		200	250	300	325	350	375	400	425	450	500	>500	
LogP (up to)	-1	1,000	1,000	4,901	2,521	2,330	1,117	1,112	1,000	1,000	1,000	1,000	9,752
	0	15,000	15,000	20,189	13,548	14,387	1,000	1,000	1,000	1,000	1,000	1,000	48,124
	1	38,000	47,000	95,466	70,768	75,829	48,000	37,310	35,210	30,000	30,000	30,000	242,063
	2	88,000	111,000	249,868	207,766	154,078	138,000	90,000	82,400	65,000	61,000	61,000	611,712
	2.5	138,000	165,000	184,539	167,482	223,298	143,000	92,000	82,887	65,000	61,000	61,000	575,319
	3	160,000	175,000	182,044	173,561	266,904	159,000	100,000	109,000	87,000	88,000	88,000	622,509
	3.5	180,000	175,000	147,211	160,083	242,579	207,000	145,000	140,000	110,000	100,000	100,000	549,873
	4	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	0
	4.5	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	0
	5	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	0
	>5	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	180,000	0
Totals, by Weight		0	0	884,216	795,729	979,405	0	0	0	0	0	0	2.7M Protomers 332 Tranches

Fig. 7.1 ZINC 15 Tranche of chemical libraries of A) all chemicals and B) selected subset. Physical-chemical space was split into 11 hydrophobicity-polarity bins, calculated logP values given in vertical direction, and the molecular weight is presented in a horizontal direction.

7.3. Protein Selection and Preparation

The stable 3D structure of 5-LOX protein with a PDB ID 3O8Y has been selected as the target for this study. Chapter 3 provides a detailed description of the reasons and significance of choosing this protein structure (3O8Y) for molecular docking. The 3D structure of the protein is shown in Figure 7.2. In its native state, the protein structure obtained from the PDB was not acceptable for molecular docking. Therefore, it was essential to optimize, refine, and minimize the protein. For each docking program, protein is prepared separately with its graphical interface. Details of each method and software for protein preparations were presented in section 7.4.

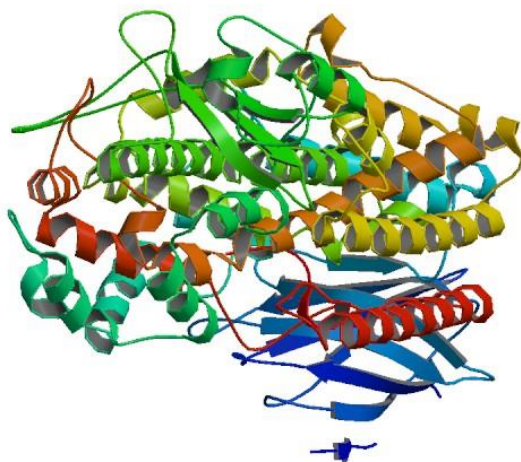


Fig. 7.2 3D structure of 5-LOX enzyme with a PDB ID 3O8Y.

7.4. Molecular Docking Programs

Identification of actives enriched subset of compounds from a big, chemical varied library based on predicted interaction with a target binding site is the primary objective of the docking-based VS program. A fundamental understanding of the benefits and limitations of each docking program is mandatory to conduct more reasonable studies on docking and docking-based VS. Previous research has documented several comprehensive evaluations of docking programs on a diverse set of protein-ligand complexes before virtual screening [11–13]. In this study, four popular docking algorithms such as DOCK 6, Glide XP, AutoDock Vina (exhaustiveness = 8), and LeDock were used to assess the prediction accuracy of ligand-binding poses and discrimination capability of docking-based VS. Each molecular docking methods provide its scoring function, which are mathematical functions used to approximately predict the binding affinity between protein and ligand after they have been docked. A consensus scoring model is also generated, which integrates the scores from the above-mentioned docking programs that may outperform individual programs in terms of VS enrichment. Because of the variability in the performance of the different score functions, the rates of enrichment are reduced by the blind choice of scoring functions for consensus scoring. Besides, the four scoring functions that we chose were independent of each other. It is reasonable to expect that an effective consensus scoring scheme would combine complementary scoring functions rather than highly correlated.

So, in this study, we have evaluated the capabilities of 4 individual scoring and a consensus scoring model to predict the ligand binding poses (sampling power) and rank the binding affinities (scoring power). The docking and scoring programs included here are routinely used for VS, and most of them are widely available to academic research groups. The detailed description of the docking algorithms used for this study is given below.

1. AutoDock Vina

AutoDock Vina [14] is an open-source docking program that uses a sophisticated gradient optimization method in its local optimization procedure. A protein preparation wizard 'AutoDockTools (ADT)' [15,16] provided by script research institutes have used for protein refinement. This process includes format conversion, removal of water, the addition of hydrogens, assignment of Gasteiger charges, and clean-up of unwanted elements. Finally, the dimension of the grid box is found and assigned to $20 \times 20 \times 25$ Å with center at $-8.374, 66.379, -1.009$ for x, y, and z, respectively. The docking scores were calculated by the default scoring function, and the best docking score for each molecule was saved.

2. Glide

The most popular docking algorithm 'Glide' [17,18] uses a series of hierarchical filters to approximate systematic search of positions, orientations, and conformations of the ligand in the receptor-binding site. The protein preparation wizard module in Schrödinger

2018 helps to protein preparation and refinement. This process includes adding hydrogens and disulfide bridges, removing crystallographic waters and ions, fixing bond orders, assigning partial charges with the OPLS force field. Initially, H-bond was optimized, then the whole protein structure was allowed to relax, and subsequently, the receptor protein was minimized by applying the OPLS 2005 force field. The binding box was constructed on the binding cleft that is identified by the SiteMap with the size of $10 \times 10 \times 10 \text{ \AA}$ generated by using the Receptor Grid Generation component of Glide. For the docking calculations, extra precision (XP) scoring functions [19] of Glide is used.

3. DOCK 6

The first introduced docking program 'DOCK' by Irwin "Tack" Kuntz's Group [20] uses geometric algorithms to predict the binding modes of small molecules [21–23]. The flexibility of ligand is accounted for by using an algorithm called anchor and grow. Two versions of the docking program are actively developed, which is DOCK 6 and DOCK 3, but the current study uses the DOCK 6 [24] program for molecular docking analysis. The protein preparation has been carried out by removing nonpolar hydrogen and adding Gasteiger charges, *etc.*, using Chimera. The molecular surface was then generated using the DMS program in the DOCK 6 suite with a probe radius of 1.4 \AA . The negative binding site space was defined using the SPHGEN program from the molecular surface file input. Contact, energy, and bump grid files were generated with grid_spacing

argument set to 0.3 Å. The spheres were selected within 6 Å from the ligand, and a 5 Å box margin was employed for the energy grids. Anchor and grow docking were performed with default parameters, except for max_orientations, which was increased from 1000 to 2000. Van der Waals atom definitions were taken from vdw_AMBER_parm99.defn file included with the DOCK 6 installation tree. Grid score was used for the docking score calculation.

4. LeDock

The 'LeDock' docking program [25] is based on a combination of simulated annealing and evolutionary optimization algorithm of ligand pose and its rotatable bonds, using a hybrid scoring scheme derived from prospective virtual screening campaigns [11]. LeDock is free of charge for educational use maintained by the Lephar Research Group. Protein preparation has been done by using ADT as same as that of in Vina docking. Binding pocket is defined by x_{\min} , y_{\min} , and z_{\min} with a value of -18.374, 56.379, and -13.509, respectively, while x_{\max} , y_{\max} , and z_{\max} with a value of 1.626, 76.379 and 11.491 respectively. Docking scores were calculated by the default scoring function.

5. Consensus Scoring (CS)

There are many scoring functions in existence, and their performance varies from case to case. So, finding better scoring methods are still a major goal for the researchers who are working on structure-based drug design. 'Consensus scoring' [26] is a strategy that

can be used to improve the predictive power of the molecular docking. CS involves formulating a new score by combining multiple individual docking scores from different docking programs, and the resultant score may largely reduce the false positives in virtual library screening, and hence, the hit rates were improved. Nevertheless, some studies have shown that consensus ranking does not surpass the best individual scoring function [27,28]. Because of this, we need to check the fate of Consensus scoring in the virtual screening of 5-LOX inhibitors. The binding scores provided by the different scoring functions are typically given in different units, it is almost impossible to compute consensus scores by merely summing up the binding scores determined by each of the individual scoring functions. Furthermore, merely scaling the scores from the methods do not adequately account for the variability and dynamic range of the different techniques. Therefore, we scaled the binding scores of each scoring function to unit variance and centered. The Z-scaled scoring function values (ZScore) are computed by Equation 7.1:

$$Zscore = \frac{f_i - \mu}{\sigma} \quad (7.1)$$

Where f_i is the scoring value of i^{th} ligand in the database of a certain scoring function, μ is the mean s , that is the mean score of all the compounds in the database, and σ is the standard deviation of this scoring function obtained from the method. The final score is the average of the scaled-score among all of the scoring functions.

7.5. Performance Evaluation of Docking Programs

7.5.1. Actives and Decoys Selection

The performance of docking programs was accessed by the capability of the program to distinguish between known actives and decoys for the 5-LOX crystal structure. The decoys are compounds that are chemically distinct from active compounds but have resemblance in the physical properties of the same so that they are likely to be non-binders. The 11 clinically approved known antagonists of 5-LOX were retrieved from the Drug bank database and considered as actives. The chemical structure of the actives (5-LOX inhibitors) used in this study is given in Figure 7.3.

