PREDICTION OF CYTOCHROME P450-RELATED DRUG-DRUG INTERACTIONS BY DEEP LEARNING

by

Shan Lu

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Industrial and Physical Pharmacy

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Tongleu Li, Chair

Department of Industrial and Physical Pharmacy

Dr. Gregory Knipp

Department of Industrial and Physical Pharmacy

Dr. Sandro Matosevic

Department of Industrial and Physical Pharmacy

Dr. Vitaliy Rayz

Weldon Scool of Biomedical Engineering

Dr. Yoon Yeo

Department of Industrial and Physical Pharmacy

Approved by:

Dr. Rodolfo Pinal

To my family and Joe

ACKNOWLEDGMENTS

I would like to acknowledge and express my warmest thanks to those who have helped and supported me throughout the course of my graduate work. This dissertation would not have been possible without you.

First, I would like to give a special thanks to my committee chair Dr. Tonglei Li, for being a great mentor who guided me through my graduate career with invaluable support, countless hours and great patience. It has been an incredible journey having you as my advisor who always sharing the unique interdisciplinary views, helping me to expand my scientific knowledge, and challenging me to become a better researcher to always think broader and dig deeper. I appreciate for everything I have learned since the beginning of my graduate learning. The dedicated hard working to challenge for a better scientific understanding has invaluable inspiration to enable me becoming the scientist I am today. I have been incredibly privileged to have you as my mentor.

To my committee group, each of you has unique contributions, invaluable views, and generous support throughout my graduate study. To Dr. Gregory Knipp, thank you for sharing your unique perspective to my project, helping me to expand the applications and discovering greater values of my work, and steering me towards the right path. Also, generously sharing the cell lines supporting my research scientific exploration. To Dr. Sandro Mateosevic, I appreciate all the kindness and friendly support you have brought to my graduate process. Your broad knowledge basis inspired me to think critically have provided invaluable help throughout my project. To Dr. Vitaliy L. Rayz, thank you for contributing valuable advice from a different perspective and the inspiration of power of model and mathematics has been incredible. Dr. Yoon Yeo, thank you for your advising, trainings and generously sharing the laboratory equipment all the time. Your guidance and support are sincerely appreciated. Also, I appreciate all the training and help from Andrew.

Also, I would like to thank all the current and past members and visiting scholars from Dr. Tonglei Li's lab. Especially to Clairissa Corpstein, Nick Huls, Peace Umoru, and Yue Li. Thank you for your friendship and support. It has been a good journey with the laughs and pains shared. To visiting scholars Drs. Xingzheng Huang, Zhaohuan Lou, Zhengjie

Meng, and Xiaoyu Yu, I am very fortunate to have the opportunity to meet and work with everyone and grateful for all the experimental design training, research advices, and hands-on supports. The contributions in my early years of graduate training and research inspiration are invaluable.

To my classmates, who have been both colleagues, mentors, partners, and friends every step we take together. Thank you, Matthew Beyher, for the career experience and development advise shared. Tram Doc for always sharing the valuable information and opportunities. Dhawal Chobisa and Jianping wang for helping me with your research perspectives and sharing scientific thoughts all the time.

Finally, I would like to thank my family and friends. I am grateful to each of you and for your unconditional love and support. To my mother, Yuhong, the most compassionate, adventurous, and curious woman I have ever met. You have taught me how to face the challenges in life, to learn from mistakes, and to move forward with confidence. To my father, thank you for the endless support you have provided me throughout my life and decisions, whether they ended in triumph or disaster. To all my friends, who have always been there to support in various aspects and brought tons of joy through challenging times. Your grace continues to be my inspiration.

To Joe Chu, my best friend and beyond. It has been an incredible journey over the four years since we met. You have been the rock and foundation over these years. It would be impossible for me to stay sane over those days and nights without your support and endless calls. I am in awe of your intelligence, your hardworking, and your braveness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
CART	Classification Regression Tree
CDFT	Conceptual Density Functional Theory
CoMFA	Comparative Molecular Field Analysis
CYP	Cytochrome P450
DDI	Drug Drug Interaction
DFT	Density functional theory
DL	Deep Learning
DT	Decision Trees
ECFP	Extended-connectivity Fingerprint
ESP	Electrostatic Potential
FCFP	Functional-Class Fingerprint
qHTS	Quantitative High-throughput Screening
QSAR	Quantitative Structure Activity Relationship
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HSAB	Hard and Soft (Lewis) Acids and (Lewis) Bases
HTS	High-throughput Screening
KNN	k-nearest neighbor
LBVS	Ligand-based Virtual Screening
MACCS	Molecular ACCess System
MD	Molecular Dynamics
MEMS	Manifold Embedding of Molecular Surface
MLR	Multiple Linear Regression
NB	Naive Bayesian
NeRV	Neighbor retrieval visualizer
NIH	National Institutes of Health
NN	Neural Network

PC	Principal Component
PCA	Principle Component Analysis
PLS	Partial Least Squares
RBF	Radial Basis Functions
SAS	Solvent Accessible Surface
SBVS	Structure-based Virtual Screening
SES	Solvent Excluded Surface
SMILES	Simplified Molecular-input Line-entry System
SMF	Substructural Molecular Fragments
SNE	Stochastic Neighbor Embedding
SOM	Self-organizing Map
SPE	Stochastic Proximity Embedding
SPP	Similarity Property Principle
SVM	Support Vector Machines
UHTS	Ultra High-throughput Screening
VS	Virtual Screening
VWS	Van Der Valls

ABSTRACT

Drug-drug interactions (DDIs) occur when multiple drugs are used concurrently. Caused by one drug inhibiting or inducing the metabolism of a second drug, DDIs often alter plasma concentrations and could seriously impact efficacy and safety of co-administered medications. Cytochrome P450 (CYP), a superfamily of enzymes, plays an important role in metabolizing a majority of FDA approved drugs currently on the market. 70% of predicable DDIs are associated with CYP enzymes inhibition. In-silico methods are increasingly adopted as a cost-effective complement to guide and prioritize efforts in drug discovery. Recent emerging applications of artificial intelligence algorithms have demonstrated promising results capable of prioritizing the selection of large chemical libraries, thereby outlining the future of in-silico methods assisting in drug discovery. Nevertheless, current methods rely on molecular descriptors that almost exclusively focus on chemical properties and atomic structures that fail to capture critical conformation and biological interaction related properties. There is also a lack of trainable molecular descriptors with feature specificity that reflect detailed protein-ligand binding energy and enable biological activity prediction. The overall objective of this dissertation is to understand molecular biological binding activity through electronic structure-based local descriptors derived from quantum based conceptual density functional theory (CDFT). This method will be used to assess the correlation of intermolecular interaction energy with ligand-protein binding with 2D feature maps reduced from the 4D molecular surfaces of the binding site and ligand (3D molecular surface with 1D electronic property). Additionally, it will be used to explore the possibility of predicting CYP related DDIs using descriptors generated using first principles including protein-ligand binding with specificity and strength and deep learning algorithms. Using quantum chemistry to interpret topological molecular information residing on 3D molecular surface permits the extraction of interacting features directly from the ligand structure. To achieve that, a set of curatable data containing consistent measurements was accessed through publicly accessible libraries. A series of novel Manifold Embedding of Molecular Surface (MEMS) descriptors were generated containing local electronic properties residing on the 3D molecule structure surface of each ligand using manifold learning. Major information were captured featuring electronic characteristics on the molecular 3D surface. Shape context was employed to derive transnational invariance feature vectors from MEMS with high granularity, thus preserving molecular information with specificity. DeepSet was utilized to perform permutation equivariance model training and validation. Powerful model learning is observed with an F-measure for all targets above 75% with the highest of 87% from external testing. Despite their promising prediction performance, molecular conformation changes and analytical featurization methods need to be implemented to expand model applicability and improve model reliability.

KEYWORDS: Cytochrome P450, Deep learning, Quantum chemistry, Drug-drug interactions, Ligand-based virtual screening

1. INTRODUCTION

1.1 Drug-drug interaction and adverse drug reactions

Drug-drug interactions (DDI) describe a modification of the pharmacological effect of one drug by the prior or concomitant administration of another. When two or more drugs are used concurrently, one drug can inhibit or induce the metabolism of a second drug, thus either increasing or decreasing the effectiveness and side effects of a drug, or introducing a new side effect that was not previously observed (especially for drugs with narrow therapeutic windows). The elongated life expectancy and continuous advancement in chronic conditions management have also contributed to the increase of polypharmacy among other factors. Nowadays, with the rapidly growing usage of prescription drugs, the need to use multiple medications concurrently is common, as patients often develop comorbid multiple chronic diseases, especially in the elderly population [1], [2]. Statistics show that patients aged 70–79 years have a 34% risk of DDIs when they are prescribed with two or more drugs concomitantly [3]. Furthermore, it was estimated that 36% of elder Americans regularly used at least five medications or supplements simultaneously in 2010-2011 [4]. An increased potential adverse drug reactions due to DDIs have been emerging and becoming an unignorable risk for public health. It could pose serious impact on the safety and efficacy of affected medications [5], [6].

As the risk of DDIs increases, the incidence of severe adverse drug events including fatal cases also increases. Currently, study shows there are about two million severe adverse drug reactions reported in the United States alone. Among them, approximately 26% are shown to be attributed to avoidable DDIs [7]. Thus, early detection of potential DDIs is highly desirable for both pharmaceutical industry and pharmacovigillence systems to avoid possible failures in drug development and improve patient health. Pharmaceutical companies have been conducting many screening tests for new drug candidates to avoid unexpected DDIs in pre-clinical, clinical and post marketing phases during the research pipeline [8], [9]. Thanks to continuously developed data bases and increasing knowledge of DDI mechanisms, several literature reviews have described the possibility of using and applying data based in-silico modeling and applications of in-silico methods developed to predict drug absorption, distribution, metabolism, excretion and toxicity (ADMET) [10]–[13]. Our understanding of DDI at the molecular level has been yielding a large amount of experimental data and great understanding of mechanism pathways.

A variety of drug metabolizing enzymes and transporters have been identified and modeled as the key factors for DDIs over the past couple of decades. Cytochrome P450 (CYP), a superfamily of enzymes, have been identified as important enzymes for the metabolism of xenobiotics and most drugs that are currently on the market. Adverse CYP enzyme DDIs caused by drug co-administration have led to several withdrawals of commercial drugs from the market during the past several decades, such as Seldane, Posicor, Duract, Hismanal, Propulsid, Lotronex, Baycol, and Seraone [14]. US Food and Drug Administration (FDA) and the Pharmaceutical Research and Manufacturers of America (PhRMA) published several guidelines for industry, providing guidance for the application of both in-vitro and in-vivo studies, as well as in silico analysis that should be conducted during early stage of drug discovery to evaluate the cytochrome P450 enzyme and transporter-mediated drug interactions, particularly for members of the CYP superfamily [15], [16].

In an effort to reduce time and cost, various in-silico methods are continuously being considered and widely used in the pharmaceutical industry to provide more possibilities in assisting with decision-making in the drug discovery pipeline [17], [18]. Many screening tests for new drug entities to avoid unexpected DDIs in the pre-clinical, clinical and post marketing phase during the research pipeline have been designed and conducted [8], [9]. As such, our understanding of DDIs at the molecular level has been yielding a large amount of experimental data and a greater understanding of interaction mechanisms. Various in-silico methods are continuously considered and widely applied by scientists to provide possibilities assisting with decision-making and compound prioritization in the drug discovery pipeline [17], [18].

1.2 Role of CYP450 enzyme

1.2.1 Definition of CYP450 enzymes

The name "cytochrome P450 enzymes" (CYP450) is originated from the characteristics of the proteins. Containing heme pigment, chrome and P, they are bound to the membranes of cell. Further, when exposed to carbon monoxide, they absorbs light at a wavelength of approximately 450 nm. Be able to detoxificate foreign chemicals, these CYP450 enzymes have been identified as an essential enzyme family responsible for the metabolism of xenobiotics and most other drugs that are currently on the market. Among 57 CYP isoforms that have been identified in humans, six isoforms (CYP1A2, CYP2B6, CYP2C9, CYP2C19, CYP2D6, and CYP3A4) have proven to be accountable for the metabolism of 90% of FDA approved medications [17]. Most DDIs associated with CYPs are caused by CYP inhibition (73%) or induction (27%) [17].

1.2.2 Characteristics of CYP enzymes induction

Many of the human liver CYP enzymes, including CYP1A, CYP2B, CYP2C, and CYP3A families, are responsible for the metabolism of a large portion of pharmaceutical drugs. They are found to be inducible by a diverse array of compounds including medications, natural products, synthesized chemicals, and ethanol [19]. Induced enzyme activities can enhance metabolism of pharmaceutical drugs and alter drug plasma concentrations. Major advancements in understanding the mechanisms for induction of CYP enzymes have been made. Based on current understanding, CYP induction is proven to be a relatively slow phenomenon involving gene regulation activities and protein synthesis processes, such as nuclear receptor activation and the synthesis of mRNA and enzymes [20]. Similar mechanisms have been found for many CYP induction process. Alternative mechanisms of CYP induction is found to be related with chemicals that are able to stabilize the translation of an enzyme and inhibit protein degradation pathway [21].

1.2.3 Characteristics of CYP enzymes inhibition

Mechanisms of CYP inhibition is usually classified as two categories: reversible or irreversible (which is also known as quasi-irreversible). Reversible inhibitions are those caused by rapid association and dissociation between the enzyme and drugs or ligand substrates that can be categorized as competitive, non-competitive, or un-competitive inhibitors [21], [22]. Reversible competitive inhibition is associated with multiple substrates that can be metabolized by the same binding pockets of the CYP isoform. It happens when more than one substrates were used together and compete for thee same active binding regions on CYPs. When two substrates with competitive behaviors are co-administered, the competition can be understood as the function of their concentrations and respective binding affinities of the present ligands for the same active site of enzymes [23]. For competitive inhibition, a stronger substrate will have the ability to replace the substrate with a relatively weaker binding affinity. The metabolism of the weaker drug will have a decreased affinity and thus slower metabolism rate as well as clearance [24], [25]. Non-competitive inhibition happens when one of administered substrate binds to the allosteric site of an enzyme and cause a conformation change of the enzyme's structure and alter the active binding site for the other compound [26]. Although there is no direct competition between the two ligands, this type of inhibition is often concentration independent and will usually last longer. Un-competitive inhibitors bind only to the enzyme–substrate complex and lead to a stable form. While this is a less common phenomenon Un-competitive inhibitor is still observed in many cases and demonstrated great effects [27]–[29]. Another set of inhibitors is known as mixed inhibitors, which refer to ligands that can bind to both the active pockets and the allosteric sites. With those circumstances, both competitive and non-competitive inhibition can happen at the same time. Greater DDI effects are usually seen than those from competitive or non-competitive inhibition activities [23], [27].

Mechanism-based inhibition is an irreversible process commonly observed with pharmacokinetic interaction studies. Integrated by NADPH, mechanism-based inhibition is proven to be both time and concentration dependent enzyme inactivation and substrate metabolization [23], [30]. Mechanism-based inhibitions happen when the CYP enzyme activates the ligands to a reactive metabolite which can bind with the enzyme active site and form a stable complex. Those activities can be quasi-irreversible or irreversible. Some substrates can be metabolized by CYP isoforms and produce intermediate metabolites that are inhibitory to enzyme activities. In irreversible cases, the intermediate metabolites can form covalent bonds with the heme prosthetic site which result in irreversible inhibitions [31]. For quasi-irreversible inhibitions, the metabolites will usually develop a stable complex with the heme prosthetic site of the enzyme and form a metabolite-intermediate complex that force the enzyme to be inactive [32], [33]. Those type of formation can be disrupted, however, the exact mechanism involved in this process is not yet discovered.

Mechanism-based inhibition is usually time-dependent, where a plateau in the metabolization pathway of entities and inhibition rate plateau can often be reached over time depending on the concentration of ligands and available enzyme [34], [35]. Suicide inhibition is another scenario which describes the process by which highly reactive intermediates are formed during metabolic activation [36], [37]. The interactive intermediate complex is formed through strong irreversible covalent bonds with the targeted enzyme in an unusual pathway. The conformation of the enzyme will then alter significantly during the process and become inactivated and lead to acute inhibition activities [38], [39].

1.3 Computer-Aided Design in DDIs Prediction

The discovery of new drug entity begins with the identification of targets for a disease of interest with the receptors that the drug should act upon. Then, high-throughput screening (HTS) experiments are necessary to be applied to compound libraries to select the hit compounds that are active against the targets [40]. It is followed by an optimization process to obtain lead compounds with drug-like properties and increased potency. Drug absorption, distribution, metabolism, excretion and toxicicity (ADMET) are carefully evaluated during this process. The drugs that reach safety and efficacy goals will be brought to clinical trials after completing pre-clinical studies.