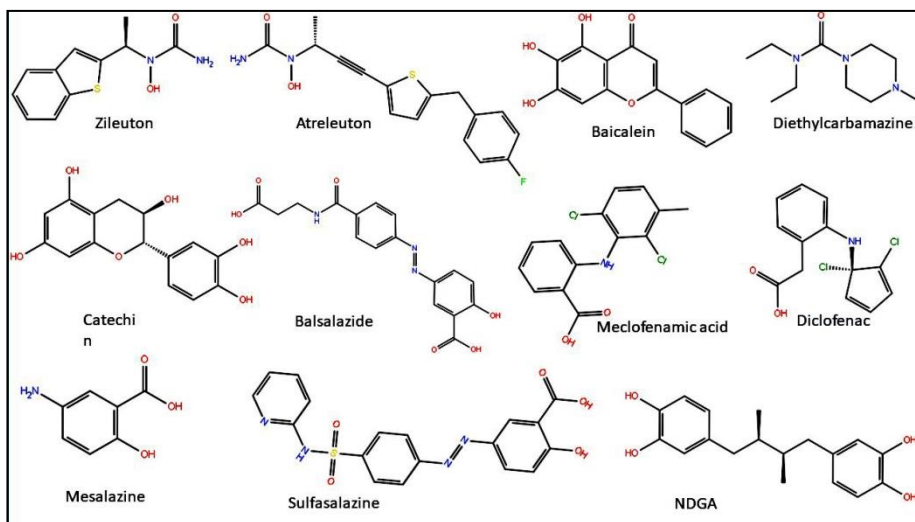


Fig. 7.3 Chemical structure of the actives (5-LOX inhibitors) used in this study.

The inactive compounds, also known as 'decoys' were generated from an online decoy database named as Directory of Useful

Decoy (DUD-E), which contains 550 compounds, and also from Schrodinger's universal decoy set which contains 1000 compounds. DUD-E decoys have been retrieved as SMILE format, which was then converted to mol2 format using Open Babel software. The Schrödinger drug-like decoys set consisted of 1000 drug-like compounds with an average molecular weight of 400 Daltons and were downloaded as a 3D SD file from the Schrödinger website. This collection of ligands was created by selecting 1000 ligands from a one million compound library that was chosen to exhibit "drug-like" properties [17,18]. All actives and decoys were prepared using the LigPrep module in Schrodinger by adding 3D coordinates. Two validation sets are created in which one contains 11 actives plus DUD-E decoys, and another one contains 11 actives plus Schrödinger decoys. Then, in order to evaluate the screening power and to distinguish known antagonists from decoys, all the molecules in the validation data sets were docked into the 5-LOX crystal structure and ranked by the docking scores.

7.5.2. Evaluation Metrics

The performance of virtual screening methods can be evaluated by different metrics including Enrichment Factors (EFs) [18], Receiver Operating Characteristics (ROC) curves [29], the Area Under the ROC Curve (ROC AUC) [29], the Boltzmann-Enhanced Discrimination of ROC (BEDROC) [30] and the Robust Initial Enhancement (RIE) [31]. The most commonly used methods are ROC and EF. Enrichment factor at 1% was computed using Equation 7.2:

$$EF1 = \frac{a/n}{A/N} \quad (7.2)$$

Where n = number of compounds in 1% of the database, a = number of actives in the top scoring 1% of the database, A = number of actives in the database, N = total number of compounds in database. Higher EF values indicate more actives found within a defined "early recognition" fraction of the ordered list relative to a random distribution. Enrichment was calculated at 1, 2, 5, 10, and 15% of the total decoy set and is given as histogram in Figure 7.4. From Figure 7.4, it is understood that force-field based scoring function like DOCK 6 and empirical scoring function like Glide XP performed well in each of the two validation sets. Both scoring functions provide high EF value for Schrodinger and DUD-E validation sets. Also, the Vina score function showed average performance for the discrimination of active from decoy while knowledge-based scoring functions' LeDock score' performed poorly for each validation set. No actives are present in the first 5 and 2% of the validation dataset containing Schrodinger and DUD-E decoy set, respectively. However, evaluation using EF only is not an accurate way, because the maximum value for EF is strongly dependent on the number of actives and inactives.

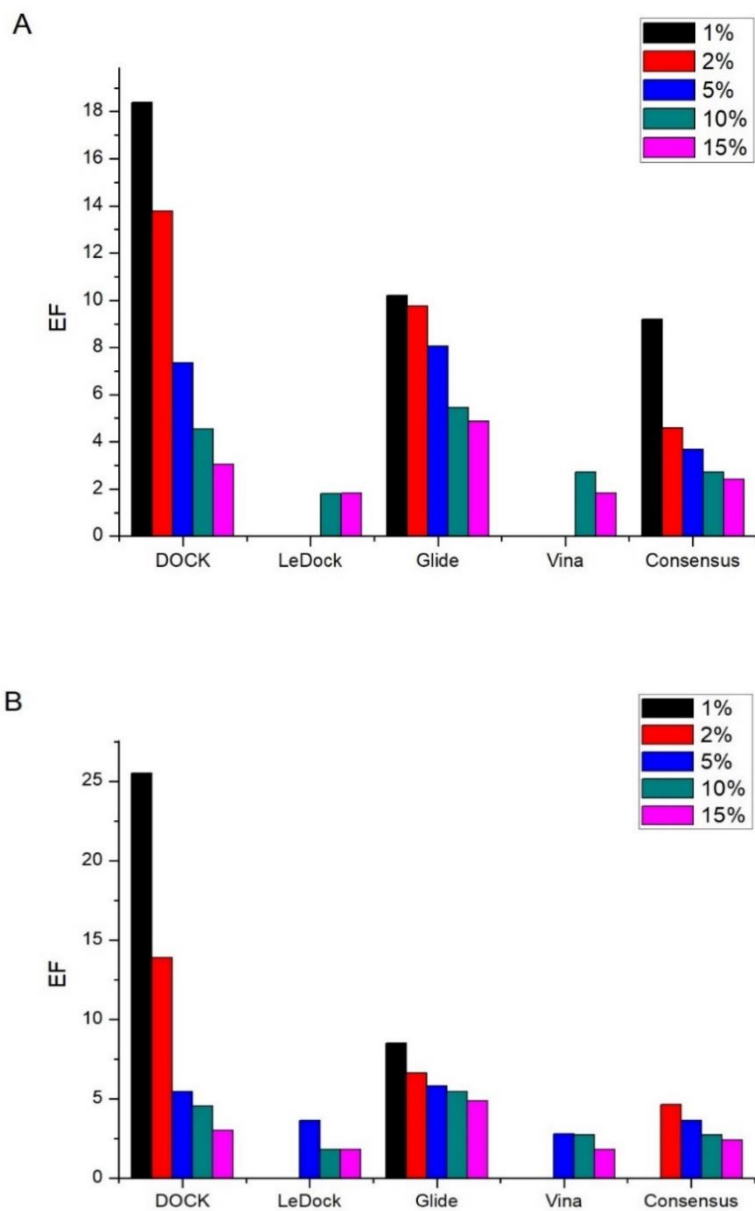


Fig. 7.4 Histogram of EF calculated for a dataset contains 1, 2, 5, 10, and 15% of the total A) Schrödinger decoy set +actives and A) DUD-E decoy +actives.

The area under the curve (AUC) of a receiver operating characteristic (ROC) was also used to measure the overall performance of docking enrichment and is independent of the number of actives. That is, the probability of active compounds being ranked earlier than decoy compounds are obtained from the area under the ROC (AU-ROC) curve, and it has a value ranging from 0 for a complete failure to 1 for a perfect enrichment. Performance evaluation matrices of each docking program are recorded in Table 7.1. All individual scoring functions were successful in the discrimination of active compounds over inactive with an AU-ROC score higher than seven (DOCK 6: 0.733, Glide: 0.917, Vina: 0.709) except LeDock score. However, Truchon and Bayly explained the inability of AU-ROC to address the "early recognition" problem specific to VS [30]. A fundamental requirement for the accomplishments of VS is that active compounds should be ranked very early because only a few compounds can be tested experimentally. Even if the VS approach is outstanding in the first half of the data set, it is useless if the early recognition is poor. Truchon and Bayly have also shown that the exponential weighting schemes BEDROC and RIE provide proper "early recognition" of actives [30]. BEDROC is derived from ROC generalization, but it tackles both the question of "early recognition " and RIE. So, the BEDROC value of DOCK 6, Glide, and Vina are 0.477, 0.689, 0.334, respectively, indicating the satisfaction of early recognition of active compounds, especially for Glide. The consensus study shows that the combination of the four methods using the same z-scores did not result in better enrichments than Glide and DOCK 6. Vina and consensus

scoring displayed an average performance, but consensus scoring slightly outperformed Vina in the conditions of this experiment. All of the validation parameter values obtained for Glide docking procedure is better than that obtained for the consensus level of four independent docking program, thus suggesting that consensus docking procedures are not able to filter the enriched database as efficiently as the Glide docking approach. Consensus scoring could not, in this context, increase the performance of the VS.

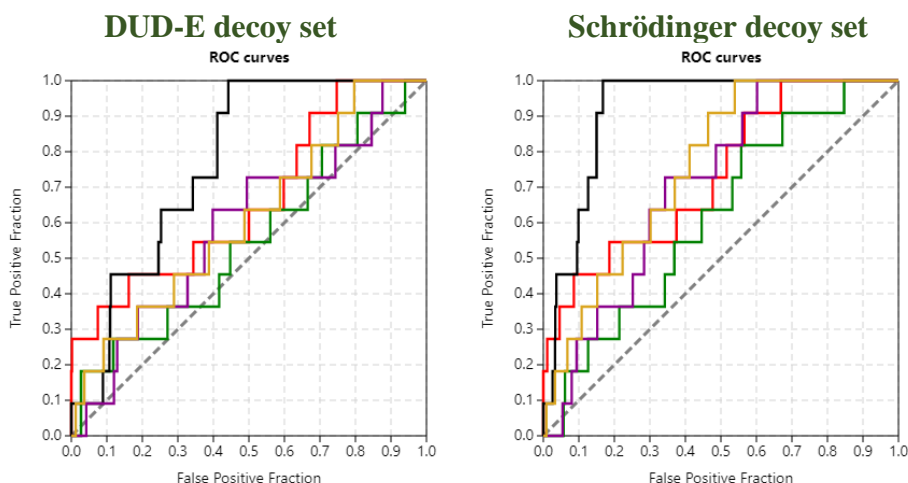
Table 7.1 Evaluation Metrics of docking programs

	ROC AUC	Total Gain (TG)	RIE	BED ROC	Avg. rank of actives	Maximum reachable EF
Schrodinger decoy set						
DOCK 6	0.733	0.401	5.986	0.477	272.55	91.91
LeDock	0.616	0.159	1.282	0.273	390.00	91.91
Glide	0.917	0.652	6.236	0.689	89.45	91.91
Vina	0.709	0.304	1.358	0.334	297.45	91.91
consensus	0.756	0.381	3.184	0.418	249.64	91.91
DUD-E decoy set						
DOCK 6	0.661	0.283	5.554	0.442	192.64	51.00
LeDock	0.547	0.08	2.229	0.274	255.18	51.00
Glide	0.771	0.363	2.680	0.418	132.00	51.00
Vina	0.587	0.208	1.128	0.269	232.91	51.00
consensus	0.609	0.184	2.565	0.316	221.00	51.00

Each docking score function follows the same order but differs in magnitude when it comes to selecting the active from decoy set. Overall, the Schrödinger decoy validation set yielded the highest outcomes for enrichment, while the DUD-E decoy validation set yielded the worst results. The docking programs may have the most

difficulty in distinguishing the active compounds from the decoy set if similar in size and lipophilicity; however, this tendency was not seen in our enrichment result. The decoy set of Schrödinger differentiates the most from the active compounds but returns only slightly better enrichment outcomes than DUD-E, which have closest parameters.

Graphically assessing the quality of various docking algorithms for virtual screening is now on the trend. Among them, the Receiver Operating Characteristics (ROC) curves [29], enrichment curves, and the newly implemented Predictiveness Curves (PC) [32] offer good logical visualization, handle different characteristics of the results and present them intuitively. Figure 7.5 provides a visual comparison of the performances of individual scoring functions via ROC curves, PC, and enrichment curves.



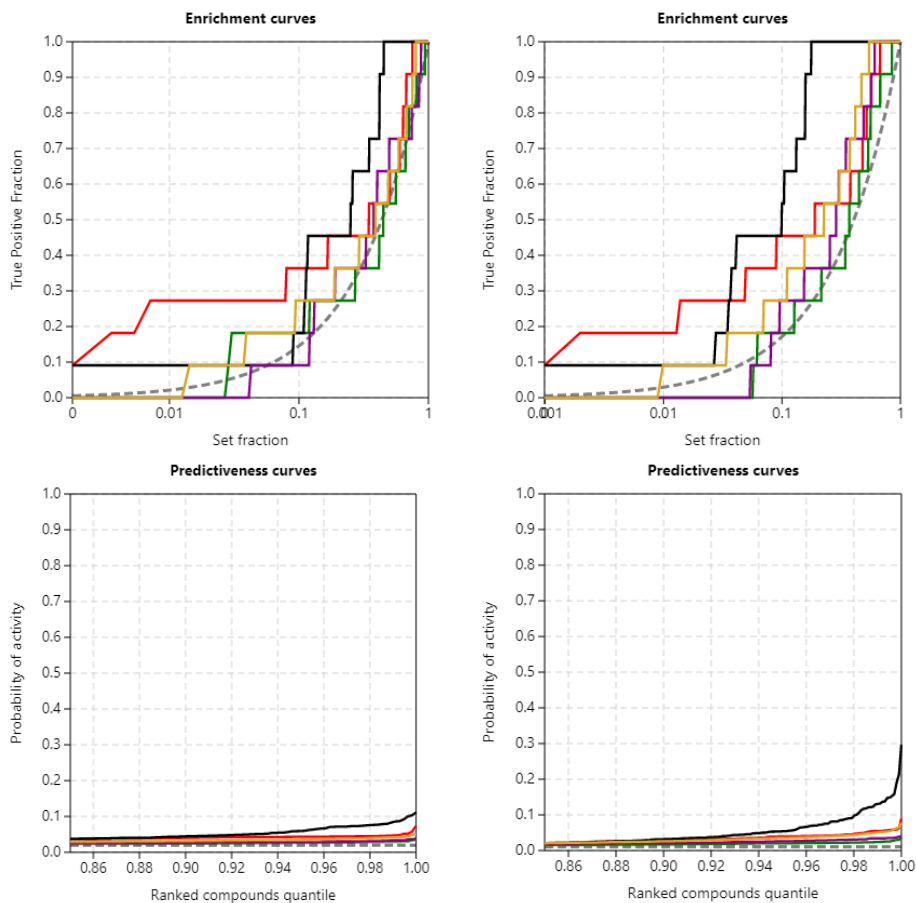


Fig. 7.5 Predictiveness, ROC, and enrichment curves for the virtual screenings of 5-LOX inhibitors from the DUD-E and Schrodinger decoy datasets using Glide XP, DOCK 6, Consensus, Vina and LeDock score with a color code of black, red, yellow, violet and green, respectively.

The ROC curve is a chart of the TPF (True Positive Fraction) versus the FPF (False Positive Fraction) for all compounds in an ordered dataset [29,33]. Here TPF is a fraction of the active compound, whereas FPF is a fraction of the inactive compounds. The ROC curves assess the overall success of a method in the ranking of active compounds. A ROC curve goes through the top-left corner of the plot

where the TPF is one, and the FPF is zero to indicate ideal discrimination. The nearer the curve is to the upper left corner, the higher the overall accuracy of the test. A 45° diagonal line shows no discrimination. Here, it can be seen that the Glide XP score can able to perform perfect discrimination compared to other scoring functions because its ROC curve for each validation sets passes through closest to the upper left corner of the plot. ROC curve of Vina score (violet) and DOCK 6 score (red) was also found to be far away from the diagonal and closer to the upper left corner, indicating the quality of these scores in discriminating 5-LOX inhibitors from decoy especially from Schrodinger decoy.

The predictiveness curve was built as a cumulative distribution function (CDF) of activity probabilities, and activity probabilities are derived from the scores obtained by the compounds in a virtual screening experiment using generalized linear models with a binomial distribution function and the canonical log link [32]. PC enables the identification of potential score gaps and variations caused by a score function in the monitoring of active compounds, which correspond to activity probabilities gaps. PC might align to a horizontal line at the level of activity pervasiveness in a completely uninformative model, while conversely, steep curves allow an inflection point from which the curve rises to be observed. That is, good predictions from virtual screening methods produce steeper PC curves that correspond to broader variations in activity probabilities. Here, the PC curve of the Glide XP score shows a steeper curve, which again supports the good predictive and discrimination power of this score. The standardized

Total Gain (TG), an output metric from the PC, summarizes the discrimination of active compounds imputable to the variation of the scores over a complete molecular dataset [32]. Also, TG values over 0.25, together with ROC AUC values over 0.5, typically indicate that the score variations in activity discrimination are relevant and that their performance is comparable. Also, the experimental conditions would be reproducible. The total gain of DOCK 6, Glide, and Vina methods respectively are 0.401, 0.652, and 0.304, indicating these methods produced meaningful score variations in the detection of the actives. Enrichment curves or accumulation curves are used to assess the early recognition of active compounds by envisioning the TPF (y-axis) for each fraction of the ordered dataset on a logarithmic scale (x-axis) [30]. This plot also concludes the high discrimination power of the Glide XP score as compared to other scores and medium discriminating power of Vina scores.

7.6. Application of Virtual Machines in Virtual Screening

Virtual screening of big databases requires high computational power, which can be either in the form of a supercomputer with high processing speed and capacity or can be clusters of small computers with medium processing capacity. It requires a high cost and facility in both ways. Therefore, we need to establish a viable method of computing that can reduce costs and improve the efficiency of virtual screening. The emerging technology such as virtual machine (VM) and cloud computing revolutionizing the area of computer science may also be beneficial to the field of chemical science, especially in virtual

screening study [34]. Previous research has documented the same [35,36].

7.6.1. Application of VMs in Research

Virtualization of the software platform (virtual machine system) enables a single host operating system to run various guest operating systems without rebooting on the same computer. For example, a computer running LINUX can run an independent Microsoft Windows operating system in a separate window and vice versa. Generally, programs compiled for a particular operating system can only operate on the same operating system. But virtualization provides a platform to pick up any operating system, depending on the user's choice. These are the significant characteristics of the virtualization of the platform. Another benefit is to test software in a working production environment without installing the software. Among fifty commercial or free open-source VMs that are currently available, VMware Workstation for Windows, as well as LINUX, is commonly used. Literature showing virtualization concepts in chemistry are extremely scarce, however, Bullard D *et al.*, mention the importance of virtualization software in the pharmaceutical industry for virtual screening and lead optimization in a grid-like environment [37].

7.6.2. Engineering Infrastructure

To run virtual screening using all docking programs, we have required a LINUX platform with more numbers of CPU because some of the docking programs may not work in the windows platform. But in our Laboratory, we have only 12 windows machines with a dual Intel Xeon E5-2640 quadcore CPU (2.40 GHz) system with 32 GB RAM and 24 × 1TB Seagate Barracuda ES.2 hard disks. So, we have installed VMware workstation 14.1.1 in each system. Memory and processors were set to 16 GB RAM and 16 processors for the virtual machine because all docking programs may take benefit of multiple CPUs or CPU cores on the system to shorten its runtime considerably and then installed Ubuntu 16 as a Guest operating system.

7.6.3. Benchmarking Computational Power

We have investigated probable speed loss during the use of virtual machines using a series of scientific benchmarks. VMware virtual machine also allows multiple CPU setups; therefore, all results are based on maximum CPU speed instead of utilizing one single CPU. We selected a list of 1000 random molecules from our screening database and docked using respective software. To evaluate the performance of the various hardware and operating systems, including VMware, we noted the number of compounds docked per hour by each docking program across a variety of different architectures. The details of the benchmarking are given in Table 7.2.