While HTS experiments are powerful techniques, they are often time- and resource-heavy require several thousands of compounds and advanced facilities, which is one of the major challenges in drug development process. More than 2 billion US dollars is needed on average for a new drug to be developed, and it takes at least 10 years to complete the journey [41]. Still, compounds could fail during any stages of development. As our understanding of the mechanisms of DDIs and the ability of predicting those interactions have advanced over the past years, the FDA has included various computational decision methods to assist the decision-making process. Computer aided design is an effective alternative that has been widely explored and incorporated in the process to guide and prioritize the process and minimize the effort in the past decades. In silico prediction methods are continuously being considered to evaluate DDIs by both pharmaceutical companies and worldwide pharmacovigillence systems [17], [18].

Virtual Screening (VS) is a computational technique used to identify structures that are likely to bind to a drug target by searching small molecules libraries based on biological structures [42]. It has become an important tool that many scientists choose to implement during the drug development process to facilitate with both in-vitro and in-vivo experiments. Whether used in conjunction with HTS or applied alone, the in-silico process provides a quick and economical alternative by searching over millions of potential compounds against interested compound libraries. VS is designed to effectively scan large compound libraries in a timely manner with managebal cost and prioritize the most potent compounds for further investigation.

VS methods are often summarized as structure-based methods, and ligand-based methods. Structure-based methods calculate the free energy of binding estimates based on protein structure whereas ligand-based methods perform molecular similarity searching based on ligand structures.

1.3.1 Structure-based method

Structure-based virtual screening (SBVS) methods attempt to predict the interaction and binding affinities between ligands against a molecular target [43]. To perform the calculation, 3D structure of the targeted protein is required to estimate interactions between the target and each chemical compound. To understand the binding activities and generate reliable predictions, information about the receptor structure will be needed. Compounds will be ranked and selected from the library based on their binding affinity calculation for certain receptor sites.

In 1982, Kuntz et al first developed the algorithms to explore the feasibility of geometrically assess the alignments between ligands and receptors for existing target structures [44]. It later became an increasingly important tool with constant advancement in technology and computing powers known as molecular docking (MD). MD is capable with predicting the interactions between ligands and target proteins at the atomic level. It can be used to characterize the interaction behavior of interested compounds with the context of target proteins. The main objective is reproducing experimental binding scenarios and evaluate the strength of ligand-receptor complex structures using in-silico prediction techniques.

Starting with sampling of the ligand conformations and orientations within the targeted sites of protein, molecular docking assesses their interaction strength with scoring function algorithms [45]. Various sampling methodologies have been developed to explore different positions of ligands at the binding site with six transnational and rotational degrees of freedom as well as different possible conformations. A large amount of possible binding complexes can be generated between ligands and targets, thus usually requireing large computational powers[46], [47]. Searching algorithms are usually widely incorporated into molecular docking software to improve searching efficiency. The incremental construction method is a systematic searching algorithm that exploits the degrees of freedom of the molecules in a fragment-based manner and docks these incrementally [43]. During the process, the interested molecules will first be divided into fragments with respect to their rotatable bonds. Then, docking will be performed with one of the randomly selected fragments against the known target active site. Other fragments will be incrementally added to the previously chosen fragments with different transformations in an attempt to fit the active site. Several software programs like DOCK 4.0 [48], FlexX [49], Glide [50], eHiTS [51], and so on have this type of algorithm available. Another commonly used algorithm is the stochastic search algorithm that changes one degree of freedom at a time randomly in the spatial conformation of ligand. Functions based on molecular physics are used to measure the free energy of the conformation to evaluate how favorable the conformation is regarding to the binding target site. The Monte-Carlo method, genetic algorithms are examples of this type of search. Software programs like Glide [50], MOE [52], GOLD [53] and AutoDock4 [54] all have those methods implemented. Deterministic search is another type of search algorithm that focuses on small conformational changes at the atomic level to analyze how the whole molecule behaves and to find different molecular dynamics simulations.

As a widely used deterministic simulation method, MD simulation takes into account of both the ligand flexibility and protein flexibility more effectively comparing with other algorithms [55]. MD simulations also bears with some disadvantages as calculation can take a long time and require great computational power. Also, MD simulations can be trapped in a local minima due to the smaller steps it takes. Besides that, they have been recognized as an efficient tool to perform local optimization which can be used as a good strategy following the random search techniques to capture subtle conformational changes.

Scoring functions are developed to estimate the interaction forces between ligands and molecular targets to capture the correct docking poses. It ranks the interested ligands with their single or multiple binding possibilities in the receptor structure based on their estimated binding affinity. Traditionally, scoring functions are divided to force-field-based, empirical-based, and knowledge-based scoring functions. Force-field-based scoring functions calculates the binding energy of intermolecular interactions between the ligands and targets. Typical forces including electrostatics, Van der Waals, hydrogen bonds, solvations, and entropy obtained from experimental data and molecular mechanics principles. Evaluation tools including software like DOCK [48], GOLD [53] and AutoDock4 [54]. Empirical scoring functions estimates binding energy by decomposing complexes into several components based on their force type. A final score is determined based on a coefficient assigned to each group. LUDI [56], PLP [57], ChemScore [58] and DOCK6 [59] have different empirical scoring functions ready to use. Knowledge-based scoring method can be used to screen large compound databases with statistical analyses that calculates the binding interactions of the protein atoms and molecular target [60]. These methods are developed based on the assumption that intermolecular interactions occurs more frequently near certain types of atom and some types of functional groups are likely to contribute more to the binding activities than other types. ParaDock5 [61], PMF [62], and DrugScore [63] are examples for those types of calculations. As advantages and disadvantages exist for different scoring functions, an emerging technique has been used known as consensus docking to apply different approaches to increase the accuracy of predictions. However, inaccurate prediction of binding energies still exists. Sampling algorithms and scoring functions calculate binding activities between the ligands and targets, and yield compounds that have higher likelihood to interact with designated targets.

1.3.2 Ligand-based method

The ligand-based virtual screening (LBVS) is another computational approach that uses only ligand information to perform predictions. These methods including pharmacophore maps and quantitative-structure-activity-relationship (QSAR) which can be applied to identify a lead only using chemical structure information based on molecular similarity [10], [64]. Usually, information can be extracted with known ligand categories and applied to both lead identification and optimization process. With an assumption that undiscovered active ligands would share similar functional groups and chemical features with the known active ligands. These types of methods begin with developing molecular representation of each of the known input molecules followed by evaluating similarities between known active ligands and unknown candidates to rank and identify potential compounds.

First developed in 1964 by Hansch and Fujita, QSAR has become an critical method in the pharmaceutical industry [65]. It is an efficient mathematical calculation to characterize both classification and regression prediction based on ligand structural similarity. Biological activities and physio-chemical properties are usually correlated with ligand structural properties. QSAR methods prioritize potential compounds based on their desirable activities predicted with in-silico models. These methods have been greatly advanced in the last decades with the development of machine learning techniques that utilize both supervised and un-supervised methods to identify ligands with higher similarities. Different dimensionality of the molecular representations have been utilized to identify relevant aspects of molecular properties and evaluate chemical similarities [66], [67]. Advanced statistical methods have been explored to enable rapid predictions with improved accuracy. More detailed introduction can be found in later sections.

1.3.3 Databases for virtual screening

Data availability and quality is critical for computer aided designs. Over the past decades, the development of high-throughput screening (HTS) and ultra-high-throughput screening (UHTS) have enabled aggregation of a large volume of data regarding binding activities of drug ligands to different targets [68]. Several databases, both from public resources and internal experimental assays were selected and applied to build machine learning based models to explore both ADME and several toxicities (hepatotoxicity, cardiotoxicity, renal toxicity, genotoxicity) properties. Public libraries and national laboratories along with research institutes, as shown in Table 1.1, have been developing and managing ultra-large virtual molecular libraries through open chemical space which provides indispensable experimental basis for computational modeling.

1.3.4 Molecular fingerprints

With the numerous datasets and entries available, the next question is how to extract and convert the most relevant information to something computer interpretable. In the early 1990s, the similarity property principle (SPP) was proposed with molecular similarity analysis. SPP states that structurally similar compounds should have similar properties, with the most frequently studied property being biological activity [88]. Molecular descriptors, which are mathematical representations of molecules' properties for computer based studies, are typically developed to represent features based on SPP. Various descriptors have been calculated covering 0D - 4D feature for compound screening. Evaluated from the chemical structure, molecular descriptors provide an abundance of crucial chemical and physical information about the compounds with mathematical vectors and matrices. Molecules of interests are characterized and described by symbols or vectors to effectively execute the prediction tasks. Several commonly proven reliable approaches to calculate molecular fingerprints are as shown in Table 1.2.

Substructure keyed-based fingerprints transformation is calculated based on molecule substructures or features that is presence in the compounds [97]. Given a list of key properties for the sub-structural information, a set of bits and bit strings could be calculated by splitting a molecule into several fragments characterized as different descriptors. These types of molecular descriptors are useful when the molecules of interest are most likely covered by previously existing structural keys. However, those compound features are less informative if the compounds do not contain the designated structural keys. Some of the commonly used substructure keys-based fingerprints are listed in Table 1.2 and elaborated below.

Molecular ACCess System (MACCS) keys is another type of the most commonly used structural key based calcualtion. Also referred to as the MDL keys, their name is derived from the developer team (the MDL Information Systems (now BIOVIA)) [89]. Those descriptors come with two sets, one with a set of 960 keys and the other with 166 structural keys. The subset with 166 keys version is more accessible to the public and therefore more commonly used. It is relatively smaller in size, yet still is able to cover most of the interesting chemical properties for drug compound discovery. The calculations are implemented using open source software packages, like RDkit [98], OpenBabel [99], and so on.

PubChem fingerprint is used to find identical PubChem records to identify different levels of "sameness" through consideration of structural connectivity and the occurrence of isotopic and stereo-chemical information [90]. It contains 881 structural keys that cover a set of extensive chemical substructures and molecule features. Besides PubChem database implementation, it can also be accessed with ChemFP [100] and CDK [101].

ISIDA SMF descriptors are substructural molecular fragments (SMF) based descriptors that were developed with an in-silico design and data analysis (ISIDA) software for in-silico calculation based on fragment and pharmacophoric descriptors [91]. This approach would divide molecule of interest into fragments and characterize them as a set of descriptors. The mathematical representation for the descriptors could then be generated based on the sum of existences over these fragments within each of the molecules. Two fragment types are considered and generated as ”sequences” and ”augmented atoms”. Three sub-fragment types, A (atom only), B (bond only), and AB (atom and bond) can be defined depending on the interest and existing understanding.

BCI fingerprints are dictionary-based fingerprints with 1052 standard bits structure. Modification about bits size can be made by the user and the calculation is available with BCI toolkits from Digital Chemistry [92].

Topological fingerprints are able to analyze the whole molecular structure with fragments based manner following specific pathways of a certain parameters and hash the paths too generate a unique fingerprints with predetermined parameters. There are two paths the analyzation can follow, either a linear path or a circular path, which is also know as circular fingerprints [102]. Unlike the substructure keys based method, topological fingerprints are applicable for all type of molecules and are able to calculate meaningful fingerprints. These types of calculations can also be applied for sub-structural property searching and keys filtering. The fingerprints are hashed, meaning that with a single bit from the molecule fingerprints one cannot trace back to the given feature it represents. More than one features can be hashed to one bit, which is referred to as ”bit collision”.

The most prominently used topological fingerprints that look at the molecule with linear path is the Daylight fingerprints, which consists of up to 2048 bits encoding connectivity pathways through a molecule up to a given length [93]. The algorithm is molecule based and implemented in most commonly used software packages like RDkit [98].

Instead of following the atom in a linearly manner, circular fingerprints interpret the molecules with a pre-determined radius per the user’s preference. The surrounding of each atom up to the radius will be memorized and calculated. These types of molecular descriptors are widely used for full structural similarity searching.

Extended-connectivity fingerprints (ECFPs) is one of the circular topological fingerprints for molecular characterization developed in 2010 by Rogers et al [66]. Unlike other substructure similarity calculated topological fingerprints developed based on linear paths, ECFPs were calculated based on Morgan algorithm that can can be developed in the following three steps: 1. Assign an integer identifier to each atom. 2. Update each atom identifier to characterize neighbors of each atom. 3. Remove the duplicated identifiers calculated when the same atom feature occurred more than once. ECFPs are fingerprints designed specifically for the structure–activity based modeling calculating with circular manner and can be used to extract an infinite number of different molecular features. Those fingerprints represent atomic neighborhoods within the circular radius and is able to consist of different lengths. With above mentioned flexibility, this method is easy to customize and calculate a varietytypes mathematical representations describing particular sub-structures; those features represent the existence of particular sub-structures which allow easy interpretation of the analysis results. Calculation of ECFPs can be done using RDkit [98], CDK [101], and Chemaxon's JChem [103].

Functional-Class Fingerprints(FCFP) are a similar class to ECFPs [66]. As the name states, the FCFP rule is generated based on the functional atom groups with a molecule. Instead of assigning integer identifiers to the atoms, FCFP indexes the atom's rule within the molecules. Generated with pharmacophoric sets of initial atom identifiers that are identified to be similar with the catalyst pharmacophore identifier, FCFPs can be better representatives of where abstraction can be useful when two atoms play equivalent roles within the functional group. When calculating FCFPs, each atom is recognized with a 6-bit code associated with different roles including hydrogen-bond acceptor and donor, negatively and positively ionizable, aromatic, and halogen. If associated roles are identified, the given bits will be labeled. FCFP calculation is supported with open-source software packages as well [98], [103].

Hybrid fingerprints are also developed that combine different aspects and bits strings using different approaches. Those fingerprints are usually more flexible and can be modified based on user's need. Some of the commonly used fingerprints are summarized in Table 1.2 and discussed below.

Dragon descriptors are a reference as a variety of molecular descriptors (NO. of 1664 to date known as Dragon descriptors), covering 0D - 4D descriptors arranged into 20 blocks. Derived from different molecular representations, Dragon software produce a user-friendly interface for researchers [94]. The 0D descriptors contain any molecule properties that are independent of information regarding molecular structure. Examples of 0D descriptors are total atom count, number of specific atom types, number of specific bond types, and so on. 1D descriptors are derived from one - dimensional characteristics of a molecule including a series of fragments or functional groups existed. 2D descriptors are calculated based on two - dimensional topological information extracted from the molecule. The 3D descriptors are characterized based on three-dimensional representations of the molecule. WHIM descriptors are typical 3D descriptors calculated from three - dimensional molecular structures describing molecular properties from different perspective. They contain information for the whole 3D structure including their size, shape, symmetry, and atom coordinates [104]. Four - dimensional descriptors, also know as grid based descriptors, include an extra fourth dimension upon molecular 3D geometry. The additional dimension provides more information better characterizing the biological activities between liands and targets. With the extra information, 4D descriptors are able to differentiate activities for structurally different molecules but are also more difficult to compute.

Another type of descriptor is known as E-state descriptors that use a combination of electronic and topological properties to describe molecular properties. It calculates ”E-state indices” for boht atom type and molecular bond to generate relevant information for molecular descriptions as stated in the articles published by the original inventors Hall and Kier. An E-state variable is a representation for each atom in the molecule and encloses the molecular electronic properties of each atom as their response to perturbation of other atoms in the molecule. Topological characteristics of the molecule were also considered [95]. The descriptor take the sum of calculations for all the atoms and bonds in the molecule to describe a whole structure and can be calculated with Molconn-Z software.

Other hybrid fingerprints like Unity 2D that is a 988-bit long developed based on structural keys and connectivity path fragments properties [96].

Despite the abundance of the descriptors that have been developed, these descriptors are untrainable and do not have specificity. Choosing the correct combination of descriptors is an unsolved problem with no clear process. Obvious limitations have been observed with the current descriptors [105]. Mounting evidence suggests that compounds that might not be considered similar by a chemist can have similar biological activity and vice versa [106]–[108]. Minimal chemical structurally differences can have large activity differences, like the ”magic methyl effect” where replacing a hydrogen atom by a methyl group can cause a 10-fold activity boost, even 100-fold in some cases [109]. Trainable descriptors is needed to capture major information describing interactions with biological targets.

1.4 Machine Learning in Virtual Screening

Machine learning, a branch of artificial intelligence, performs data analysis that automates model developing. The goal for machine learning is to have the systems learn from given data and be able to extract certain patterns for problem solving with minimal human interventions.