Table 7.2 Benchmarking result of computational power for virtual screening

CPU type	Used CPU and RAM	Operating system	Compounds docked per hour			
			Glide XP	DOCK 6	Vina	LeDock
Intel Xeon E5-2640	1 CPU 2 GB	Ubuntu 16.04 LTS	20	17	20	15
Intel(R) Core (TM) i5-8250U	8 CPU 8 GB	Windows 10	35	-	144	120
Intel Xeon E5-2640 (quadcore)	20 CPU 32 GB	Windows 10	47	-	300	240
Intel Xeon E5-2640 (quadcore)	16 CPU 16 GB	VMware work station with Windows 10 host and Ubuntu 16.04 LTS guest	76	275	416	400

The docking processes and the scoring functions that we have implemented are endowed with distinct computational time demands, which are inversely correlated with their accuracy. DOCK 6 program is not run in the Windows platform, so corresponding results are not included here. VMware work station with Windows 10 host and Ubuntu 16.04 LTS guest with 16 CPU and 16 GB has maximum speed than windows with 20 CPU 32 GB RAM indicates virtualization reduces the time of docking for all docking program. Also, the number of docked compounds per hour is larger for the Vina compared to the other docking programs. That is, the average total CPU time required for the processing of 1, 000 compounds were 1 hour for the vina docking while, 24 hours for Glide XP docking. So, in the early phases of virtual screening campaigns, the quickest and the coarsest docking program AutoDock Vina with average docking accuracy is decided to

implement in evaluating the data of millions of compounds; in contrast, the slowest and most meticulous one Glide XP is agreed to apply to later phases of the campaigns when data sets have already decreased considerably.

7.7. Virtual Screening of ZINC 15 Database

Virtual screening of lead and hit compounds based on molecular docking is one of the sophisticated approaches in the drug design process. In this study, the necessary inputs are experimentally solved target structure and a 2.7 million compound library of small molecules available by purchase derived from the ZINC 15 database. Figure 7.6 depicts Scheme, which represents the protocol used in this study for virtual screening. Three sequential docking protocols were performed to find a novel and potent 5-LOX inhibitors.

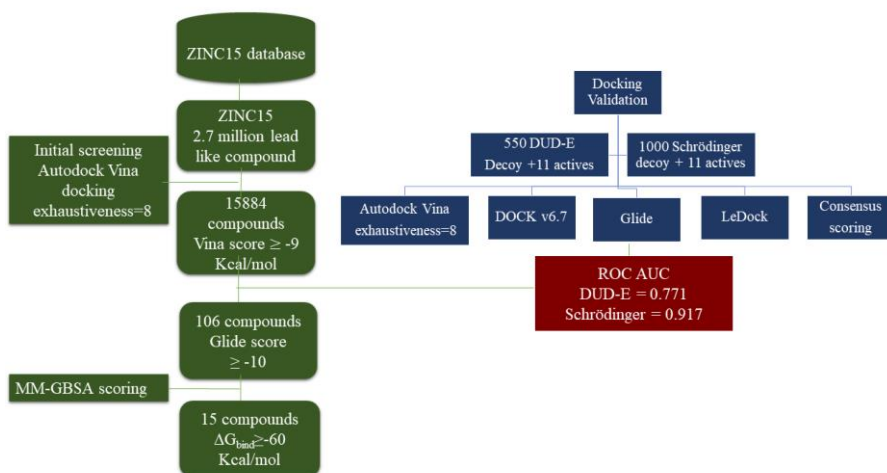


Fig. 7.6 Virtual screening and docking performance evaluation scheme used in the study.

All the compounds were screened with stepwise filtering strategy; initially, the 2.7 million lead like compounds are filtered from the ZINC 15 database contains over 1130 million purchasable compounds. Then these 2.7 million compounds were docked into the active site of 5-LOX. This preliminary database screening was performed using AutoDock Vina software to screen compounds at a faster rate. Evaluation of various molecular docking programs explained in section 7.5 shows that the docking program 'Vina' is well performed in enriching actives in the dataset and discriminating actives from the decoy. Computational time is taken for vina docking is lower than all other dockings. Also, by the Comparative Assessment of Score functions (CASF) benchmark in 2013, AutoDock Vina is listed among the high-ranking scoring functions for docking power and screening [38]. Therefore, initial screening was performed with Vina docking in a VMware work station with Windows 10 host and Ubuntu 16.04 LTS Guest. The distribution of 2.7 million compounds over a range of docking scores is shown in the frequency distribution graph in Figure 7.7. From this figure, it can be observed that over a large number of compounds have a binding affinity score of -6 to -7 kcal/mol. Molecules with binding affinity lesser than -9 kcal/mol, were selected for further screening. A total of 15884 compounds were found with a binding affinity higher than -9 kcal/mol. These 15884 compounds were virtually screened using a more accurate docking method 'Glide XP' because this method is shown to yield enrichments superior to the other four alternative methods consistently. The top 106 compounds with a high Glide XP score (≥ -10) were obtained as a

result of this step. These compounds should be further screened or re-docked to get an accurate and most sophisticated binding affinity using the post-docking processing techniques. The following section will describe the details of the same.

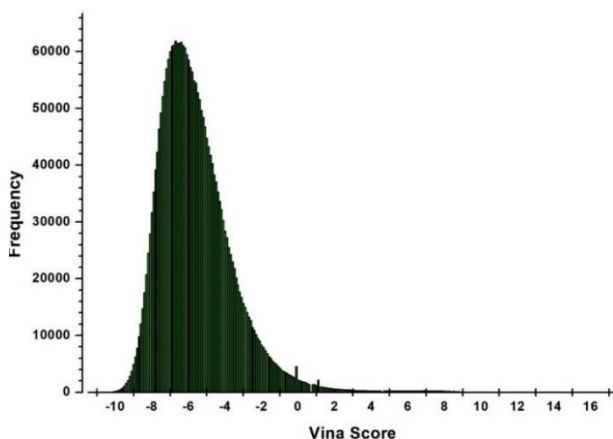


Fig. 7.7 Frequency distribution graph showing the distribution of ~2.7 million compounds over the range of docking scores (Scores are in kcal/mol).

7.8. Binding Free Energy Calculation

Since large numbers of compounds to be screened in a reasonable amount of time, we need to use the approximate scoring functions that will result in the non-correlation of docking scores and experimental affinities [39]. Also, it is a challenging and target-dependent task to sample the conformational space accessible to ligand-target complexes in an induced-fit context [39]. These two effects may generate a few false-positive and false-negative hits in the library of the screened compound, which then requires careful assessment and additional post-docking analyses. Docking

performance should, therefore, be enhanced, employing more robust post-docking processing techniques. The binding free energy obtained from MM-GB/SA calculation is one of several post-processing strategies built to resolve docking constraints.

This study, too, used Prime MM-GB/SA's binding free energy as a post-docking scoring protocol to correctly rank the potent inhibitor molecule against target protein [40]. MM-GB/SA uses molecular mechanics, the Generalized Born model, and Solvent accessibility method to obtain free energy from structural information to determine the relative binding free energies (ΔG_{bind}) in biomolecular complexes [41,42]. The binding energies derived via the MM-GB/SA OPLS-2005 are known to be far more reliable and precise than the XP GScore [43]. With the MM-GB/SA method implemented in the Schrödinger software suite in the Prime program, all 106 compound docking poses are subjected to rescoring. This process then leads to minor changes of the ligand conformations within the receptor site, which results in the ranking of ligand-based on calculated binding energies (MM-GB/SA ΔG_{binds}) using Equation 7.3.

$$\Delta G_{bind} = \Delta E + \Delta G_{solv} + \Delta G_{SA} \quad (7.3)$$

Where ΔG_{SOLV} is the difference in the GB/SA solvation energy of the protein–inhibitor complex and the total of the solvation energies for the unbound 5-LOX and inhibitor; ΔG_{SA} is the change in surface area energies for the complex and the totality of the surface area energies for the unbound 5-LOX and inhibitor. Where ΔE is the difference in the minimized energies between the 5-LOX–inhibitor

complex and the total of energies of unbound 5-LOX and inhibitor shown in Equation 7.4.

$$\Delta E = E_{complex} - E_{protein} - E_{ligand} \quad (7.4)$$

While the simulation process, the ligand strain energy was also taken into consideration. Based on results, MM-GB/SA's binding free energy calculation, 106 molecules again ranked, and best 15 molecules with high ΔG_{bind} and Glide XP scores were then screened. MM-GB/SA's binding free energy and contribution from each energy parameter like coulombic, covalent, Hydrogen bond, and lipophilic interaction of 15 potential hits are given in Table 7.3, and their respective chemical structures are shown in Figure 7.8. All energy values are in kcal/mol. The more negative the MM-GB/SA binding energies indicate stronger the binding. So compound ZINC00238144370 has maximum binding energy with a Glide Gscore of -10.102, followed by ZINC000225607571 with the binding energy of -67.248 kcal/mol and a Glide Gscore of 10.148 kcal/mol. The powerful lipophilic interaction (ΔG_{bind} Lipo) and intensified electrostatic interaction (ΔG_{bind} Coulomb) are the major contributors to the strong binding of ligands to 5-LOX. Contribution to the free energy from the hydrogen bond is too shallow. The categorization of ligands based on calculated binding energy (MM-GB/SA ΔG_{bind}) is maybe reasonably consistent with the categorization based on the experimental binding affinity. Therefore, these fifteen molecules are expected to have good antagonist activity against 5-LOX enzyme and could be used as a potential hit for the lead development of 5-LOX

inhibitors. Nature of interaction of these molecules to the active site amino acid of 5-LOX enzymes explained in the next section.

Table 7.3 Final hits and their MM-GB/SA's binding free energy, energy components of the ligand-5-LOX complexes, and Glide Gscore values (all values are in kcal/mol).

Ligand	ΔG_{bind}	Coulomb	Covalent	Hbond	Lipo	Glide Gscore
ZINC000238144370	-68.89	-35.46	2.00	-1.66	-24.88	-10.10
ZINC000225607571	-67.25	-24.73	4.58	-1.93	-30.66	-10.15
ZINC000065552536	-66.90	-24.78	3.83	-1.64	-29.50	-10.69
ZINC000206717530	-66.66	-11.61	2.41	-1.01	-30.32	-10.15
ZINC000408513761	-65.26	-28.92	4.57	-2.07	-30.01	-10.16
ZINC000225607740	-64.93	-27.07	5.22	-2.32	-28.27	-10.81
ZINC000085560727	-63.41	-16.49	1.14	-1.53	-31.32	-10.25
ZINC000081818151	-63.13	-23.84	7.47	-2.19	-33.02	-10.53
ZINC000031097555	-62.78	-26.11	6.95	-2.10	-23.89	-10.34
ZINC000069778142	-62.38	-28.26	7.03	-1.94	-27.99	-11.55
ZINC000095417240	-61.99	-26.92	6.64	-2.09	-28.65	-11.84
ZINC000075155788	-61.36	-20.66	5.14	-1.42	-30.86	-10.05
ZINC000121756318	-60.97	-21.70	2.07	-1.98	-31.31	-10.31
ZINC000095533957	-60.93	-33.59	11.63	-2.36	-28.14	-10.20
ZINC000040059693	-60.87	-35.62	7.71	-1.85	-27.05	-10.20

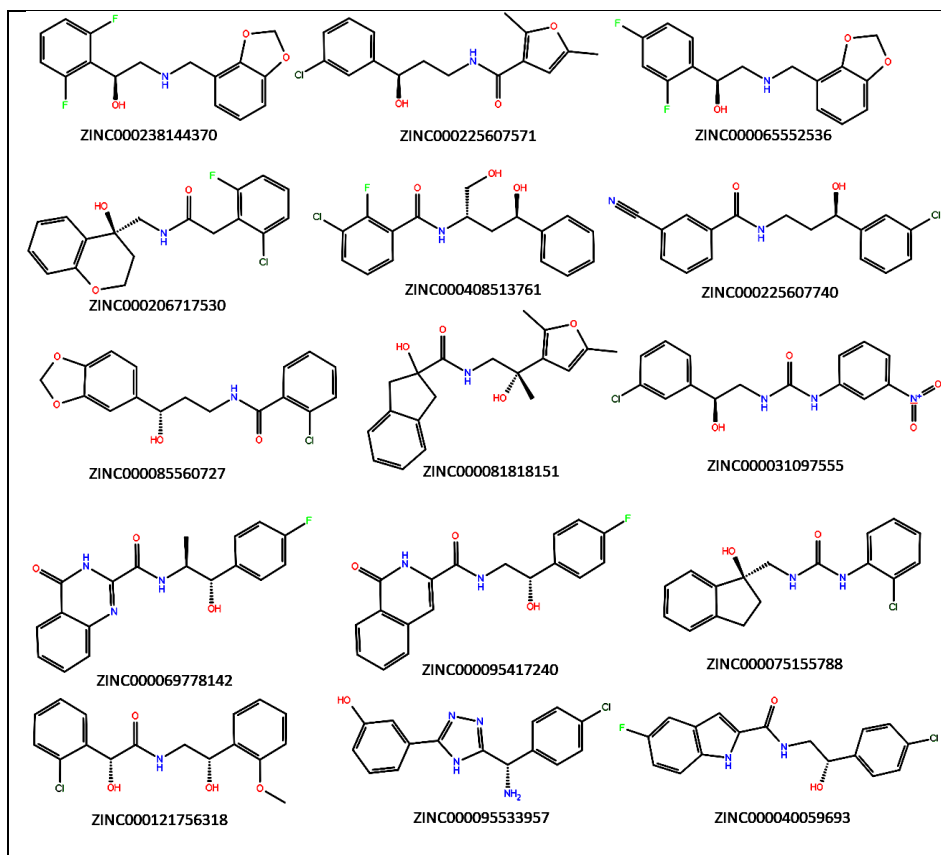


Fig. 7.8 Potential virtual hits of 5-LOX inhibitors identified through virtual screening of the ZINC 15 database.

7.9. Interaction of Virtual Hits at the Active Site

The interaction between protein 5-LOX and virtual hits has been investigated in this section to identify the nature of bonding and other non-covalent interaction that helps ligand to situate in the binding cavity. 2D interaction maps of these docked complexes are shown in Figure 7.9. The binding mode between virtual hits and 5-LOX reveals that interaction is almost similar to the interaction of reference compounds (well-known 5-LOX antagonist), Zileuton, and

NDGA with 5-LOX (Figure 3.3) But these molecules do not obey the classical inhibitory mechanism of a good 5-LOX inhibitor that should have a polar head and tail and a hydrophobic body. However, the majority of interactions are hydrophobic and polar. Amino acids like Phe 177, Ala 410, Leu 607, Val 604, Ala 603, Ile 415, Leu 414, Phe 359, Pro 569, Trp 599, Ala 424, Tyr 181, Phe 421, Leu 420, Leu 368, Ile 673 and Ile 406 are the crucial residues for non-bonded hydrophobic interactions while amino acids like His 600, Asn 425, Gln 363, His 367, Gln 557, His 372, Asn 407, Asn 425 and Thr 364 are the crucial residues for polar non-bonded interactions. Both potential hits and reference compound NDGA has π - π stack interaction with amino acid His 367. Likewise, ligand molecules and reference compound zileuton form H-bond with His 363. ZINC000238144370 and ZINC000225607571 also form H-bond with His 367. Interestingly, most of the potential hits have either an amide group or urea group in the middle and aromatic or heterocyclic ring present at each end. Generally, C=O groups or NH groups present in the ligands form H-bond with polar amino acids like Gln 363, His 367, His 372, Asn 425, Tyr 181, Ile 673, His 600, *etc.*, present in the active site. Also, aromatic ring or heterocyclic aromatic ring form π - π stack interaction with aromatic amino acids like His 372, His 367, Tyr 181, Phe 421, *etc.*, present in the active site. Thus, these 15 virtual hits which exhibit better Gscores greater than -10 kcal/mol and MM-GB/SA's binding free energy greater than -60 kcal/mol with a stronger H-bond and other non-bonded interaction compared to reference known inhibitors are considered as potential lead compounds. Pharmacokinetic and toxicity

risk assessments of these 15 potential hits are performed to account more on druggability, and the result is provided in the following section.

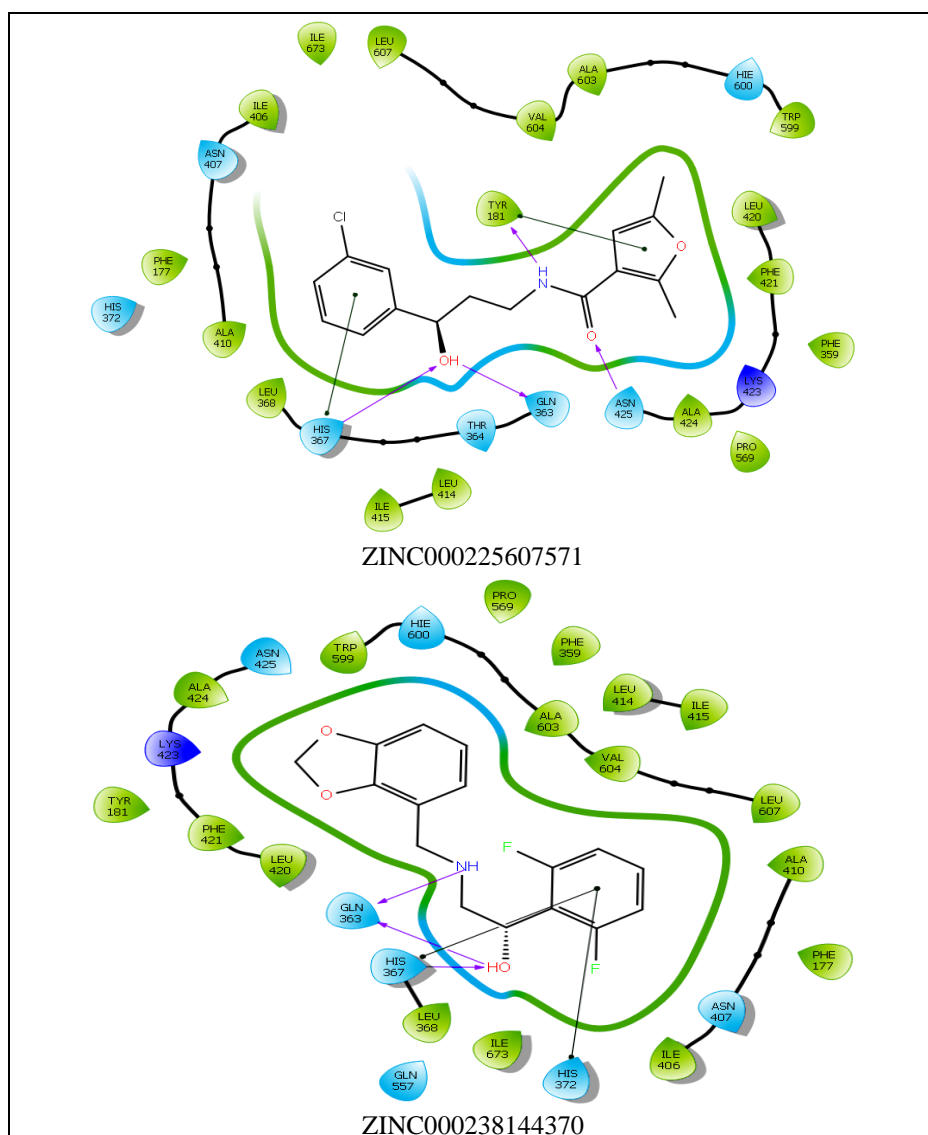


Fig. 7.9 2D view of the binding interaction of two virtual hits with 5-LOX active site amino acids.

7.10. ADME Property and Toxicity Analysis

ADME is an acronym for "absorption, distribution, metabolism, and excretion" in pharmacokinetics and pharmacology and depicts the disposition of a drug within an organism [44]. The goal of ADME analyses is to achieve an early assessment of the human pharmacokinetic and metabolic profiles. The weak pharmacokinetic characteristics of the stronger active compound make it much less active *in vivo* analysis. In addition to ADME studies, early toxicology and safety studies also considered for filtering out compounds before lengthy and expensive clinical trials [45]. It is possible to assess the safety of potential drug candidates by investigating the genotoxicity, mutagenicity, safety pharmacology, and general toxicology of the candidate. Along with toxicity assessment, ADME parameters of the potential drug candidates (ADME/T properties) should be considered and tested before the lead optimization process for successful drug discovery.