In the field of virtual screening, machine learning assembles a predetermined compounds as training set with known activities [110]. It is known to be a cost-effective and fast achieving approach for predicting substrates that have the potential to pose biological activities for drug candidates during the early stage of drug discovery process. Usually, molecules with historically categorical or continuous information are often used as data objects to perform classification or regression problems. Classification models can be used to discriminate substrates from non-substrates and inhibitors from non-inhibitors with model performance evaluation. Regression approaches use a series of quantitative experimental values such as IC50, K_m, K_i to performance binding affinity prediction of substrates and inhibitors of interest. These methods are also known as supervised machine learning, and they differs from un-supervised learning in that they emphasizes expertise to complete a task without an expected label.

Supervised machine learning is trained with data containing both input and output targets. With input information available for each of the molecules, the model is able to generate some function by comparing the predicted output with an actual label. The function can then be applied to data without known output. Some earlier methods that have been applied include various linear regression approaches, such as multiple linear regression (MLR), partial least squares (PLS), classification regression tree (CART) and etc have being applied. As the continuous development and advancement of machine learning algorithms along with the accumulated experimental datasets continues, recent methods have been explored to predict DDIs, such as support vector machines (SVM), neural networkS (NN), and so on. [111]–[121]. Some of the representative methods that have not been proven effective will be introduced.

1.4.1 Linear Models

Linear regression is one of the most commonly used supervised machine learning algorithms developed during the early days. The objective of the model is to identify the best linearly fitted relationship between the dependent and independent variables and minimize the error between predicted and actual values. A simple linear regression contains only one independent variable, whereas, a multiple linear regression (MLR) contains two or more independent variables. MLR has been widely used in molecular virtual screening to analyze the effect of one or more explanatory variables on a single response variable such as potency [111], [122]–[125]. Given multiple independent variables x_1, x_2, x_3, ..., x_n, dependent variable y is predicted with the weighted equation 1.1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + ... + \beta_n x_n + \epsilon \tag{1.1}$$

where each of the β_n represents a slope coefficient of independent variable x_n with ϵ representing residuals. In MLR, the least squares method is typically used to minimize the sum of the squares of the residuals. MLR can be a good method to implement when the factors are non-redundant with well established relationships with the responses. However, in the field of virtual screening, the variable relationships are not always well understood. Some commonly used 3D QSAR techniques, like Comparative Molecular Field Analysis (CoMFA), generate descriptive values calculated at each point of their 3D grid of the molecular structures and often have a large number of variables. Those variables can be highly correlated and makes MLR inefficient and lead to over-fitting. Partial least squares (PLS) is an advanced linear method to implement when col-linearity is observed. First developed in the 1960's as an econometric technique by Herman Wold, it was well adopted in many different areas. When a large number of independent variables are applied, PLS methods reduce the explanatory information to smaller size and linear combinations. Dependent variables are evaluated to identify only a few underlying factors that account for most of the variation in the response. Chohan and Roy etc showed examples of using PLS methods to perform predictions on small datasets with oral drugs and Flavonoids, respectively [111], [112]. Besides those, Gedeck and colleagues mentioned the issues for data accessibility and published models developed with a large data collections from company internal resources using the PLS method [113]. Nine different 2D and 3D descriptors were applied and yielded the highest r^2_{pred} yielded was 0.91.

1.4.2 Naive Bayesian

Naive Bayesian (NB) is one of the machine learning classifiers derived based on Baye's theorem [114]. It describes the outcome probabilities of dependent variables using conditional probability as equation 1.2 describes:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \tag{1.2}$$

The NB classifier is generated given a set of training instances from a certain class, e.g., inhibitor or non-inhibitor. Given class A, a bag of features is employed to understand the instances. The probability of occurrence for a given feature B is calculated based on information in the training dataset. The model counts the number of appearances of a given instances for each classes. Assuming the descriptors in the training sets are equally important and independent of each other, each descriptor is considered to be proportional to the ratio of different classes and overall probability for each class is evaluated by the product of each descriptor probability as shown in equation 1.3:

$$\Pr(A \mid B_{1,2,3,\ldots,n}) = \Pr(B_1 \mid A) * \Pr(B_2 \mid A) * \ldots * \Pr(B_n \mid A)\frac{\Pr(A)\Pr(A)}{\Pr(B_{1,2,3,\ldots,n})} \tag{1.3}$$

Categorical outcome is required for this type of classifier. Compared with clustering methods, this method can effectively avoids "the curse of dimensionality" observed in high dimensional spaces which becomes a good alternative for LBVS [126]. Molecular descriptors like LogP values, molecular weight and those calculated with software like Discovery Studio [127], MOE [128], RDkit [98] can be used as attributes to perform classification. NB classifiers can also identify highly correlated features to each class which allows better guidance for compounds design.

A number of VS modeling techniques have been developed using NB classifiers. The first model using Bayesian inference techniques was introduced in 1999 by Labute using Binary QSAR to classify 1,947 small molecules to either active or inactive [115]. Sean and co-workers describe their attempts of using NB classifiers to illustrate the prediction of hERG channel blocking and inhibition activities for cytochrome P450 2D6 [13]. The model correctly predicted more than 80% of compounds. Other approaches, including support vector machines and neural networks were comobined with NB classifiers to enhace model performance. Some structure-based VS studies were carried out by Jacobsson and colleagues [121]. Three types statistical methods, PLS discriminant analysis, rule-based methods, and Bayesian classification, were applied in their study to developed to construct classifiers to distinguish active and inactive compounds for the estrogen receptor α (ER), matrix metalloprotease 3 (MMP3), factor Xa (fXa), and acetylcholine esterase (AChE), respectively.

1.4.3 Nearest Neighbors

Nearest neighbors are one of the simplest classification methods to generate activity prediction of a molecule [129]. The algorithm assumes that similar compounds exist in close proximity. When data are projected into a multidimensional feature space, the class of interested molecules are likely to be the same as its nearest neighbors. 'k-nearest neighbor' (KNN) is an extension to the scheme using k neighbors.

When searching a database for molecules, a set of feature vectors describing molecular properties is obtained. The training set molecules are mapped to a high dimensional space as well as the test set. kNN performs classification for the test sets by assigning a test point to the class that is most similar to itself out of the k closest points. The critical part is to choose the right features and k value. K is a predetermined number by experimenting with different values in the algorithm and cross validating to reduce the error encountered. A typical range is between 1 and 10 [130]. Odd numbers can be used to prevent ties. Low values for K can be subject to outliers where a large K value would smooth over things; however, if too large, it will be out voted by other categories [130]. Euclidean distance is the most commonly used to measure distance between points, although sometime Manhattan distance can be used as well. Several attempts have been made to apply kNN in ligands classification. Different weighing functions have been applied to evaluating the distance between the molecule and its neighbors. Jensen et al. used Gaussian kernel weighted KNN models to classify CYP2D6 and CYP3A4 inhibitors [131]. Their model calculated Tanimoto distance based on extended connectivity fingerprints (ECFP) and FCFP. 865 and 1037 compounds were analyzed with CYP2D6 and CYP3A4, respectively. The model was able to correctly predict 83% for the training set, 82% of the test set compounds for CYP2D6 and 87% and 88% for CYP3A4. Additionally, 14 CYP2D6 external compounds were tested with 6 out of 14 from medium inhibitors failed to classify.

1.4.4 Support Vector Machine

First introduced in 1992 by Vapnik and colleagues [132], [133], support vector machine (SVM) became one of the most robust prediction algorithms that can be used for both classification and regression predictions. The idea is developed based on statistical learning frameworks also known as Vapnik–Chervonenkis theory proposed by Vapnik and Chervonenkis [134]. In general, input data can be represented as features vectors. A set of hyperplanes with n-dimensional is then constructed by an SVM to separates data into different spaces, which can be used for classification and regression [135]. Ideally, a good hyperplane separation in the feature space would be able to have all the points belong in to one class on the same side with linear dataset and distances between the nearest training data points of any class maximized. The subset of training points, also known as support vectors, are usually used to define the boundary.

Non-linear SVM problems can be efficiently performed using the ”kernel trick” that is developed from Mercer's theorem [132]. Commonly used kernel functions are polynomial, radial basis functions (RBF), and sigmoid functions. With the kernel functions, SVM algorithms can be applied to fit the maximum-margin hyperplane for nonlinear transformed high dimensional feature space.

In 2001, Burbidge and co-workers first explored it with drug molecule design predicting the inhibition of dihydrofolate reductase by pyrimidines [136]. Since then, there have been significant developments and applications in the pharmaceutical industry. In machine learning for VS experiments, many models have been successfully developed for compound identification. SVM models is shown to be able to minimize empirical classification errors and maximizes geometric margin. Further, SVMs can achieve reliable model prediction even if the descriptors are redundant [137].

1.4.5 Decision Tree

Decision Trees (DTs) are a type of non-parametric supervised learning method developed to design rapid interpretable filters for classification and regression problems [138]. With a known set of criteria or decision rules, DTs are composed using branch nodes that split paths [139]. Starting from the very top node of the tree (root node), two more more nodes are linked based on classification rules. The other linked nodes are internal nodes containing feature points. The un-linked nodes are leaf nodes that determine the classification. A tree can be treated as a piece-wise constant approximation. The path of the decision follows with series of ”if...then” decisions. It is crucial to organize more discrimination features to be applied first to split the dataset with lower impurity. Gini Impurity is a one of the most popular feature selection methods that measure the likelihood of incorrect classification with random variables. With C total classes, $p(\mathrm{i})$ is the probability of picking a datapoint within class i, the Gini Impurity (G) score can be calculated as seen in Equation 1.4:

$$G = \sum_{\mathrm{i}=1}^{C} p(\mathrm{i}) * (1 - p(\mathrm{i})) \tag{1.4}$$

A perfect split would yield 0 impurities meaning that there are no chances that any random sample will be mis-classified with the split decision, whereas a gini impurity of 1 implies an absolute mis-classification [140], [141]. Gini impurity for each node is calculated and the lowest score is selected as the leaf nodes. Other feature evaluation methods were also developed including information gain, bootstrap aggregation, and so on. Calculations for feature determination is repeated to achieve purer subsets until predetermined termination conditions are met. DT models are usually sensitive to the descriptors chosen and the composition of datasets. The advantage of using such model is their great interpretability coming from decision features and thresholds, therefore they have attracted great attentions in molecular VS models, especially in some of the earlier works.

Ekins et al. presented some of the typical rapid and simple filters developed with DT methods. In 2003, they reported models developed based on a commercially accessible data set containing more than 1750 molecules in the training set. In-silico models to predict inhibitors and non-inhibitors for for CYP2D6 and CYP3A4 were developed [142]. Atom augmented descriptors were applied with models with 20 random trees. With simple rankings, they were able to reach the result of Spearman's rho of 0.61 and 0.48 for CYP2D6 and CYP3A4, respectively for their test sets with R^2 0.88 and 0.82. Using a similar algorithm, they developed another model based on 875 structurally diverse molecules with in-vitro clearance data for general human liver microsomal metabolic stability prediction [143]. The model generated $r^2 = 0.71$ for the training set and r^2 =0.34 for testing with Spearman's rho -0.64. Both of those models demonstrated the ability of performing efficient calculation with DTs for compounds especially with high potency.

Burton et al. developed a study exploring the effects of the data quality used combining with a statistical training method to predict inhibition activities for the CYP1A2 and CYP2D6 inhibition [125]. The data were curated from literature by Aureus−Pharma containing well constructed databases with high quality of experimental measurements concerning inhibition information for the targets. With DTs and MOE descriptors, the authors were able to generate models with accuracy greater than 80% for the training set. Their CYP2D6 datasets were applied to develop 11 models with accuracies of over 80%. The CYP1A2 datasets generated total of 5 models with high-accuracy. Among those, the best result for validation set yielded 88% and 81% accuracy for CYP2D6 and CYP1A2, respectively. Their results demonstrated the importance of data quality. Comparing with data collected in traditional ways, simple DTs can achieve promising results.

1.4.6 Neural Network

Neural networks (NNs) attempt to identify underlying relationships embedded in a dataset by activates or deactivates a series of neurons mimicking the knowledge acquisition process of the human brain. Inspired by the mechanism of how a biological neuron interpret sensory information and pick up a task, neural network architectures contain layers of interconnected nodes to perform mathematical calculations and signal passing. Each node is also known as perceptrons and collect input signals and transform them into output signals. A basic NN is composed of an input layer, an output layer, and one or more hidden layers in between. Those layers of perceptrons are linked with each other in different ways to construct unknown complex relationships from input to output. The output information generated from one layer of perceptrons is then passed on as an input information for the next layer. Each perceptron connects to each another with an associated trainable weight. and bias is activated with enhanced signals if feedback is prompted. NNs rely on feeding in training signals to learn and improve prediction accuracy by adjusting weights through trials and to generate predictions for unseen inputs once properly tuned. The weights are trained through backpropagation, the essential mechanism of NN, whose aim is to minimize a loss function toward global minimization of the output error from the predictions and true values calculated based on training dataset. Gradient descent is the key to minimizing the prediction error and predict closely to the true value by evaluating the derivative of the loss function. Gradients will be positive as the loss increases with an increase in weight and negative if the loss decreases with a decrease in weight. With that mechanism, the model is always tuned toward negative gradients to minimize the loss. A large number of neurons are connected and tuned automatically, models can take an input and generate reliable output response in an efficient way with limited human intervention. In 1988, Hornik demonstrated the ”Universal Approximation Theorem” proving the universality of NNs which states that there is always a network that can approximately approach the result no matter what the function is [144]. It provides great advantages for NNs in solving complicated non-linear relationships with a minimum knowledge of the mechanism. There have been many successful applications of NN in drug discovery including classification and regression prediction of biological activities [145]–[147], pharmacokinetics activities [148], [149], and compound discovery with multi-dimensional data analyses [150]–[152].

Obviously, no perfect approach has been developed. Unexpected DDIs still can occur with respect to several aspects that we cannot measure or have not yet discovered. Several concerns have been raised together with the development. of these predictive models. One of the possible pitfalls lies with the data content and data design, especially with reference to meta analysis. Molecular/target interpretation and feature extraction has continue to be a challenging aspect. Despite the various published molecular fingerprints and descriptors, descriptor performance has been much more difficult to evaluate. The question, then, is how do we prioritize the molecular/target properties? Which aspects weights more than others? What does biological similarity mean? The definition of pharmacophore features and similarity comparison needs to be carefully evaluated. Evidence has shown that atoms in the same structural group do not have to own the same function in different structures. For example, oxygen atoms does not behave as an HBA when it is presents in an ester, or embedded with a furan ring, etc [153]. Not all molecular fingerprints can properly capture those properties and the importance of the information. Chemical properties may not represents biological similarity and activity. Although complicated and comprehensive molecular descriptors have been developed, the importance of each property has not been carefully looked at. It is unclear what parameters are deemed critical which may result in noise and sometime false signals for the prediction.

Another common critique for the current advancing machine learning and deep learning methods is the black box nature that limits the clear interpretation and identification of what features/fragments are critical. Novel and comprehensive learning frameworks have been introduced to predict drug target interactions operating on graph-structured data with convolution and attention mechanism based deep neural network architectures. The development of advancing algorithms, which have already provided promising results in various areas (like character recognizing and voice recognizing), demonstrate great potential in DDI prediction and drug discovery. Methods using various descriptors and fingerprints that result in the features used in modeling are not yet interpretive. Obviously, many improvements can be made with the current methods to better characterize the phenomenon and identify potential drug candidates or interactions with acceptable ADMET properties for drug discovery. One can definitely design a program against any target set and library with the right tool, which means opportunity and risk at the same time. The challenging objective is to develop an applicable model with trainable features to describe biological activity with specificity and strength.

Besides the current abundant report of dosing accident and clinical adverse reactions, there are also cases where DDIs can be put to good uses to enhance the drugs' pharmacokinetics that are extensively metabolized by CYP P450 enzymes. There are also occurrences where certain inhibitors are co-administered to prevent undesired peripheral metabolism [14], [154], [155]. Without a doubt, all of those activities require a comprehensive understanding of enzyme metabolism and compound DDIs. The application of in-silico prediction of DDIs should certainly be expanded. Despite the fact that computational methods that have been developed already are greatly contributed to our knowledge base and current understanding. They have been applied to discover many new inhibitors and ligands for series of biological targets activities that can be applied to further extended current understanding of drug-target and drug-drug interactions as well as even guiding new compounds design [156].

Table 1.1. Examples of public databases integrating ultra-large molecular data for computational modeling.