Although initial screening of ZINC 15 database is based on the selection of the subset of lead-like compounds with a molecular weight between 250 and 350 g/mol, predicted partition constant ($x\text{LogP}$) ≤ 3.5 , and the number of rotatable bonds (RBs) ≤ 7 , still need more emphasis on ADME/T to explore the pharmacokinetics toxicological nature of the potential hits. So, in this section, the ADME properties of selected virtual hits were analyzed using the QikProp tool of the Schrödinger suite. It is used to predicts ADME properties and pharmaceutically relevant physicochemical descriptors of all ligands. Tables 7.4 and 7.5 gives the QikProp results of the 15 top-ranked ligands. For each descriptor, the range satisfying 95 % of known drugs is also provided

for comparison. The QikProp descriptors calculated are molecular weight (MW), total solvent accessible surface area (SASA), the hydrophobic component of the SASA (FOSA), the estimated number of hydrogen bond donor (HBD) and acceptor (HBA) in aqueous solution, and Human Oral Adsorption (HOA). The values of QPlogS, QPlogBB, QPlogPo/w, QPlogPw, and QPlogPoct, respectively, are the predicted partition coefficients of aqueous solubility, brain/blood, octanol/water, water/gas, and octanol/gas whereas QPlogKp is predicted skin permeability.

Table 7.4 QikProp results of the top-ranked ligands

Ligand	MW	SASA	FOSA	HBD	HBA	HOA
Reference range	130–725	500–2000	0–750	0.0–6.0	2.0–20.0	1: L 2: M 3: H
ZINC000238144370	307.30	532.71	142.45	2	4.7	3
ZINC000225607571	307.78	686.37	308.91	2	4.7	1
ZINC000065552536	307.30	537.35	146.43	2	4.7	3
ZINC000206717530	349.79	591.34	139.92	2	4	3
ZINC000408513761	337.78	589.32	88.51	3	5.9	3
ZINC000225607740	314.77	609.93	75.41	2	5.7	3
ZINC000085560727	333.77	567.90	147.85	2	5.7	3
ZINC000081818151	329.40	605.83	303.08	3	4.5	3
ZINC000031097555	335.75	613.58	45.88	3	4.7	3
ZINC000069778142	341.34	588.04	94.29	2.25	6.95	3
ZINC000095417240	326.33	606.41	45.85	2.25	5.95	3
ZINC000075155788	316.79	580.17	117.02	3	2.75	3
ZINC000121756318	335.79	594.85	140.25	3	6.65	3
ZINC000095533957	300.75	547.12	18.12	4	3.75	3
ZINC000040059693	332.76	605.47	45.89	3	4.2	3
NDGA	302.37	582.64	164.29	4	3	3
Zileuton	236.29	452.29	81.63	3	3.7	3

Table 7.5 QikProp results of various predicted partition coefficients and predicted skin permeability parameters of the top-ranked ligands.

Ligand	QPlogS	QPlogBB	QPlogKp	QPlogPo/w	QPlogPw	QPlogPoct
Reference range	-6.5-0.5	-3.0-1.2	-8.0-1.0	-2.0-6.5	4.0-45.0	8.0-35.0
ZINC000238144370	-2.38	0.34	-2.87	2.62	9.47	14.97
ZINC000225607571	-6.27	-0.96	-2.32	3.68	9.71	17.37
ZINC000065552536	-2.53	0.37	-2.94	2.69	9.38	15.03
ZINC000206717530	-4.36	-0.16	-1.24	3.45	10.97	16.57
ZINC000408513761	-3.95	-0.58	-1.36	3.00	12.35	19.10
ZINC000225607740	-5.40	-1.34	-2.72	2.86	11.36	18.45
ZINC000085560727	-3.89	-0.40	-1.39	2.93	10.79	17.02
ZINC000081818151	-3.76	-0.74	-2.02	2.65	12.28	17.96
ZINC000031097555	-4.26	-1.86	-3.84	1.80	13.60	19.69
ZINC000069778142	-3.98	-1.06	-2.80	2.20	13.23	19.00
ZINC000095417240	-4.55	-1.35	-2.92	2.34	12.74	18.20
ZINC000075155788	-4.41	-0.36	-1.63	3.23	11.43	17.23
ZINC000121756318	-2.63	-0.94	-1.81	1.78	15.36	18.95
ZINC000095533957	-2.83	-0.57	-4.94	1.90	12.84	19.45
ZINC000040059693	-5.19	-0.68	-2.11	3.46	11.64	18.74
NDGA	-3.64	-1.90	-3.62	2.55	10.79	17.53
Zileuton	-1.50	-0.71	-3.09	0.89	13.21	14.71

All the 15 molecules fall within the recommended ranges of properties and have a high human oral absorption range, thus indicating their potential as a drug-like molecule. All six predicted partition coefficients fall within the recommended ranges. Uniquely, medium predicted aqueous solubility (QPlogS) of all compounds indicates that they are more soluble and more absorbable, and it may decrease the quantity of drug prescribed to achieve the desired pharmacological effect while minimizing the risk of side-effects and toxicity. The three partition coefficients, QPlogPo/w (octanol/water),

QPlogPw (water/gas), and QPlogPoct (octanol/gas), together imply that 15 potential compounds have medium lipophilicity which can cause to medium gastrointestinal absorption through passive diffusion. The predicted brain/blood partition coefficient (QPlogBB) of all compounds is also within the recommended range, indicating the proper brain penetration of the hit molecules. The number of hydrogen bond donors are less than 5, and hydrogen bond acceptors are less than 10, so the compounds will satisfy Lipinski's rule for drug likeliness.

Table 7.6 Toxicity prediction Results from DataWarrior.

Molecule Name	Mutagenic	Tumorigenic	Reproductive Effective	Irritant
ZINC000238144370	none	none	none	none
ZINC000225607571	none	none	none	none
ZINC000065552536	none	none	none	none
ZINC000206717530	none	none	none	none
ZINC000408513761	none	none	none	none
ZINC000225607740	none	none	none	none
ZINC000085560727	none	high	none	none
ZINC000081818151	none	none	high	none
ZINC000031097555	none	none	none	none
ZINC000069778142	none	none	none	none
ZINC000095417240	none	none	none	none
ZINC000075155788	none	none	none	low
ZINC000121756318	none	none	none	none
ZINC000095533957	none	none	none	none
ZINC000040059693	none	none	none	none
ZINC000040059693	none	none	none	none

The toxicity risk assessment aims to identify substructures that are representative of a toxicity hazard within one of four main toxicity

groups, such as mutagenic, tumorigenic, reproductive effective, and irritant within the chemical structure. Toxicity Prediction results are obtained from Data Warrior [46] software are given in Table 7.6. The red box with a 'high' mark indicates compound with high risks of unwanted effects such as mutagenicity or low intestinal absorption whereas the green color box with 'none' writings suggest that the compounds have no toxicity risk at all and the yellow-colored box with 'low' writing indicates a low toxicity risk compound. Most of the compounds show green color, indicating free of fragments that are within one of four major toxicity classes. But ligand ZINC000085560727 and ZINC000081818151 show high tumorigenic and reproductive toxicity risk, respectively, while ZINC000075155788 is might be a mild irritant. Only these three compounds show some degree of toxicity, so special care is needed, and advanced toxicity studies should be carried out before optimizing the lead. To conclude, these studies show that all 15 virtual hits can be more potent lead compounds with the best ADME/T score and show stronger 5-LOX binding interactions than currently known compounds.

7.11. Conclusion

In this Chapter, we have conducted a virtual screening of 2.7 million ZINC 15 compounds to identify novel potential 5-LOX inhibitors. For this, we have done a comparative assessment of four commonly used docking programs such as Glide, LeDock, DOCK 6, and AutoDock Vina before the virtual screening. Additionally, a consensus model that combines all of the four docking programs is also developed. The assessment was based on the scoring reliability of each scoring function to recognize the known active from decoys

based on various matrices like ROC, ROC AUC, BEDROC, RIE, and EF and the various curves such as ROC curves, enrichment curves, and PC. The result indicates that the Glide XP score has high discrimination power as compared to other scores. The effect of the inclusion of a virtual machine to speed up virtual screening is assessed by benchmarking various computer power. The result confirms that VMware work station with Windows 10 host and Ubuntu 16.04 has maximum speed indicates virtualization reduces the time of docking for all docking programs, and the computational time is taken for vina docking is exceptionally lower than all other dockings. So, In the early phases of virtual screening campaigns, the quickest and the coarsest docking program AutoDock Vina with average docking accuracy is implemented, resulting in 1588 molecules with greater than -9 kcal/mol got screened. While most slower and most meticulous, one 'Glide XP' is then used to filter out 109 virtual hits with Glide Gscore greater than -10 kcal/mol. Rescoring of these compounds has done with MM-GB/SA's binding free energy calculation, and the best 15 molecules with high ΔG_{bind} (greater than -60 kcal/mol) were then screened. These 15 virtual hits then subjected to ADME and toxicity analysis resulting in good ADME score and good toxicity result except for three compounds. They need to undergo advanced toxicity assessment before further analysis. Overall, this study suggests that fifteen potential hits are expected to have good antagonist activity against 5-LOX enzyme and could be used as a potential lead for lead development of 5-LOX inhibitors.