Database	Description	Reference
BindingDB	BindingDB is a freely accessible database containing protein-ligand complexes binding affinities measurement extracted from literature focusing on interactions of drug-targets proteins with small molecule ligands.	[69]
BioGRID	BioGRID is a publicly accessible biomedical interaction repository containing protein and genetic interactions curated from literature for multiple species. Bio-Grid content includes both low-throughput and high-throughput studies, gene interactions and post-translational modifications.	[70]
ChEMBLE	ChEMBL is a collection of manually curated drug-like molecules with their chemical properties of small molecules and their biological activity extracted from literature. Biological data are also exchangeable with other databases like BindingDB.	[71]
Chembridge	Chembridge libraries is a private product for the public that offers a 50,000 diverse small molecule dataset for drug discovery chemical compounds.	[72]
DrugBank	DrugBank is a freely accessible web-based database with comprehensive information about drug compounds and targets. The library includes broad information ranging like chemical properties, drug metabolites, drug gene expression and protein expression.	[73]
Electron Microscopy Data Bank (EMDB)	EMDB is a public database contains electron microscopy density maps of macromolecular complexes and sub-cellular structures of biological specimens.	[74]
Enamine REAL database	Enamine Real database currently the largest catalog containing synthetically drug-like molecules complying the ”rule of 5”.	[75]
Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database	A knowledge base resource storing genome sequencing and HTP dataset for biological system.	[76]
Online Mendelian Inheritance in Man (OMIM)	OMIM is a comprehensive collection of human genes and genetic disorders focusing on the relationship between gene and phenotype.	[77]
PDBbind	PDBbind is a public database collects binding affinity measured from protein-ligand experiment lined with Protein Data Bank (PDB) and provide information between protein-ligand complexes structure and interaction energy.	[78]
PharMapper	PharMapper is a publicly accessible tool that can be used to identify potential binding sites on target surface for the query molecule and perform ranking for druggability.	[79]
Protein Database (PDB)	PDB summarizes 3D shapes information of proteins, nucleic acids, and complex designed to facility with various scientific aspects with protein synthesis and structural information.	[80]
PubChem	Pubchem is a public collection of small molecules containing chemical properties and biological information from experiments by organization and individual contributors.	[81]
STRING	STRING is a database containing protein-protein interactions from both published literature source and model predictions to facilitate with a systematic understanding of cellular processes.	[82]
TCM-Database@ Taiwan	The TCM database is the world largest freely accessible small molecular repository constructed based on publications about traditional Chinese medicine.	[83]
Therapeutic Targets Database (TTD)	TTD is a comprehensive free accessible library provide curated data including discovered therapeutic protein and nucleic acid targets.	[84], [85]
UniPort	UniPort is a collection of information on proteins with protein sequence and functional knowledge mainly derived form gene sequencing and literature.	[86]
ZINC	ZINC database curates over commercially available molecules including their 3D structures and is designed toward to use of virtual screening. ZINC is heavily investigated by both pharmaceutical industry and research universities.	[87]

Table 1.2. Summary of typical molecular fingerprints for similarity search in virtual screening published in literature.

Class of Fingerprints	**Type of Fingerprints**	**Accessibility**	**Reference**
Substructure keys-based fingerprints	MACCS	RDkit, OpenBeble	[89]
	PubChem Fingerprints	ChemFP, CDK	[90]
	ISIDA SMF	ISIDA	[91]
Topological fingerprints	BCI	Digital Chemistry	[92]
	Daylight Fingerprints	RDkit	[93]
	ECFP	RDkit, CDK, Chemaxon's Jchem	[66]
	FCFP	RDkit, CDK, Chemaxon's Jchem	[66]
Hybrid fingerprints	Dragon descriptor	Dragon	[94]
	E-state Descriptor	Molconn-Z	[95]
	Unity 2D	Certara	[96]

2. AIM OF THIS STUDY

DDIs are often caused by concurrent use of multiple drugs, in which one drug inhibits the metabolism of another administered drug and causes an altered plasma concentration. CYP450 plays an important role for the metabolism of most drugs that are currently on the market. The risk of DDIs could pose serious impact on the efficacy and safety of co-administered medications needs to be carefully evaluated. Commonly applied target ineraction identifying methods include: LBVS, which uses pharmacophore maps and QSAR to identify a lead with 2D chemical structure information; and SBVS, which relies on the x-ray 3D structural information interpreted for the protein targets. Considering the flexible nature of the structure of CYP enzymes, the protein structure is not always exclusive. Therefore, we attempt a LBVS based approach to perform the prediction. Many machine learning algorithms have been constructed, outlining the future of artificial intelligence in assisting drug development. However, prediction limits have been associated with current practices, mainly due to critical structure- and conformation-related properties that conventional molecular descriptors fail to capture. Different 1D to 4D molecular descriptors and fingerprints have been utilized to identify different aspects of molecular properties and evaluate chemical similarities [66], [67]. Novel molecular fingerprints are needed especially intended for data mining and deep learning that could capture detailed protein-ligand binding information with specificity and enable accurate property prediction with strength.

A common goal for descriptor calculations is to extract critical characteristics of the molecule with mathematical calculations from the chemical properties and structures. However, it is an unknown myth as to what extent and how accurately information can be generated and utilized. Currently, there are more than 5000 molecular descriptors have been developed to featurize molecules [157], and rely on different dimension following various theorems of molecular structural information, including bond lengths, number of rotatable bonds, H-donors, H-acceptors, molecular volume and so on. Despite the abundance of the descriptors that have been developed, they are untrainable and lack of specificity. Furthermore, the descriptors are often correlated, making choosing the correct combination of descriptors an unsolved problem with no clear guidance. Usually, different descriptors were tested in parallel with several statistical methods for the best performance. Our laboratory has previously studied the intermolecular interactions of organic compounds through electronic structure-based local descriptors derived from conceptual density functional theory (CDFT) [158]. Local hardness and softness, facilitated by Hirshfeld analysis, have proven to quantitatively characterize the strength and locality of intermolecular interactions [158]. Significant correlations between the local electronic properties and intermolecular interactions can be explained by information of electronic properties residing on interacting contact motifs. Relatively large area have been proven to play significant roles. Local softness and hardness descriptors, derived from hard and soft (Lewis) acids and (Lewis) bases (HSAB) principle, can characterize both the locality and strength of intermolecular interactions [159].

Based on precious investigations, the innovative descriptors can address the gap by developing a novel 3D molecular fingerprints based on first principles of molecular interaction that includes molecular information with specificity and strength. Using quantum chemistry to interpret topological molecular information that resides on 3D molecular surfaces permits extraction of features directly from molecular 3D structures. This research can help improve deep learning algorithms to mine large amounts of protein-ligand binding experimental data and achieve better prediction performance.

In this thesis, I will present the plan to test our hypothesis on the predictability of using above-mentioned principles and from quantum chemistry perspectives with the following steps to address the above-mentioned challenge:

- **Data mining and curation.** Drug-like molecules with experimental measurements of CYP-binding activities will be filtered and curated from the literature and public databases.

- **Electronic calculation and manifold embedding.** Computer programs developed in house will be applied to generate low-dimension manifold of molecular surface that is mapped by electronic properties.

- **Development of Deep Learning (DL).** Computer programs developed will be applied to generate low-dimension manifold of molecular surface that is mapped by electronic properties.

We first mine a set of model ready data and generate molecular descriptors by extracting local properties directly from 3D molecule structure surface of ligands. Then, the features are projected into a low-dimension manifold to enable low dimensional calculation. From a library of small molecules or functional groups, we can first calculate the electronic structures and derive the local properties of molecules. 3D molecular surfaces can be generated with their 2D maps calculated by manifold learning and dimensionality reduction. Then, based on the projected 2D maps of the ligands, we can incorporate a deep neural network architecture that operates on set-structured data with attention mechanisms to recognize the surface features that the maps of the fragments can match. Thus, local interactions and fragments interactions can be characterized. Our pioneering method to assess the molecular interaction between a molecule and protein binding sites is based on first principles that define the origin of molecular interactions. Deep learning can evaluate the binding fitness between a ligand and target by machine learning by interpreting and extracting the low-dimension manifolds directly from the molecular surface.

With that, we expect our models to predict the ligand-target binding affinities and test our hypothesis that electronic surface properties can be applied to predict intermolecular activity. Understanding this interaction mechanism and predicting enzyme binding affinity can further advance the investigation and evaluation of DDI for drug development process.

3. DATA MINING AND CURATION

3.1 Introduction

Data collection is essential to evaluate the adequacy of modeling criteria for possible predictions. The outcome determination varies across different laboratories, assays and measuring techniques. Several limitations have been implicated with the lack of consistent protocol. It is critical to have experimental binding affinities and molecular properties of a set of both active and inactive compounds as reference to achieve a reliable prediction. High-throughput screening (HTS) assays can be used to test a large amount of chemical substances and are extensively applied in various areas including the pharmaceutical industry [160], [161]. PubChem is a public database developed and maintained by the U.S. National Institutes of Health (NIH) containing a large number of chemical properties. Large amount of data generated by HTS assays are submitted to and maintained by PubChem to facilitate the development of more-robust models [162].

PubChem database (AID: 1851), containing 17143 diverse compounds, was measured under a consistent experimental conditions with standardized protocol conducted by Dr. Auld and collegues [163]. A standard in vitro bioluminescent assay measuring the dealkylation of various pro-luciferin substrates to luciferin was applied against five major recombinant CYP isozymes (1A2, 2C9, 2C19, 2D6 and 3A4) at different concentrations [164]. The quantitative high-throughput screening (qHTS) assay was performed with bioluminescent-based detection method in 1536-well plates following an automated protocol. CYP inhibitory activities were determined measuring luciferin–luciferase bioluminescence reactions. When combined together, the luciferase acts as an enzyme, allowing the luciferin to release energy as it is oxidized to emit light. Luminescence signal intensity can be used to determine luciferin concentration. When inhibition activities are seen, production of luceferin will be limited and a reduced luminescence of luciferin would be observed. Inhibitory activities for CYP450 isozymes can be evaluated with respect to the intensity of luminescence. For this study, concentrations between 0.24 nM and 40 μM for each compound weere measured to calculate the concentration–response curves [165]. IC50s were calculated with concentration-response points fitted to Hill equation. Compounds were classified as Active, inconclusive, and inactive classes based on data quality and fitted experimental measurements.

3.2 Method

To extract a set of consistent data for model prediction, several filters were applied to the dataset. Figure 3.1 shows the data preparation pipeline. For our purposes, the key is to identify organic drug-like molecules with neutral charge and molecular size between 200 to 500Da. Metals and inorganic compounds were removed from the datasets since our prediction is dependent on surface electronic properties. Structural duplicated compounds were also removed during data pre-processing.

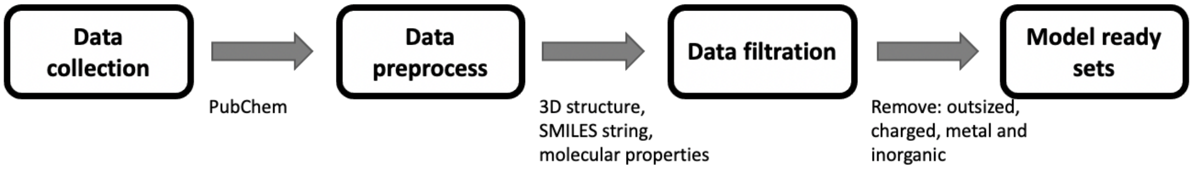

Data collection
PubChem
Data preprocess
3D structure, SMILES string, molecular properties
Data filtration
Remove: outsized, charged, metal and inorganic
Model ready sets

Figure 3.1. Data collection and preparation pipeline for DDI prediction

Both binary and continuous measurements were collected for each compound. ”Activity Outcome” and ”Activity Score” were collected for prediction and assessment based on the cutoff criterion provided by PubChem BioAssay library [162], [163]. AC50 is the compound concentration as 50% of the activity for an inhibition is reached. Compound activity outcome was determined as CYP inhibitor, non-inhibitor, or inconclusive with respect to there AC50 measurement. When AC50 is $\leq 10\mu M$, the compound is treated as CYP inhibitor. Compounds with AC50 >57 μM, are assigned as non-inhibitors. Compound with AC50 value between 10 to 57 μM are classified as inconclusive compounds. The value for AC50 below 10 μM was not reported. A normalized activity score with value between 0 and 100 were assigned to each compound with the most active inhibitors have a higher score and inactive compounds have a lower value. In addition to the training set, which contains 80% of available data entries, a diverse testing set and validation set, 10% each, were also gathered from the PubChem BioAssay database to verify the robustness of prediction models.

The same pre-processing procedure like described for training set was applied to testing and validation sets.

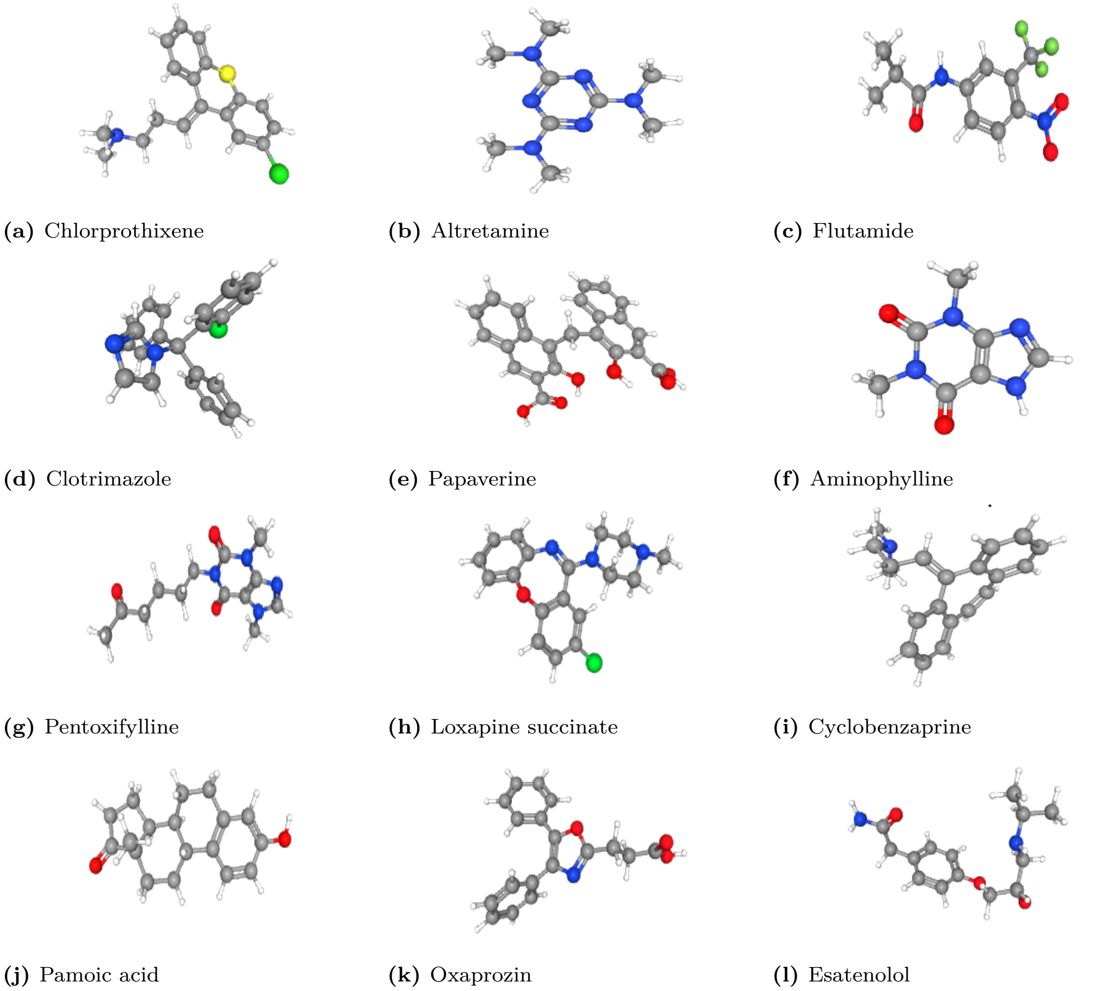
3.3 Results and Discussion

3.3.1 Data assessment

PubChem Dataset consisted 16697 unique compounds. 314 inorganic or charged compounds were removed to avoid confusions in model training with electronic properties. 1816 compounds that has molecular weight outside predefined range (200-500Da) were also removed to yield entire data set with good drug-likeness. Table 3.1 demonstrate an example of 10 compounds information extracted from PubChem library. pIC50 was converted from potency. Tanimoto index was calculated as 0.216, 0.209, 0.215, 0.212, and 0.218 for each of the target demonstrating good diversity in prepared data set. Final data sets to construct models for each enzymes are developed with specificity. Figure 3.2, 3.3, 3.4, 3.5, 3.6 are examples of respective 3D ligand structures for CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4, respectively, containing active, inconclusive, and inactive compounds.

3.3.2 Ligand information and 3D structures

Final data sets to construct models for each enzymes were developed with specificity. The entire set were divided into 5667 inhibitors, 3274 inconclusive compounds, and 5626 non-inhibitor compounds for CYP1A2; 5502 inhibitors, 3208 inconclusive coompounds, and 5857 noninhibitors for CYP2C19; 3689 inhibitors, 3849 inconclusive compounds and 7029 noninhibitors for CYP2C9; 2420 inhibitors, 2817 inconclusive compounds and 9330 non-inhibitors for CYP2D6; and 4668 inhibitors, 3676 inconclusive and 6223 noninhibitors for CYP3A4 in the pre-processed data set for each CYP450 end point as listed in Table 3.2.



(a) Chlorprothixene

(b) Altretamine

(c) Flutamide

(d) Clotrimazole

(e) Papaverine

(f) Aminophylline

(g) Pentoxifylline

(h) Loxapine succinate

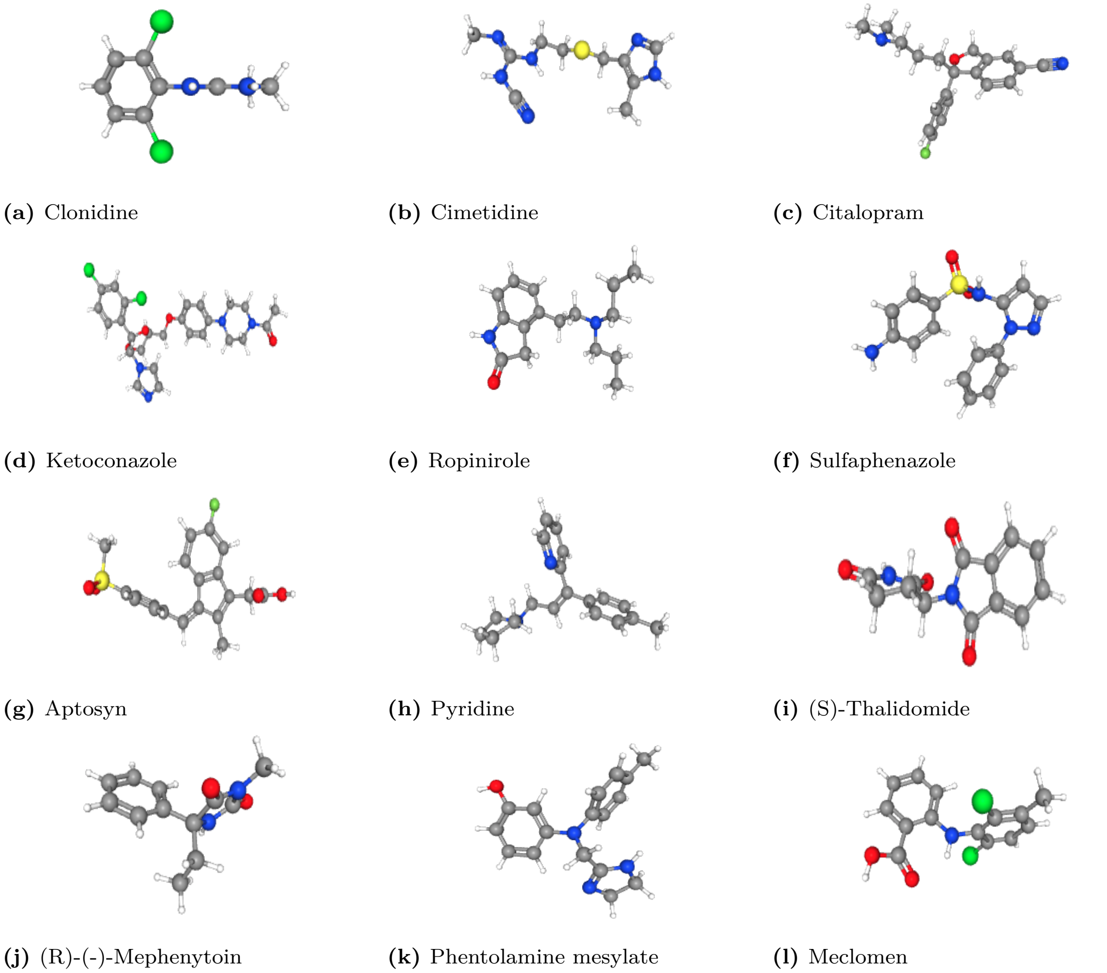
(i) Cyclobenzaprine

(j) Pamoic acid

(k) Oxaprozin

(l) Esatenolol

Figure 3.2. 12 of the representative ligands 3D structures extracted from PubChem Dataset against CYP1A2 with active ligands: (a) - (d); inconclusive ligands: (e) - (h); inactive ligands: (i-l) [166]



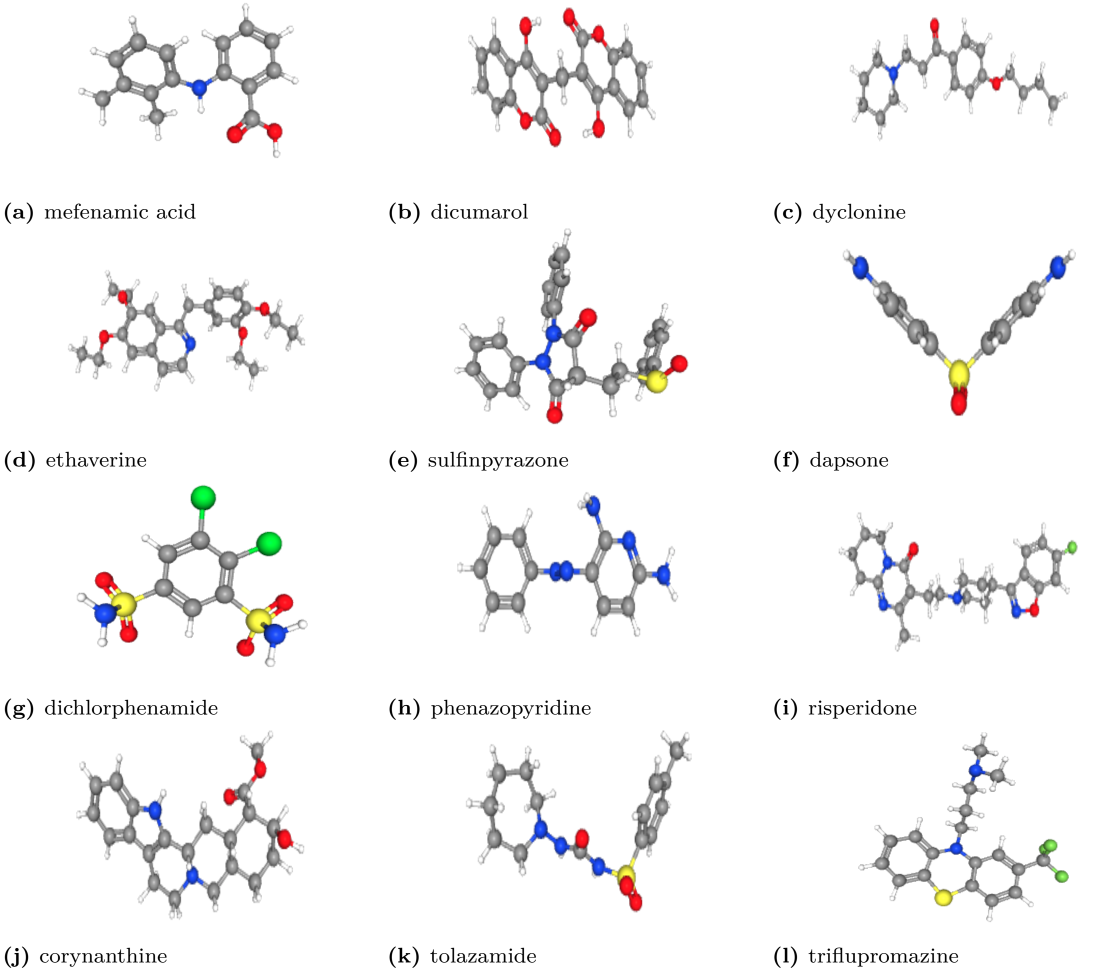
(a) Clonidine **(b)** Cimetidine **(c)** Citalopram

(d) Ketoconazole **(e)** Ropinirole **(f)** Sulfaphenazole

(g) Aptosyn **(h)** Pyridine **(i)** (S)-Thalidomide

(j) (R)-(-)-Mephenytoin **(k)** Phentolamine mesylate **(l)** Meclomen

Figure 3.3. 12 of the representative ligands 3D structures extracted from PubChem Dataset against CYP2C19 with active ligands: (a) - (d); inconclusive ligands: (e) - (h); inactive ligands: (i-l) [166]



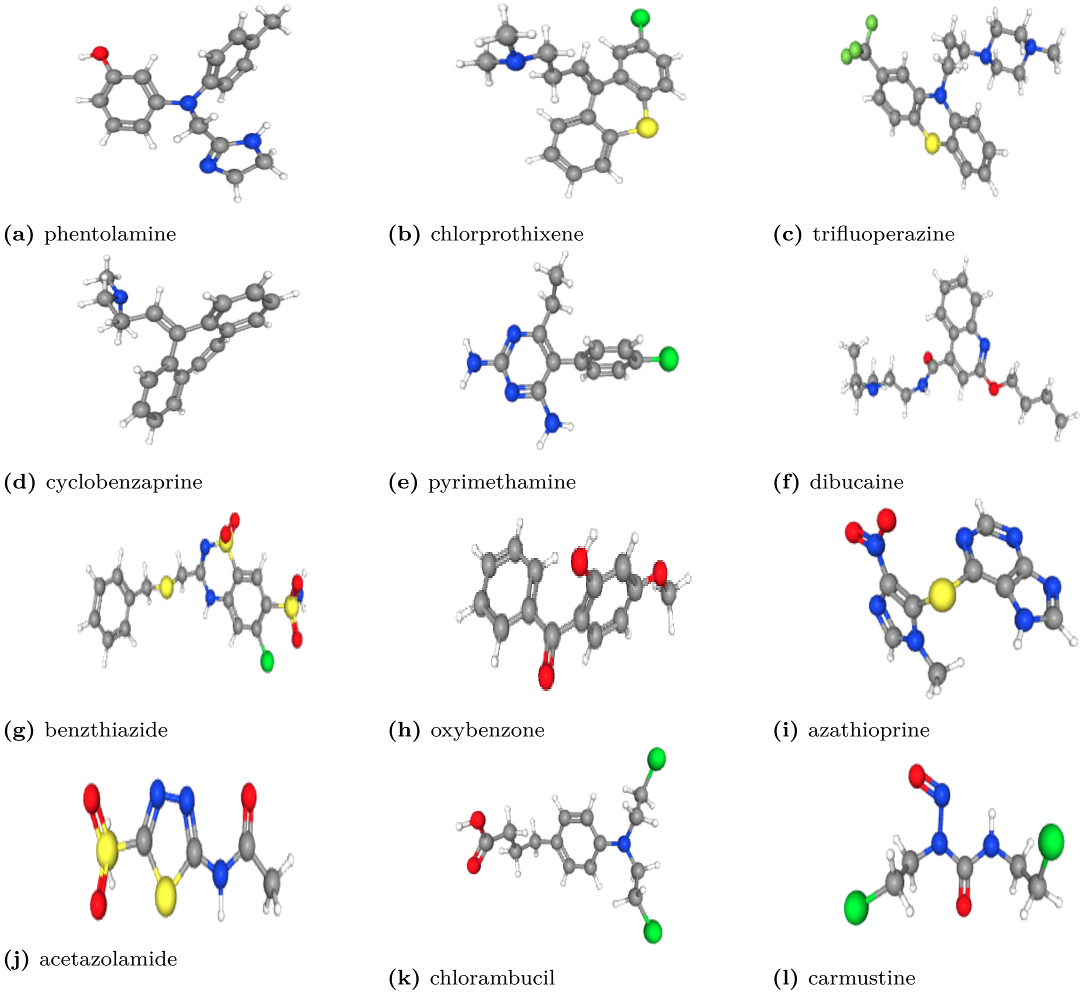
(a) mefenamic acid **(b)** dicumarol **(c)** dyclonine

(d) ethaverine **(e)** sulfinpyrazone **(f)** dapsone

(g) dichlorphenamide **(h)** phenazopyridine **(i)** risperidone

(j) corynanthine **(k)** tolazamide **(l)** triflupromazine

Figure 3.4. 12 of the representative ligands 3D structures extracted from PubChem Dataset against CYP2C9 with active ligands: (a) - (d); inconclusive ligands: (e) - (h); inactive ligands: (i-l) [166]



(a) phentolamine

(b) chlorprothixene

(c) trifluoperazine

(d) cyclobenzaprine

(e) pyrimethamine

(f) dibucaine

(g) benzthiazide

(h) oxybenzone

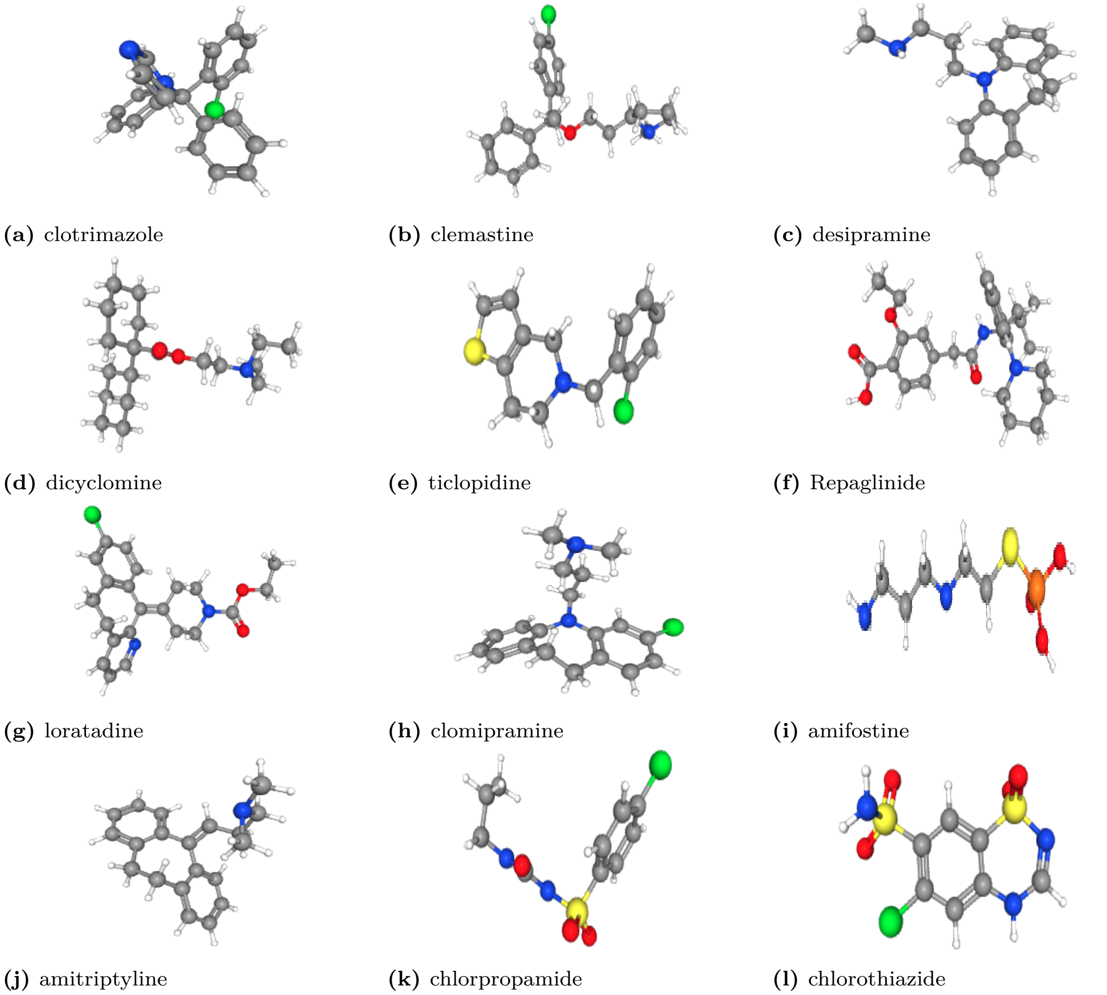
(i) azathioprine

(j) acetazolamide

(k) chlorambucil

(l) carmustine

Figure 3.5. 12 of the representative ligands 3D structures extracted from PubChem Dataset against CYP2D6 with active ligands: (a) - (d); inconclusive ligands: (e) - (h); inactive ligands: (i-l) [166]



(a) clotrimazole **(b)** clemastine **(c)** desipramine

(d) dicyclomine **(e)** ticlopidine **(f)** Repaglinide

(g) loratadine **(h)** clomipramine **(i)** amifostine

(j) amitriptyline **(k)** chlorpropamide **(l)** chlorothiazide

Figure 3.6. 12 of the representative ligands 3D structures extracted from PubChem Dataset against CYP3A4 with active ligands: (a) - (d); inconclusive ligands: (e) - (h); inactive ligands: (i-l) [166]

Table 3.1. Example of 10 PubChem data entries curatecd for CYP1A2 activity. (AS: Activity score; MW: molecular weight).