References

- [1] S. Kar, K. Roy, How far can virtual screening take us in drug discovery?, *Expert Opin. Drug Discov.* 8 (2013) 245–261. doi:10.1517/17460441.2013.761204.
- [2] G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe Jr, Computational methods in drug discovery, *Pharmacol. Rev.* 66 (2013) 334–395. doi:10.1124/pr.112.007336.
- [3] A. Gimeno, M.J. Ojeda-Montes, S. Tomás-Hernández, A. Cereto-Massagué, R. Beltrán-Debón, M. Mulero, G. Pujadas, S. Garcia-Vallvé, The Light and Dark Sides of Virtual Screening: What Is There to Know?, *Int. J. Mol. Sci.* 20 (2019) 1375. doi:10.3390/ijms20061375.
- [4] E. Lionta, G. Spyrou, D.K. Vassilatis, Z. Cournia, Structure-based virtual screening for drug discovery: principles, applications and recent advances, *Curr. Top. Med. Chem.* 14 (2014) 1923–1938. doi:10.2174/1568026614666140929124445.
- [5] S.-Y. Huang, Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges, *Brief. Bioinform.* 19 (2017) 982–994. doi:10.1093/bib/bbx030.
- [6] S. Sinha, M. Doble, S.L. Manju, 5-Lipoxygenase as a drug target: A review on trends in inhibitors structural design, SAR and mechanism based approach, *Bioorg. Med. Chem.* 27 (2019) 3745–3759. doi:https://doi.org/10.1016/j.bmc.2019.06.040.
- [7] C. Pergola, O. Werz, 5-Lipoxygenase inhibitors: a review of recent developments and patents, *Expert Opin. Ther. Pat.* 20 (2010) 355–375. doi:10.1517/13543771003602012.
- [8] B. Hofmann, D. Steinhilber, 5-Lipoxygenase inhibitors: a review of recent patents (2010 – 2012), *Expert Opin. Ther. Pat.* 23 (2013) 895–909. doi:10.1517/13543776.2013.791678.
- [9] T. Sterling, J.J. Irwin, ZINC 15--Ligand Discovery for Everyone, *J. Chem. Inf. Model.* 55 (2015) 2324–2337. doi:10.1021/acs.jcim.5b00559.
- [10] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, *J.*

Cheminform. 3 (2011) 33. doi:10.1186/1758-2946-3-33.

- [11] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, T. Hou, Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power, *Phys. Chem. Chem. Phys.* 18 (2016) 12964–12975. doi:10.1039/C6CP01555G.
- [12] E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, Comparative evaluation of eight docking tools for docking and virtual screening accuracy, *Proteins Struct. Funct. Bioinforma.* 57 (2004) 225–242. doi:10.1002/prot.20149.
- [13] K. Onodera, K. Satou, H. Hirota, Evaluations of Molecular Docking Programs for Virtual Screening, *J. Chem. Inf. Model.* 47 (2007) 1609–1618. doi:10.1021/ci7000378.
- [14] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.* 31 (2010) 455–461. doi:10.1002/jcc.21334.
- [15] M.F. Sanner, Python: a programming language for software integration and development, *J. Mol. Graph. Model.* 17 (1999) 57–61. <http://europepmc.org/abstract/MED/10660911>.
- [16] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J. Comput. Chem.* 30 (2009) 2785–2791. doi:10.1002/jcc.21256.
- [17] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.* 47 (2004) 1739–1749. doi:10.1021/jm0306430.
- [18] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening, *J. Med. Chem.* 47 (2004) 1750–1759. doi:10.1021/jm030644s.

- [19] R.A. Friesner, R.B. Murphy, M.P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, P.C. Sanschagrin, D.T. Mainz, Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes, *J. Med. Chem.* 49 (2006) 6177–6196. doi:10.1021/jm051256o.
- [20] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, *J. Mol. Biol.* 161 (1982) 269–288. doi:https://doi.org/10.1016/0022-2836(82)90153-X.
- [21] T.J.A. Ewing, S. Makino, A.G. Skillman, I.D. Kuntz, DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases, *J. Comput. Aided. Mol. Des.* 15 (2001) 411–428. doi:10.1023/A:1011115820450.
- [22] D.T. Moustakas, P.T. Lang, S. Pegg, E. Pettersen, I.D. Kuntz, N. Brooijmans, R.C. Rizzo, Development and validation of a modular, extensible docking program: DOCK 5, *J. Comput. Aided. Mol. Des.* 20 (2006) 601–619. doi:10.1007/s10822-006-9060-4.
- [23] P.T. Lang, S.R. Brozell, S. Mukherjee, E.F. Pettersen, E.C. Meng, V. Thomas, R.C. Rizzo, D.A. Case, T.L. James, I.D. Kuntz, DOCK 6: combining techniques to model RNA–small molecule complexes, *RNA.* 15 (2009) 1219–1230. doi:10.1261/rna.1563609.
- [24] W.J. Allen, T.E. Balius, S. Mukherjee, S.R. Brozell, D.T. Moustakas, P.T. Lang, D.A. Case, I.D. Kuntz, R.C. Rizzo, DOCK 6: Impact of new features and current docking performance, *J. Comput. Chem.* 36 (2015) 1132–1156. doi:10.1002/jcc.23905.
- [25] N. Liu, Z. Xu, Using LeDock as a docking tool for computational drug design, *IOP Conf. Ser. Earth Environ. Sci.* 218 (2019) 12143. doi:10.1088/1755-1315/218/1/012143.
- [26] J.-M. Yang, Y.-F. Chen, T.-W. Shen, B.S. Kristal, D.F. Hsu, Consensus Scoring Criteria for Improving Enrichment in Virtual Screening, *J. Chem. Inf. Model.* 45 (2005) 1134–1146. doi:10.1021/ci050034w.
- [27] C. Konstantinou-Kirtay, J.B.O. Mitchell, J.A. Lumley, Scoring functions and enrichment: a case study on Hsp90, *BMC Bioinformatics.* 8 (2007) 27. doi:10.1186/1471-2105-8-27.

- [28] L. Xing, E. Hodgkin, Q. Liu, D. Sedlock, Evaluation and application of multiple scoring functions for a virtual screening experiment, *J. Comput. Aided. Mol. Des.* 18 (2004) 333–344. doi:10.1023/B:JCAM.0000047812.39758.ab.
- [29] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, Virtual Screening Workflow Development Guided by the “Receiver Operating Characteristic” Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4, *J. Med. Chem.* 48 (2005) 2534–2547. doi:10.1021/jm049092j.
- [30] J.-F. Truchon, C.I. Bayly, Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem, *J. Chem. Inf. Model.* 47 (2007) 488–508. doi:10.1021/ci600426e.
- [31] R.P. Sheridan, S.B. Singh, E.M. Fluder, S.K. Kearsley, Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1395–1406. doi:10.1021/ci0100144.
- [32] C. Empeur-Mot, H. Guillemain, A. Latouche, J.-F. Zagury, V. Viallon, M. Montes, Predictiveness curves in virtual screening, *J. Cheminform.* 7 (2015) 52. doi:10.1186/s13321-015-0100-8.
- [33] W. Zhao, K.E. Hevener, S.W. White, R.E. Lee, J.M. Boyett, A statistical framework to evaluate virtual screening, *BMC Bioinformatics.* 10 (2009) 225. doi:10.1186/1471-2105-10-225.
- [34] R.P. Goldberg, Survey of virtual machine research, *Computer (Long Beach, Calif.)* 7 (1974) 34–45. doi:10.1109/MC.1974.6323581.
- [35] T. Kind, T. Leamy, J.A. Leary, O. Fiehn, Software platform virtualization in chemistry research and university teaching, *J. Cheminform.* 1 (2009) 18. doi:10.1186/1758-2946-1-18.
- [36] O. Korb, P.W. Finn, G. Jones, The cloud and other new computational methods to improve molecular modelling, *Expert Opin. Drug Discov.* 9 (2014) 1121–1131. doi:10.1517/17460441.2014.941800.
- [37] D. Bullard, A. Gobbi, M.A. Lardy, C. Perkins, Z. Little, Hydra: A Self Regenerating High Performance Computing Grid for Drug Discovery, *J. Chem. Inf. Model.* 48 (2008) 811–816. doi:10.1021/ci700396b.
- [38] T. Gaillard, Evaluation of AutoDock and AutoDock Vina on the

- CASF-2013 Benchmark, *J. Chem. Inf. Model.* 58 (2018) 1697–1706. doi:10.1021/acs.jcim.8b00312.
- [39] G. Rastelli, L. Pinzi, Refinement and Rescoring of Virtual Screening Results, *Front. Chem.* 7 (2019) 498. doi:10.3389/fchem.2019.00498.
- [40] P.D. Lyne, M.L. Lamb, J.C. Saeh, Accurate Prediction of the Relative Potencies of Members of a Series of Kinase Inhibitors Using Molecular Docking and MM-GBSA Scoring, *J. Med. Chem.* 49 (2006) 4805–4808. doi:10.1021/jm060522a.
- [41] S. Genheden, U. Ryde, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities, *Expert Opin. Drug Discov.* 10 (2015) 449–461. doi:10.1517/17460441.2015.1032936.
- [42] B.R. Miller, T.D. McGee, J.M. Swails, N. Homeyer, H. Gohlke, A.E. Roitberg, MMPBSA.py: An Efficient Program for End-State Free Energy Calculations, *J. Chem. Theory Comput.* 8 (2012) 3314–3321. doi:10.1021/ct300418h.
- [43] D.L. Mobley, K.A. Dill, Binding of small-molecule ligands to proteins: “what you see” is not always “what you get,” *Structure.* 17 (2009) 489–498. doi:10.1016/j.str.2009.02.010.
- [44] J. Vrbanac, R. Slauter, Chapter 3 - ADME in Drug Discovery, in: A.S.B.T.-A.C.G. to T. in N.D.D. (Second E. Faqi (Ed.), Academic Press, Boston, 2017: pp. 39–67. doi:https://doi.org/10.1016/B978-0-12-803620-4.00003-7.
- [45] A.P. Li, Screening for human ADME/Tox drug properties in drug discovery, *Drug Discov. Today.* 6 (2001) 357–366. doi:https://doi.org/10.1016/S1359-6446(01)01712-3.
- [46] T. Sander, J. Freyss, M. Von Kor, C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, (2015). doi:10.1021/ci500588j.



CONCLUSION AND FUTURE OUTLOOK

Pharmacological intervention of 5-lipoxygenase (5-LOX) catalyzed leukotriene biosynthesis has been widely studied as a promising therapeutic strategy for acute inflammation, allergic and respiratory diseases. Due to the toxicity effect of the marketed 5-LOX inhibitor zileuton, the scientific community is looking for novel 5-LOX inhibitors. As a result, the significant and relevant amount of structure-activity information of 5-LOX inhibitors has been released and stored in public databases. Besides, the newly resolved crystal structure of stable human 5-LOX is published recently. Varieties of computational methods that are either based on protein structural information or pharmacological information from known inhibitors can be useful in recognizing, predicting, and screening novel potential 5-LOX inhibitors. So, in the study, we have used discipline like cheminformatics and computer-aided drug design for the rapid and efficient identification and prediction of potent therapeutic agents against this protein.