CID	AS	Outcome	Potency	pIC50	MW	Charge	SMILES
667466	45	Active	4.0	5.4	315.9	0	CN(C)CCC=C1C2=CC=CC=C2SC3=C1C=C(C=C3)Cl
2123	86	Active	2.0	5.7	210.28	0	CN(C)C1=NC(=NC(=N1)N(C)C)N(C)C
3397	90	Active	0.3	6.5	276.21	0	CC(C)C(=O)NC1=CC(=C(C=C1)[N+](=O)[O-])C(F)(F)F
2812	41	Active	15.8	4.8	344.8	0	C1=CC=C(C=C1)C(C2=CC=CC=C2)(C3=CC=CC=C3Cl)N4C=CN=C4
4680	20	Inconclusive	25.1	4.6	339.4	0	COC1=C(C=C(C=C1)CC2=NC=CC3=CC(=C(C=C32)OC)OC)OC
9433	20	Inconclusive	31.6	4.5	420.43	0	CN1C2=C(C(=O)N(C1=O)C)NC=N2.CN1C2=C(C(=O)N(C1=O)C)NC=N2.C(CN)N
4740	21	Inconclusive	12.6	4.9	278.31	0	CC(=O)CCCCN1C(=O)C2=C(N=CN2C)N(C1=O)C
71399	20	Inconclusive	31.6	4.5	445.9	0	CN1CCN(CC1)C2=NC3=CC=CC=C3OC4=C2C=C(C=C4)Cl.C(CC(=O)O)C(=O)O
2895	0	Inactive	NaN	NaN	275.4	0	CN(C)CCC=C1C2=CC=CC=C2C=CC3=CC=CC=C31
8546	0	Inactive	NaN	NaN	388.4	0	C1=CC=C2C(=C1)C=C(C(=C2CC3=C(C(=CC4=CC=CC=C43)C(=O)O)O)O)C(=O)O
4614	0	Inactive	NaN	NaN	293.3	0	C1=CC=C(C=C1)C2=C(OC(=N2)CCC(=O)O)C3=CC=CC=C3
175540	0	Inactive	NaN	NaN	266.34	0	CC(C)NC[C@@H](COC1=CC=C(C=C1)CC(=O)N)O

Table 3.2. PubChem Data Set: AID 1851 (conducted by the National Center for Biotechnology Information). Inactive: AC50 below 10 μM; Active: AC50 greater than 57 μM; Inconclusive: AC50 value smaller than 10 and greater than 57 μM.

	Inactive	Active	Inconclusive
CYP1A2	5626	5667	3274
CYP2C19	5857	5502	3208
CYP2C9	7029	3689	3849
CYP2D6	9330	2420	2817
CYP3A4	6223	4668	3676

3.4 Conclusion

A total of 14567 unique compounds were filtered from PubChem library. The dataset was curated to model ready set with Active, inconclusive, and inactive groups for each target. A comprehensive understanding of available data sources was established with Tanimoto distance calculation demonstrating data diversity thus will be used for further calculations and train deep learning models.

4. ELECTRONIC CALCULATION AND MANIFOLD EMBEDDING

4.1 Introduction

4.1.1 Electronic structure of chemical compounds

Density functional theory (DFT) played significant role in the evolution of quantum chemistry during the past two decades. Developed based on Hohenberg–Kohn theorem, which states that 'the ground state of any interacting many particle system with a given fixed inter-particle interaction is a unique functional of the electron density n(r)' [167]. DFT provides fundamental basis for computational study of molecules. Gradually, DFT has developed to two branches: 1. to obtain critical information about structure, energetics, and properties of molecule [168]–[170]; 2. Conceptual Density Functional Theory (CDFT) guiding for the computation of a variety of molecular properties and understanding the relationship with electron density [171]–[173]. Nowadays, both branches are now abundantly applied and well documented.

The cornerstone of our study is based on Pearson's HSAB (hard and soft acids and bases) Principle. First published in 1968, Pearson's HSAB (hard and soft acids and bases) principle is reported with the study of generalized Lewis acid—base reactions. It states that hard acids prefer binding to the hard bases, whereas the soft acids prefer interacting to soft bases [174], where Lewis acid is any molecule or ions that accept a pair of electron and Lewis base is any piece that is electron pair donor. Being the essence of our study, HSBA utilizes the inherent electronic characteristic of polarizability to define reacting electrophiles and nucleophiles as either hard or soft, with a thermochemically based operational scale. It can be further extended to characterize intermolecular interactions, especially in organic systems. When two molecules interact, local softness and hardness determine their spatial orientation and packing motif [158]. Soft regions or functional groups like to interact with other soft regions and hard regions prefer to interact with hard regions.

Conceptual Density Functional Theory (CDFT) provides a hierarchy of well-defined chemical concepts that provide fundamental interpretations and quantification measurements for the HSAB principle [173]. It enabled quantitative definition for molecular “hardness” and “softness” through the framework of electronic properties. Using electron density as underlying properties to describe atomic and molecular ground states along with its response to perturbation to analyze molecular chemical properties. Energy derived from the electronic structure and nucler-nuclear Coulomb repulsion energy can be used to determine total energy of a system.

Supporting the HSAB principle, Parr and Yang described the bodywork of calculating electronic softness and hardness. With two interaction molecular systems, A and B, their interactions energy ΔE is described as equation 4.1 [172]:

$$\begin{aligned}\Delta E = &\left\{\int \rho_A(r)\Delta v_A(r)dr + \int \rho_B(r)\Delta v_B(r)dr\right\} \\ &+ \left\{-\frac{[\mu_B^0 - \mu_A^0 + \int f_B(r)\Delta v_B(r)dr - \int f_A(r)\Delta v_A(r)dr]^2}{4(\eta_A + \eta_B)}\right\} \\ &+ \left\{\int\int [\frac{\delta \rho_A(r)}{\delta v_A(r')}]_{N_A}\Delta v_A(r)\Delta v_A(r')drdr' + \int\int [\frac{\delta \rho_B(r)}{\delta v_B(r')}]_{N_B}\Delta v_B(r)\Delta v_B(r')drdr'\right\}\end{aligned} \tag{4.1}$$

where $\rho_A(r)$ and $\rho_B(r)$ represent the electron densities of system A and B at point r, V_A and V_B are external potentials, f_A and f_B represent Fukui functions for each system which will be described later. The mathematical theory is composed of three terms. The first component enclosed in equation 4.1 represents the energy contribution from electrostatic interactions that is derived from electron densities. The second component represents electron flow contributed covalence-type interactions. The last component calculates the polarization energy contribution. Electrostatic interaction component becomes dominant component when A and B are hard systems, whereas the later two, covalence and polarization contributions, become significant with soft-soft interactions. Chemical hardness were derived from finite difference methods. Assuming quadratic relationship between number of electrons and electronic energy, chemical ”harness” ($\eta(\mathbf{r})$) at point r can be defined with the first vertical electron affinity (IE) and the first vertical ionization energy (EA) as equation: 4.2

$$\eta(\mathbf{r}) = IE - EA \tag{4.2}$$

Local hardness and softness are interrelated with the following series of equations 4.3 [175]:

$$\int \eta(\mathbf{r})s(r)dr = 1 \tag{4.3a}$$

$$\int s(r)dr = \int Sf(r)dr = S\int f(r)dr = S \tag{4.3b}$$

$$\int \eta(\mathbf{r})f(\mathbf{r})d\mathbf{r} = \eta \tag{4.3c}$$

Equation 4.3 (a) demonstrate the inverse relationship between local hardness and softness. Integrating local softness leads to global softness S, indicated in equation 4.3 (b), while the local hardness is applied to calculate η as shown in equation 4.3 (c) [176].

Flow of electrons and Van der Waals interactions can be attributed to electron clouds polarization effect, which generate hydrogen bonding and close contacts between molecules. For organic systems, those types of soft-soft intermolecular interactions are believed to be the dominate force and are directly associated with local polarizability, also attributed as ”softness” of a molecular system.

The Fukui function describes the electron density after adding or removing certain amount of electrons, which can be used to quantify local softness as equation 4.4 shows [177]:

$$f(r) = \left[\frac{\delta^2 E}{\delta v(r)\partial N}\right] = \left[\frac{\delta\mu}{\delta v(r)}\right]_N = \left(\frac{\partial\rho(r)}{\partial N}\right)_v \tag{4.4}$$

where the term E is the total electronic energy, N describes the count of electrons, v(r) represents external electrostatic potential at position r, and μ is the external potential. f^+(r) is nucleophilic Fukui functions, corresponding to the electron distribution on the lowest unoccupied molecular orbital (LUMO) since the frontier molecular orbital is involved during electron addition. f^-(r) represents electrophilic Fukui functions, which is also the electron distribution on the highest occupied molecular orbital (HOMO) that participates in the electron depletion.

$$f^{+}(r) = \rho(N) - \rho(N-1) = \rho(LUMO) \tag{4.5a}$$

$$f^{-}(r) = \rho(N+1) - \rho(N) = \rho(HOMO) \tag{4.5b}$$

The superscript ”+” denotes to an electron addition to the system and ”-” is the depletion of an electron from the system. $\rho(r)^{+}$, $\rho(r)$, and $\rho(r)^{-}$ represents electron densities of anion, neutral, and cation, respectively [159].

As a way to combinie the two Fukui functions, the difference between f^{+}(r) and f^{-}(r) is described as the dual descriptor $f^{2}(r)$:

$$f^{2}(r) = f^{+}(r) - f^{-}(r) \tag{4.6}$$

having a positive value where electrophilic happens and a negative contribution where it is nucleophilic.

While Fukui function can be used to quantify molecular local softness describing every point in chemical space, it is difficult to utilize to characterize intermolecular interactions or to probe a specific functional group describing interacting energies. The condensed Fukui function depicts similar local chemical reactivity with calculation done per atom with the following equations 4.7 [178]:

$$f_{k}^{-} = q_{k}(N) - q_{k}(N-1) \tag{4.7a}$$

$$f_{k}^{+} = q_{k}(N+1) - q_{k}(N) \tag{4.7b}$$

Normalization of the condensed Fukui function can be applied to generate summation of atomic Fukui values as shown in equation 4.8:

$$\sum_{k=1}^{Natoms} f_{k}(r) = 1 \tag{4.8}$$

Local softness, proportional to Fukui function, can then be calculated with equation 4.9:

$$s_k^- = Sf_k^- \tag{4.9a}$$

$$s_k^+ = Sf_k^+ \tag{4.9b}$$

for both electron addition and depletion.

The local hardness is ambiguous and many ways have been developed to define it as equation 4.10 shows:

$$\eta(r) = \frac{1}{N}\int \eta(r, r^{'})\omega(r^{'})dr^{'} \tag{4.10}$$

where $\eta(r, r^{'})$ is the hardness kernel denoted as the derivative of the chemical potential with respect to number of electrons. $\omega(r^{'})$ represents continuous charge density with respect to total charge $+1$. The ambiguity of hardness residing in the randomness with the choice of $\omega(r^{'})$. Definition would vary with different choices determined. Several approximations have been developed to describe local hardness. Politzer et al. developed a model in 1983 demonstrated electrostatic potential (ESP) can be used effectively to determine the hardness of an electronic system [176], [179]. ESP is a local function describing the position of the system and its positive and negative ions, which can be expressed as equation 4.11:

$$\Phi = \sum_{k=1} \frac{Z_A}{\|R_A - r\|} - \int \frac{\rho(r^{'})dr^{'}}{\|r^{'} - r\|} \tag{4.11}$$

where Φ is the sum of two parts due to nuclei and electrons, respectively. The first term on the right side of equation 4.11 represents a positive nuclei contribution, where Z_A describes the charge on the nucleus A at location R_A. The second term is a negative term representing electronic contribution to the molecular electrostatic potential with electron density $\rho(r')$.

The Fukui potential is believed to characterize the active site for electron transfer. An electrophile is willing to accept charge whereas nucleophile is willing to donate charge. It can be used to approximate the “distribution” of hardness in a molecule. With equation 4.12:

$$v_f^{+/-}(r) = \int \frac{f^{+/-}(r^{'})}{\|r' - r\|}dr^{'} \tag{4.12}$$

It provides an alternative perspective characterizing the electrostatic potential associated with the distribution of charge contributing to the relationship of local hardness. Fukui potential can be applied to explore intermolecular interactions where electron sharing is pivotal [180].

4.1.2 Intermolecular interactions in organic crystals - Benzonic Acid

Over the past few decades, various CDFT concepts have been developed and applied for understanding inter-molecular interactions for organic systems. A recent case study was performed in our lab to investigate quantitative relationships between CDFT concepts and intermolecular interactions in benzoic acid model crystal system [158]. DFT-based concepts, including Fukui functions and ESP, have been applied to quantitatively evaluated the intermolecular interactions, especially for organic molecules. To utilize this concept and quantify local hardness and softness of the molecular system, Hirshfeld surfaces defining the boundary between two molecules in crystal-based calcualtion, were implemented for system visualization and electronic properties quantification. In that study, eight pairs of intermolecular contacts were generated for the benzoic acid system and intermolecular interaction of each pair was calculated.

Figure 4.1 and Figure 4.2 [158] demonstrate ESP and Fukui functions mapped on Hirshfeld surfaces of eight identified packing motifs with intermolecular interaction energies calculated based on both the crystal and single-molecule based structure with Gaussian 09. Arranged in the descending order of intermolecular interactions, the first contacting pair has the strongest interaction largely contributed by hydrogen bonding. 76.19 and 93.72 kJ mol—1 were yielded at MP2 (second-order Møller–Plesset) and DFT-D (B2PLYP-D) levels, respectively. Contact 3 reflects a weaker hydrogen bond between =O and -CH groups. Contacts 2, 4, and 5 are − stackings involving interactions between –COOH and phenyl (2), phenyl and phenyl (4), and –COOH and –COOH (5), respectively. Contact 6 calculates the interaction between the para-hydrogen and phenyl group. Contacts 7 and 8 are the weakest interactions from Van der Waals forces.

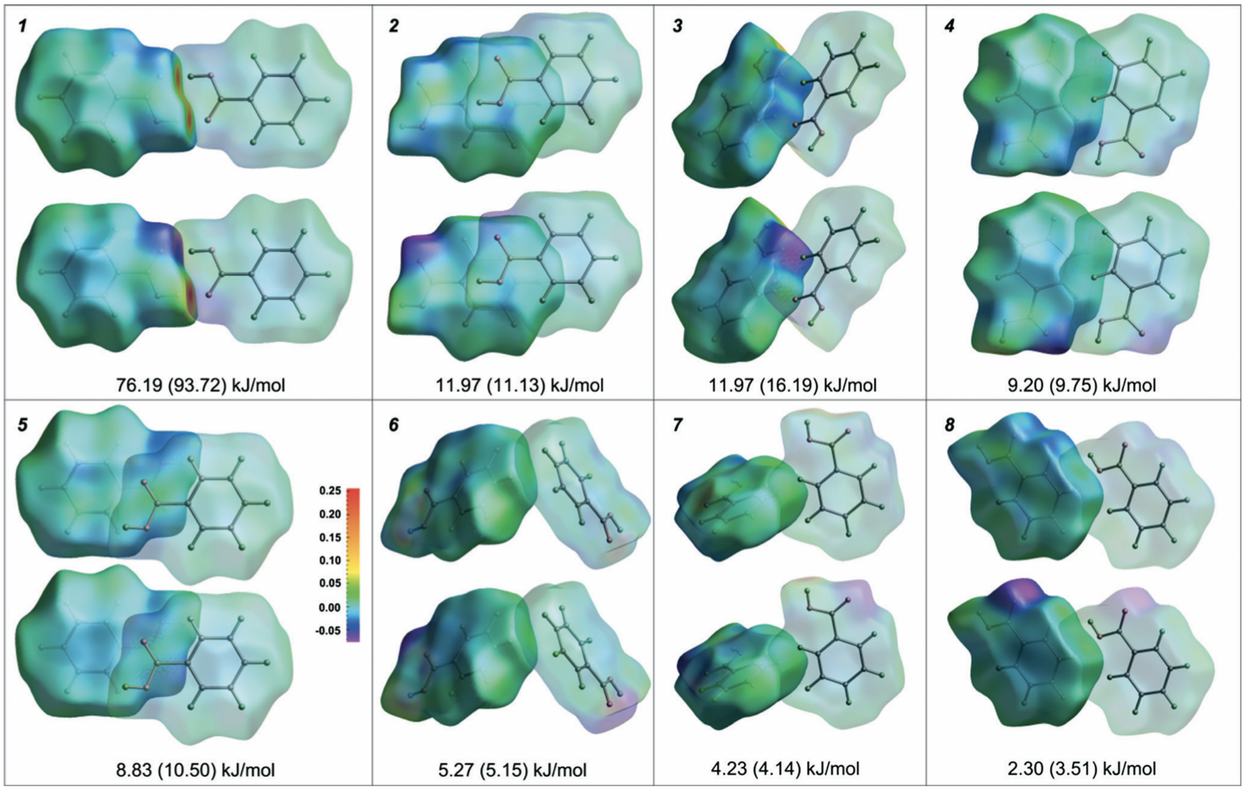


Figure 4.1. Figure is adopted directly from reference [158]. ESP mapped on hirshfeld surfaces of eight packing motifs mapped with the crystal- (top of each motif) and molecule-based calculation (bottom). Intermolecular interaction energies labeled with respect to each pair are calculated by Gaussian 09 at MP2 and DFT-D (in the parentheses) levels.

The Fukui functions and ESP mapped on Hirshfeld surfaces were integrated over the contacted area of each of the contact motif and plotted against the intermolecular interactions, as shown in Figure 4.3. Significant correlations were observed for molecular based calculation. Contact 1 show biased results with the interaction energies generated from hydrogen bonding which is several times larger than the rest of contacting motifs. Excluding contact 1, the integrated ESP demonstrated poor correlation with crystal based calculation have an R^2 of 0.25. Molecular based calculation yielded good correlation with an R^2 of 0.80.

As a local, spatial function, Fukui function indicate that the intermolecular packing motifs closely couple with local regions. Strong correlations with interactions is observed especially with f^- calculation. Quantitative trends is also observed with crystal-based calculation [158], [181]. However, the correlation is less significant when contact 1 interaction is excluded. Still,

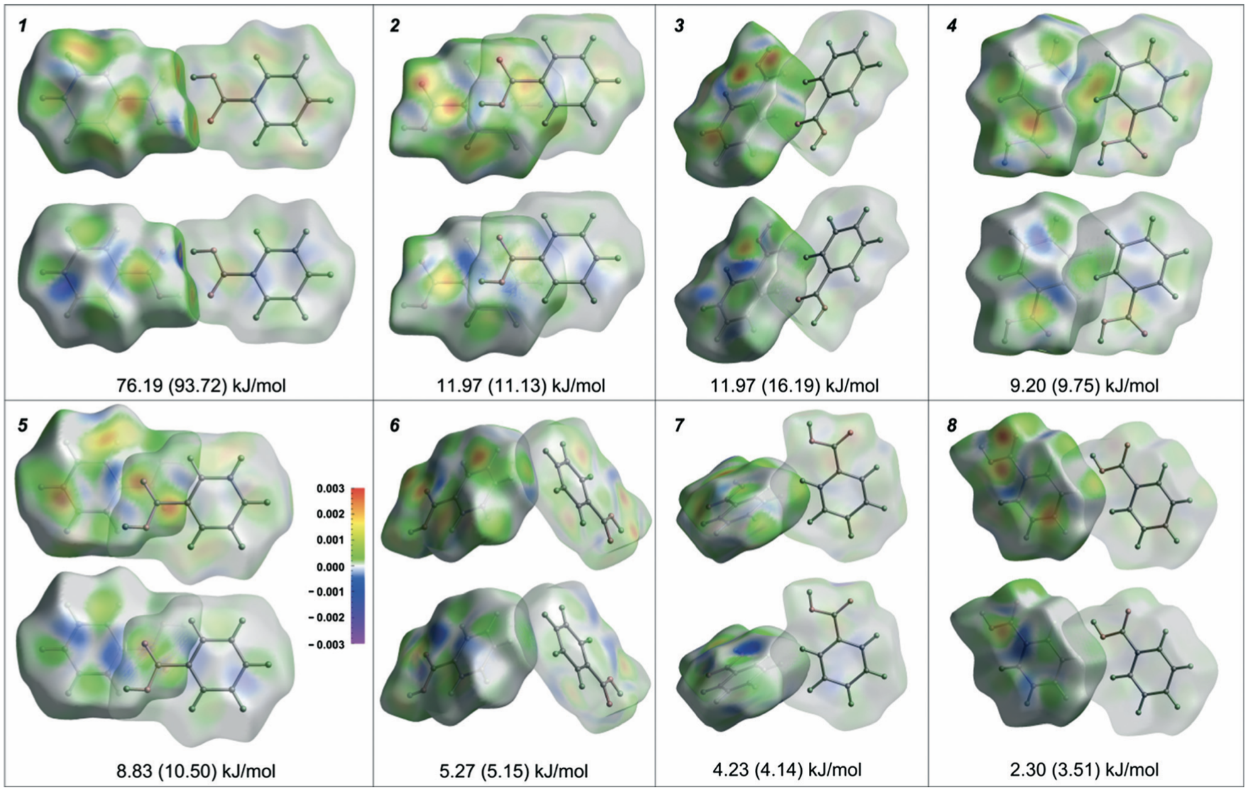


Figure 4.2. Figure is adopted directly from reference [158]. Fukui function mapped on hirshfeld surfaces of eight packing motifs mapped with the crystal- (top of each motif) and molecule-based calculation (bottom). Intermolecular interaction energies labeled with respect to each pair are calculated by Gaussian 09 at MP2 and DFT-D (in the parentheses) levels.

molecular based calculation demonstrated excellent linear relationship for both nucleophilic and electrophilic Fukui functions with $R^2 = 0.94$ and 0.74, respectively.

The study demonstrated that the intermolecular interactions in organic system are governed by electronic properties calculated on single molecule manner. The results indicate that intermolecular interaction energies is closely couple with local regions of relative strong Fukui functions. Hard-hard interaction, described by ESP, also demonstrate significant correlation when integrated over the contacting area of molecular surfaces. Matching of electronic properties decides the strength and energy of intermolecular interactions.

Besides the case study investigating ESP and Fukui functions with benzonic acid, fukui potential have also been found to containing information about intermolecular interactions

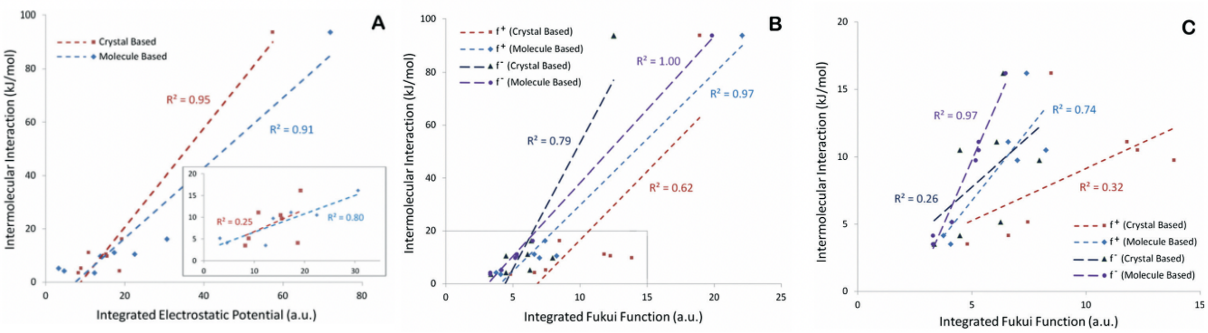


Figure 4.3. Figure adopted from reference [158]. Intermolecular interaction energy plotted as a function of integrated ESP (A) and the Fukui function (B) of contacting surface, with energy calculations for contact 2 - 8 shown in the inset of A (ESP) and C (Fukui Function) . Linear regressions of corresponding measurements are plotted with R2 denoted in figure.

and the electronic contribution to molecular electrostatic potential (MESP). Fukui potential is believed to be deterministic evaluating the regioselectivity of electron-transfer-controlled reactions [180]. A recent study explores the locality of intermolecular interactions with Fukui potential in a model systems, ROY, which bear intermolecular interactions encountered as $\pi - \pi$ stacking, close contact, and weak hydrogen-bonding interactions [159]. Evaluated at the single-molecule level, the results indicate that the Fukui potential plays an indispensable role evaluating interactions at contacting motifs where $\pi - \pi$ stacking is predominant. The sensitivity toward $\pi - \pi$ stacking carried by Fukui potential makes it worth exploring.

4.1.3 Molecular Surface

Molecular surfaces are used to represent three-dimensional geometric shapes of molecules. Various types of molecular surfaces including van der Valls (VWS), solvent-accessible surface (SAS) and molecular surface (MS) have been implemented with molecular studies [182], [183]. As demonstrated in Figure 4.4, VWS is an abstract representation for the molecule capturing the molecular surface residing on topological boundaries of Van der Waals radii of each individual atom. VWS, well delineating the interior of large molecules, describe the overall shape and can be used as a good tool for the abstract representations of molecules [184]. SAS, as shown in Figure 4.4B, is the surface area that is accessible to solvent. Defined by Lee Richards, SAS is the area traced out by a molecule the locii of the sphere center rolls over the VWS [185]. MS, also know as solvent excluded surface (SES), is a continuous surface tracing out by a probe sphere and rolls over VWS [186].

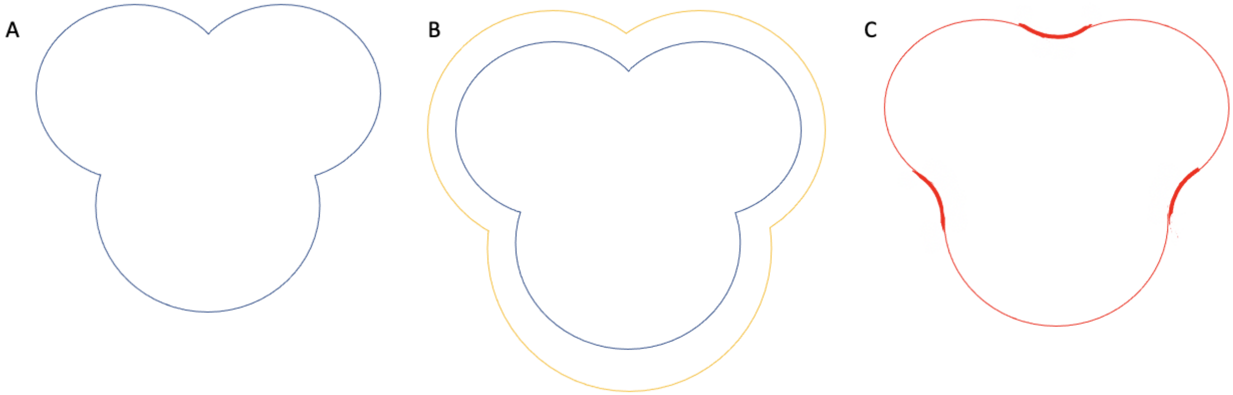


Figure 4.4. 2D examples of three molecular surface models. (A) van der Waals surface(blue); (B) solvent-accessible surface (yellow); (C) molecular surface (red).

4.1.4 Dimensionality Reduction

High dimensionality dataset contain comprehensive features well, however, become a challenge for data mining and machine learning due to the fewer direct relations the samples contain. The amount of training data needed for model training grow exponentially as the dimensionality of data increase which is known as the ”curse of dimensionality” [187]. Modeling chemical data in three dimension introduce excess degree of freedom tend to cause overfitting [187]. To foil the ”curse of dimensionality” and perform computationally tractable analyses, techniques to reduce the number of input dimensionality is needed. Dimensionality reduction method is considered as a critical data pre-processing step to eliminate irrelevant patterns from input dataset. It optimize computational power and algorithms needed for data mining and model learning.

To improve data quality and reduce data complexity, several dimensionality reduction techniques have been developed for variety data structures and successfully applied in drug discovery [188]. The goal is to transform data set represented in high dimension to low dimensional representations while maintaining the original features of the data as much as possible. During the process, data set $X = x_1, x_2, x_3, ..., x_n$ from a high-dimensional space is reduced into a low-dimensional space $Y = y_1, y_2, y_3, ..., y_n$. While the pattern from high dimensional space is eliminated, the low-dimensional representation needs to retain significant structures and crucial features of the original data for further calculations. Both linear and non-linear methods have been developed.

Principle Component Analysis

Principle Component Analysis (PCA) is the bedrock of dimensionality reduction techniques. Invented in 1901 by Karl Pearso [189], he described the goal of PCA as: ”In many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the ”best-fits straight line or plane”. PCA is an unsupervised, linear dimension reduction algorithm that reduce data dimensionality with linear calculation [190]. It provides a data-driven hierarchical coordinate system representing the statistical variations exist in the original dataset. The main idea of PCA is to analyze the high-dimensional dataset with observations described by inter-correlated quantitative dependent variables and considered maximum data variance as their new features [191]. The most representative variables are linear combinations of the original information, known as Principal Components (PCs), which are sorted to retain the maximum data variance of the original data. The first PC is supposed to have the largest variation presenting in all of the original variables, followed by subsequent PCs with decreasing variances. Each PC is uncorrelated to other PCs with the values of projected new variables, factor scores, representing the projections of PC observations. Mathematically, PCA attempts to find a set of orthogonal vectors that explain the variance of the eigenvalue of positive definite matrices.

Nowadays, PCA has been widely applied in various areas including physics, biology, chemistry, engineering and is commonly used in many machine learning applications. Similarly, it has been successfully applied in molecular modeling to facilitate analyzing and visualizing compounds dataset [192], [193]. A web-based public tool, ChemGPS-NPWeb, is introduced in 2008 for comprehensive chemical space navigation and exploration with linear projection of chemical data using PCA based global chemical positioning system [194]. Eight-dimensional (8D) maps were generated based on structure-derived physico-chemical characteristics through global mapping with the first four PCs of the map accounting for 77% of data variance. This PCA prediction system can be applied to interpret any compound with know chemical structure and has been successfully applied to various datasets including distinguishing anticancer modes; assisting with compound screening; and drug like property description of large datasets [195]. The advantage of applying PCA includes the non-iterative calculation it performs which greatly reduce time needed to perform the calculation. It can also prevent over-fitting and used as data compression tools [192]. Despite the promising applications, PCA comes with certain limitations. The algorithm is limited to linear operations when calculating eigen-matrices thus cannot produce optimal subspace for non-linear data. In addition, information losses are often seen when the PCs are not determined correctly [188]. In drug screening process, similar compounds do not always have similar biological interaction with target molecules. With highly complex and flexible protein structures, compounds with similar structural properties by chemistry definition sometimes bind with protein active site with different orientations, different conformation, or even a different protein. Furthermore, it is the neighborhood behavior that contain mainly interested features in the structure of known reference compounds [196]. PCA alone does not maintain local information from the input data and thus observe a weak correlation.

Often, non-linearity is observed with virtual screening modeling. In order to capture the indispensable nonlinear properties and assess there relations to extractble features, local based molecular descriptors that preserve essential manifolds for the molecules to restore global relationship can be invented [105]. Nonlinear projection techniques that operate without assuming global linearity can be good measures to transform nonlinear compound data facing the ”curse of dimensionality” [188]. Various nonlinear dimensionality reduction methods have been developed. Among those, self-organizing map (SOM), stochastic proximity embedding (SPE), and Stochastic Neighbor Embedding(SNE) are commonly used local manifold-based methods for understanding chemical compound structures that preserves neighborhood information [197].

Self-Organizing Map

Proposed by Kohonen in early 1980s, self-organizing map (SOM) is inspired by the cortical maps observed in human brain [198]. Like self-organizing process happen in human brain, SOM algorithm is composed of interconnected neurons in a grid structure. Neurons are represented with vectors arranged with topological structure and contain the same dimensional information as the input data. High-dimensional data can be mapped to a low dimensional manifold grid preserving their topological information. Commonly used grid structures includes hexagons, hyperbolic grids, and rectangular grids [188]. Local regions of the input space from the receptive targets are represented with a series of weights associated with each neuron. During the process, unsupervised calculation is performed to generate data clusters based on similarity between the neurons. The neighborhood in the grid is considered when each weight of the neuron is trained. Gasteiger and co-workers first applied a continuous SOM model based on topology theory to assist with compound design in 1997 [199]. The SOM algorithm has then been broadly applied to explore different questions in drug discovery [200]–[202].