To begin with, we have carefully evaluated the crystal structure of Human 5-LOX protein. Among all, 5-LOX 's stable 3D structure

with a PDB ID 3O8Y was finalized as the target structure for the whole study. The docking of known 5-LOX antagonists with the ligand-binding sites shows that most of the interactions are hydrophobic, and some pi-pi stacking interactions and H-bonded interactions are also there. The known antagonist zileuton, satisfied all the structural requirements for receptor binding. Moreover, a 3D model of the 5-LOX receptor of *Rattus norvegicus* was also constructed, and refined by energy minimization. The prepared 5-LOX model was then validated using What IF RMS Z-scores, Errat plot, and the Ramachandran plot. All the results indicated the excellent quality of the developed homology model. Prepared Human 5-LOX protein and rat 5-LOX model was used for SBVS or Molecular docking studies.

Next, we aimed at the comprehensive cheminformatic characterization of the diversity and complexity of the chemical space of 5-LOX and FLAP inhibitors by comparing it with the Approved drug space and the virtual LOX library. Property space analysis indicated that the compounds in the 5-LOX and FLAP space are, in general, less or comparable polar and flexibility similar to that of drugs in the drug database. PCA results showed that properties associated with the polarity of the compound have a significant contribution toward each PC. The visual representation of the property space indicated some compounds in the 5-LOX inhibitors space broaden the traditional medicinal space. The structural diversity of the databases was computed using complementary approaches, including PCP descriptors, molecular fingerprints, and molecular scaffold. With the apparent exception of approved drugs, the 5-LOX dataset showed

more diversity compared to FLAP and LOX library set. FLAP inhibitor set was the least diverse set. SAR of the datasets was studied using activity landscape analysis and Chemotype Enrichment and found eight important activity cliff generators and some cyclic systems with a large proportion of active molecules. The smooth SAR region present in the 5-LOX chemical space opened up the possibility of the development of highly predictive and robust QSAR models.

In the following study, we have tried to develop robust and statistically significant CoMFA QSAR models to predict the 5-LOX inhibitory potency of redox inhibitors such as 3', 4'-dihydroxyflavones, 3, 4-dihydroxychalcones and benzoquinones by exploiting smooth SAR region of the structure-activity landscape. Moreover, extracted CoMFA contour maps provided the necessary hints of modification for the design of new molecules with better activity. Molecular docking analysis has also been carried out to examine the stability and rationality of the CoMFA models. We have identified docking results coincide well with the CoMFA result. This study helped us to understand that together molecular docking results and extracted contour maps could be used to design novel inhibitors with respect to the most active compound in the dataset.

Next, we have developed some QSAR classification models by incorporating all the complex, diverse structural scaffold and related bioactivity data of 5-LOX inhibitors using non-linear machine learning algorithms. Among the 52 ML model constructed, PowerMV-IG-kNN (k=5) model gave better predictive results and was then used for the

virtual screening e-Drug3D database. As a result, 43 potential hits were identified. Furthermore, molecular docking-based virtual screenings were also performed to rank these 43 hits and identified four hits such as Belinostat, Masoprocol, Mefloquine, and Sitagliptin with high potential activity against 5-LOX protein, which can be further evaluated by biological studies.

We have always fascinated to identify novel structural scaffolds that can inhibit 5-LOX protein effectively. So, next, we have conducted a virtual screening analysis of 2.7 million compounds obtained from ZINC15 databases. To find out more accurate scoring function for virtual screening, we have done a comparative assessment of four commonly used docking programs such as Glide XP, LeDock, DOCK6, AutoDock Vina, and Consensus model before virtual screening based on various matrices and curves, also, by benchmarking different computer power. Based on this, in early phases of virtual screening campaigns, the quickest docking program Autodock Vina with average docking accuracy was implemented while slower and most meticulous one 'Glide XP' was used in later stages and the best 15 molecules were then screened. These 15 virtual hits were then subjected to ADME and toxicity analysis resulting in good ADME score and good toxicity result except three compounds. We understood that these three compounds need to undergo advanced toxicity assessment before further analysis. Overall, this study suggested that fifteen potential hits are expected to have good antagonist activity against 5-LOX enzyme and could be used as potential leads for the development of 5-LOX inhibitors.

So, in the entire study, we have tried to develop QSAR models that can predict the 5-LOX inhibitory potency of any compounds and to expand the 5-LOX chemical space by identifying novel 5-LOX inhibitors through virtual screening. Several extensions of the current work may be possible in the future.

- *In vitro* analysis of potential 5-LOX hits that are screened through *in silico* study.
- Conduct further studies on cliff generators found in the activity cliff region of the structure-activity landscape of 5-LOX and FLAP inhibitors.
- Develop QSAR models of FLAP inhibitors by utilizing a smooth SAR region of the structure-activity landscape of FLAP inhibitors.
- Expand the FLAP chemical space by identifying novel FLAP inhibitors by virtual screening.

DETAILS OF PUBLICATIONS

Peer-reviewed publications

- T.K.Shameera Ahamed, K. Muraleedharan, Towards a systematic analysis of StructureActivity Relationship of 5-LOX inhibitors through Activity Landscape and Chemotype Enrichment, *Chemometrics and Intelligent Laboratory*, 207 (2020) 104188. <https://doi.org/10.1016/j.chemolab.2020.104188>.
- T.K.Shameera Ahamed, K. Muraleedharan, A cheminformatic study on chemical space characterization and diversity analysis of 5-LOX inhibitors, *J. Mol. Graph. Model.* 100 (2020) 107699. <https://doi.org/10.1016/j.jmgm.2020.107699>.
- T.K. Shameera Ahamed, K. Muraleedharan, A ligand-based comparative molecular field analysis (CoMFA) and homology model based molecular docking studies on 3', 4'-dihydroxyflavones as rat 5-lipoxygenase inhibitors: Design of new inhibitors, *Comput. Biol. Chem.* 71 (2017) 188–200. doi:<https://doi.org/10.1016/j.compbiolchem.2017.08.010>.
- T.K. Shameera Ahamed, V.K. Rajan, K. Sabira, K. Muraleedharan, QSAR classification-based virtual screening followed by molecular docking studies for identification of potential inhibitors of 5-lipoxygenase, *Comput. Biol. Chem.* 77 (2018)154–166. doi:<https://doi.org/10.1016/j.compbiolchem.2018.10.002>.
- T.K. Shameera Ahamed, V.K. Rajan, K. Sabira, K. Muraleedharan, DFT, and QTAIM based investigation on the structure and antioxidant behavior of lichen substances

Atranorin, Evernic acid and Diffractaic acid, *Comput. Biol. Chem.* 80 (2019) 66–78. doi:<https://doi.org/10.1016/j.compbiolchem.2019.03.009>.

- T.K. Shameera Ahamed, V.K. Rajan, K. Muraleedharan, QSAR modeling of benzoquinone derivatives as 5-lipoxygenase inhibitors, *Food Sci. Hum. Wellness.* 8 (2019) 53–62. doi:<https://doi.org/10.1016/j.fshw.2019.02.001>.
- T.A. Nibila, T.K. Shameera Ahamed, P.P. Soufeena, K. Muraleedharan, Pradeepan Periyat, K. Aravindakshan, (2020). Synthesis, structural characterization, Hirshfeld surface and DFT based reactivity, UV filter and NLO studies of Schiff base analogue of 4-aminoantipyrine. *Results in Chemistry.* 2. 100062. 10.1016/j.rechem.2020.100062.
- V.K. Rajan, T.K. Shameera Ahamed, K. Muraleedharan, Studies on the UV filtering and radical scavenging capacity of the bitter masking flavanone Eriodictyol, *J. Photochem. Photobiol. B.* 185 (2018) 254–261. doi:[10.1016/j.jphotobiol.2018.06.017](https://doi.org/10.1016/j.jphotobiol.2018.06.017).
- V.K. Rajan, T.K. Shameera Ahamed, H. CK, K. Muraleedharan, A non-toxic natural food colorant and antioxidant 'Peonidin' as a pH indicator: A TDDFT analysis, *Comput. Biol. Chem.* 76 (2018) 202–209. doi:<https://doi.org/10.1016/j.compbiolchem.2018.07.015>.
- V.K. Rajan, T.K. Shameera Ahamed., K. Muraleedharan, Data on the UV filtering and radical scavenging capacity of the bitter masking flavanone Eriodictyol, *Data Br.* 20 (2018) 981–985. doi:<https://doi.org/10.1016/j.dib.2018.08.149>.

- K.P. Safna Hussan, Mohamed Shahin Thayyil, T.K. Shameera Ahamed, K. Muraleedharan, (2020), Biological Evaluation and Molecular Docking Studies of Benzalkonium Ibuprofenate. 10.5772/intechopen.90191.

Conference proceedings

- T.K. Shameera Ahamed, K.P. Nasiyya, K. Muraleedharan, DFT based investigations on structure, UV-filter, and antioxidant behavior of a tridepsides, Gyrophoric Acid, Proceedings in National seminar on Recent Trends in Material Science, NSRTMS-2019, ISBN: 978-93-5391-561-2, Govt. College, Chittur, Palakkad, Kerala - 678104, India.
- T.K. Shameera Ahamed, K. Muraleedharan, DFT, QSAR, and docking study of polyoxygenated Dibenzofuran against Staphylococcus Aureus, Proceedings in National seminar on Advances in Biomedical Science and Engineering, 2016, 23, NIT Calicut.
- T.K. Shameera Ahamed, K. Muraleedharan, Ligand-based Comparative Molecular Field Analysis on 3, 4-Dihydroxychalcones as 5-lipoxygenase inhibitors Proceedings in International Conference on Emerging Frontiers in Chemical Sciences (EFCS-2017), 2017, 101, ISBN No. 978-93-5279-617-5, Farook College, Calicut.