When applied in molecular clustering, descriptor values are generated for testing molecules with SOM nodes assigned with randomly generated values. Molecules from testing sets is then mapped to the closest node having smallest distance to the corresponding vector. Node vectors, as well as their neighborhood connection nodes are updated to increase similarity to the test molecule. With the process, test molecules with similar descriptor vectors are grouped together with similar nodes. A larger number of groups of similar nodes are then generated with reduced group size. Eventually, molecules with similar descriptor vectors are assigned to similar node areas and a large dataset is clustered based on descriptor and local features. Molecular similarity search and compound prioritizing can be performed with common pharmacophore structures. This method effectively avoids comparison with candidate reference. Several studies have demonstrated the possible applications of SOM in LBVS [203]. Gasteiger and co-workers described a method using SOM as a novelty detection device [204]. Their model was successfully applied to identify and eliminate chemical structures perceived as novel compounds that lie outside the already discovered space which are unlikely to have desired biological activity from further investigations [204]. Bonachera and colleagues demonstrated the possibility of applying SOM to accelerate VS process of a large dataset by extracting a relatively small but diverse training set to increase efficiency in model learning [205].

Despite the promising performance of SOMs, one of the major limitations of the algorithm is the number of neurons and output dimension space needed to be determined before the training process [206]. Miss determination of those two parameters can weaken the predictability and be detrimental to the model performance. Training runs are sometime non-replicable and different training bears slightly different results due to the natural of stochastic optimization process. Usually HTS compound libraries contain hundreds of millions of data which can be too big for SOM to compute in a timely manner [188].

Stochastic Proximity Embedding

Stochastic Proximity Embedding (SPE) is a manifold based self-organizing dimensionality reduction algorithm. Since 1996, SPE method have been introduced and expanded to various application in computational chemistry and biology [207]. Aiming to preserve the metric structure and pairwise proximities, SPE generates low-dimensional Euclidean embedding for a set of input observations [208]. The calculation process randomly starts at an initial configuration and use a pairwise refinement process that repeatedly select pairs of objects to adjust their initial configuration. The goal is to match their proximiteis and minimize their distance on the map.

Various applications have been found to visualize large chemical libraries for biological activity explanation, diversity analysis, and analog design [209], [210]. The advantage of SPE algorithm is that it scales linearly with respect to the size of the dataset and circumvent that complete proximity matrix which allows efficient calculation especially for large data set [211]. Instead of revealing local information over global, SPE preserves global topology and local geometry by estimating the proximities between points as lower bounds of geodesic distances. At the same time, the method impose a global structure with the mean of geodesic distances. Besides that, SPE is intrinsically programmatically simple and robust that can be applied to explore a variety of scientific questions including exploratory datasets.

SPEs bear with some limitations. First of all, the calculation is related to a large number of adjustable parameters therefore the results heavily rely on neighborhood radius [188]. Local neighborhood can capture false negative information if the neighborhood radius is set too small and lead to discontinued and fragmented manifold clusters. When the neighborhood radius is set too large, it can include false positive data entries that belong to another manifold causing predicting errors.

Stochastic Neighborhood Embedding

Stochastic Neighbor Embedding (SNE) is another nonlinear, unsupervised, and manifold-based dimensionality reduction method [212]. With SNE, high dimensional data is mapped to low dimension space based on probabilities of points being neighbors to preserve significant structure of the original dataset. Data structure is preserved by their neighborhood probabilities introduced by pairwise distances. Gaussian model is applied to calculate the distance probability distribution. Probability distribution of point i be the neighbor of point j is calculated for both input and output space. SNE introduce a cost function with Kullback–Leibler divergence aims to estimate the probability distribution of neighbor as close as possible in the low-dimensional embedding and preserve the neighbor structure as the input embedding [213].

Since it's first development in 2002, different variations of SNE algorithm have been developed. T-distributed stochastic neighbor embedding (t-SNE) is introduced by Maaten et al. in 2008, which uses a symmetric cost function that is easier to optimize [214]. Instead of using Gaussian model, a student-t distribution is applied to effectively reduce the tendency to crowd points in the center of the map.

SNE has been applied with understanding various compound libraries to reduce dimensionality for data with high-dimensional pattern. For example, Drug Discovery Maps (DDM) is a machine learning model developed to map the a activity profile of compounds across an entire protein family based on t-SNE algorithm [215]. Visualization of molecular biological similarity was generated by t-SNE algorithm.

The representation of all dimensionality reduction method is not necessarily perfect which makes assessing and measuring the goodness of representation critical. Despite the fact that significant research have been conducted with methods for generating low dimensional manifolds, there has been lack of discussion with respect to definition of a good measurement of the representation. Preserving the pairwise distances is the current widely accepted principle that can be used as goals to assess distances between data points. Neighbor retrieval visualizer (NeRV) is another variation designed based of SNE algorithm. Applying pairwise distances, the context of each pair is considered to yield a natural way to evaluate dimensionality reduction performance [216]. Further, NeRV is a method specifically designed to project high dimensional data to two-dimensional space. It demonstrated the ability of generating continuous and differentiable functions of the output manifold. With the natural way developed to evaluate output performance, NeRV demonstrated more sophisticated performance with definitive evaluation thus was chosen in this method to perform dimensionality reduction and facilate with data visualization from 3 dimension to 2 dimension manifold. As a variation of SNE method, the algorithm uses a Gaussian distribution curve to preserve both local and global features of the original dataset's local and global features in the lower dimension and provide much easier optimization and better visualizations [212], [214]. It can capture the local structure of the high-dimensional data while preserving global structure.

Distribution of distances between points in high and low dimension space is interpreted with different matrices of properties via conditional probability $p_{\mathrm{j|i}}$ and $q_{\mathrm{j|i}}$, respectively. Assuming distances are Gaussian distributed in both high and low dimensional space, given the data point location y_{i} in our output lower-dimension space, $q_{\mathrm{j|i}}$ is denoted as the conditional probability of point of user's choice y_{j}. A probabilistic model of neighborhood retrieval can be constructed using the distance as our similarity metric in projection from the high-dimensional geodesic distances with 3D input space to 2D output space as in equation 4.13 [216]:

$$q_{\mathrm{j|i}} = \frac{exp(-\frac{\|y_{\mathrm{i}}-y_{\mathrm{j}}\|^2}{\sigma_{\mathrm{i}}^2})}{\sum_{k\neq \mathrm{i}} exp(-\frac{\|y_{\mathrm{i}}-y_k\|^2}{\sigma_{\mathrm{i}}^2})} \tag{4.13}$$

σ^2 is the variance of Gaussian distribution for y_i. The positive multiplier $1/\sigma_\text{i}^2$ allows the function for each cluster to be evaluated with respect to individual density. Nearby data points have relatively higher probability, whereas further data points end up with almost infinitesimal probability values. The similarity of data point in high dimensional space x_i to data point x_j is calculated as equation 4.14 [216]:

$$p_{\text{j}|\text{i}} = \frac{exp(-\frac{d(x_\text{i}-x_\text{j})^2}{\sigma_\text{i}^2})}{\sum_{k\neq \text{i}} exp(-\frac{d(x_\text{i}-x_k)^2}{\sigma_\text{i}^2})} \tag{4.14}$$

where $p_{\text{j}|\text{i}}$ is denoted as the conditional probability of point for the original data embedded in three dimension x_j with respected to x_i. d is the difference measured in the original data. k is a user-determined upper limit for number of relevant neighbors. Here, x_i is projected to y_i in output dimension.

To preserve the similarity between the high-dimensional data points and low-dimensional data points, the two probabilities $p_{\text{j}|\text{i}}$ and $q_{\text{j}|\text{i}}$ at every data point should be equivalent. The fundamental trade-off for dimensionality reduction lies between precision and recall of information retrieval as estimating true similarities calculation and avoiding false positive similarities at the same time. The novel NeVR method was developed by Venna and co-workers where the trade-off is set only maximize recall with Kullback-Leibler divergence as equation 4.15 shows [216]:

$$D(p_\text{i}, q_\text{i}) = \sum_\text{i}\sum_\text{j} p_{\text{j}|\text{i}} log(\frac{p_{\text{j}|\text{i}}}{q_{\text{j}|\text{i}}}) \tag{4.15}$$

For any point i, the Kullback-Leibler divergence $D(p_\text{i}, q_\text{i})$ represents a generalization of points recall, and $D(q_\text{i}, p_\text{i})$ is used for generalization of points precision. The aim is to minimize the sum of Kullback-Leibler divergence over all data points using gradient descent. A total cost function E_i defining both missing similar points and retrieving dissimilar points as shown in equation 4.18 can be applied to quantify the trade off [216]:

$$E_\text{i} = N_{FP,\text{i}}C_{FP} + N_{MISS,\text{i}}C_{MISS} \tag{4.16}$$

with false positives $N_{FP,\mathrm{i}}$ as the number of samples that are in both Q_i but not in P_i. $N_{MISS,\mathrm{i}}$ are samples that are in P_i but not in Q_i. C_{FP} defines the cost for each false positive and C_{MISS}

4.2 Method

4.2.1 3D molecular map generation

Based on the HSAB theoretical framework for examining inter-molecular interactions with local softness and hardness, several steps are taken to generate reliable 3D and 2D molecular maps to decipher the spatial arrangement of molecular surface electronic properties. Figure 4.5 demonstrate the workflow we applied.

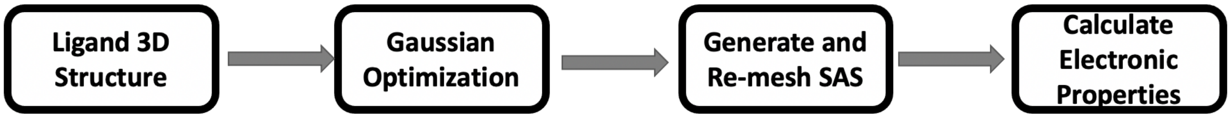


Figure 4.5. 3D molecule map generation pipeline.

To predict interactions between the molecules and target enzyme based on above mentioned first principles, 3D structures of each ligand obtained were optimized with Gaussian 16 (Gaussian Inc. Wallingford CT) applied at B3LYP/6-21G** level to calculate energies and electron densities of neutral, cationic, and anionic states. EDTSurf, which is a open source program to construct triangulated surfaces for molecules, were applied to construct triangulated solvent accessible surfaces [217]. EDTSurf was also applied to identify cavities which are inside of the collected molecules. The 3D structures were then processed with MeshLab (ISTI - CNR) to re-generate the mesh and parse out the vertices. An explicit re-meshing of a triangular mesh to improve aspect ratio and topological regularity were applied with 8 iteration and target length 1.25. Fukui functions, Fukui potentials, and ESPs were calculated at the single-molecule level and mapped to SESs of all molecules obtained from the data library for visual analyses and feature extraction.

4.2.2 Manifold Embedding of Molecular Surface

Manifold Embedding of Molecular Surface (MEMS) calculation is applied with NeRV algorithm to reduce 3D molecular surfaces of ligands to 2D space as described previously. The distance between any two atoms on a molecular surface is used as the neighborhood similarity metric for computing the probabilities of retrieval and relevance in 2D manifold. The value of the electronic property at each data point on the surface were carried directly to the corresponding data point on the 2D map. Same 2D feature maps were generated from molecular surfaces and used to map different local electronic properties (Fukui functions, Fukui potential and ESP). Pairwise probability was applied to normalize the weights:

$$\frac{\partial C}{\partial y_{\mathrm{i}}} = 2\sum_{\mathrm{j}}(k_{\mathrm{ij}} + k_{\mathrm{ji}})(y_{\mathrm{i}} - y_{\mathrm{j}}) \tag{4.17}$$

In the output space, geodesic distances was applied to calculate tarchitecturehe distance between point y_{i} and point y_{j}.

Recall and precision are applied to evaluate the quality of reduction. Expectation and the means of recall $D(p_{\mathrm{i}}, q_{\mathrm{i}})$ and precision $D(q_{\mathrm{i}}, p_{\mathrm{i}})$ were calculated as ${}_{\mathrm{i}}(p_{\mathrm{i}}, q_{\mathrm{i}})$ and ${}_{\mathrm{i}}(q_{\mathrm{i}}, p_{\mathrm{i}})$ in analogous to mean precision and recall cannot in general reach their minimum simultaneously. With cost function mathematically described as equation 4.18

$$\begin{aligned} E &= \lambda \mathbb{E}_{\mathrm{i}}[D(p_{\mathrm{i}}, q_{\mathrm{i}})] + (1-\lambda)\mathbb{E}_{\mathrm{i}}[D(q_{\mathrm{i}}, p_{\mathrm{i}})] \\ &\propto \lambda \sum_{\mathrm{i}} \sum_{\mathrm{j} \neq \mathrm{i}} p_{\mathrm{j|i}} log \frac{p_{\mathrm{j|i}}}{q_{\mathrm{j|i}}} + (1-\lambda) \sum_{\mathrm{i}} \sum_{\mathrm{j} \neq \mathrm{i}} q_{\mathrm{j|i}} log \frac{q_{\mathrm{j|i}}}{p_{\mathrm{j|i}}} \end{aligned} \tag{4.18}$$

λ was used to justify the desired trade off.

Both method minimizing precision and recall were applied. Closed map calculation is obtained to minimize the amount of misses by squashing the molecule flat to low dimensional space. This method resulted in a great amount of false positives caused by misplacing electronic properties residing on opposite sides of the molecular surface close to each other. An alternative embedding is referred to open cut calculation which simply cut open the 3D molecular structure and fold it out. The open cut calculation would eliminate false positives but introducing misses on different sides to the tear as they ended up to be far way from each other. For open-cut calculations, we randomly generated 4 cuts for each ligand in an attempt to weaken the effect of miss calculated regions.

4.3 Results and Discussion

4.3.1 3D molecule surface electronic calculation

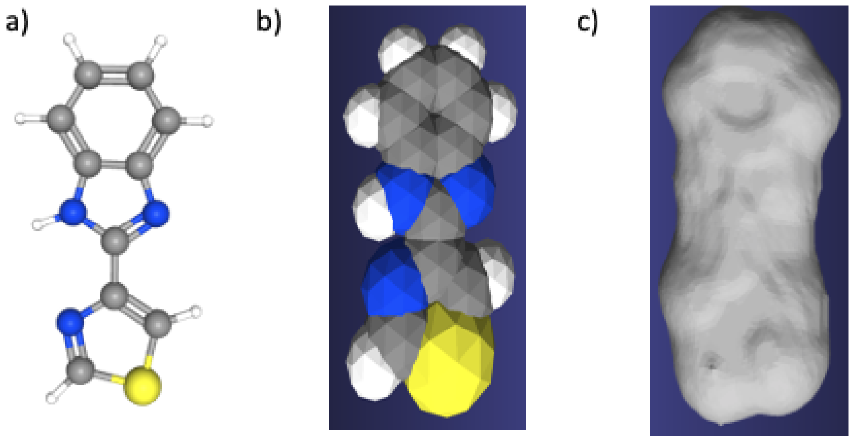


Figure 4.6. 3D molecule surface mesh generated for thiabendazole. Subfigure (a) is the ligand 3D structure obtained from PubChem (C ID: 5430); (b) is the 3D structure generated from Gaussian optimization; (c) is the Surface accessible area calculated with EDTsurf.

To demonstrate the results of 3D molecular map structure calculation for the data extracted as described in Chapter 3, calculation and analysis for one of the CYP1A2 inhibitor, thiabendazole, is demonstrated in Figure 4.6 as an example. Figure 4.6 (a) is the ligand 3D structure extracted from PubChem. Figure 4.6 (b) demonstrate the result of 3D structure optimization. Surface accessible area calculation was then performed and re-meshed as Figure 4.6 (c).

A series of 3D molecular maps featuring molecular surface electronic characteristics describing the ”softness” and ”hardness” of the ligands were generated. The results for thiabendazole is demonstrated in Figure 4.7. Local electronic properties, including Fukui potentials (Figure 4.7a), electrostatic potential (ESP) (Figure 4.7b), nucleo- and electrophilic Fukui functions (f^{+}r (Figure 4.7c) and f^{-}r (Figure 4.7d), derived from the single ligand were mapped with respect to 3D surface structure. Those properties characterize both locality and strength of intermolecular interactions. Red Region shown in ESP indicates negatively

Table 4.1. Surface electronic feature values calculated residing on 3D surface.

	Electronic value Range [high-low]
ESP	[0.055 - 0.001]
Fukui Function	[0.001 - (-0.001)]
Fukui Potential	[0.037 - (-0.04)]

charged region which brings down the potential energy to interact with a proton. Therefore the electronic contribution of ESP is always negative. Fukui potential is the positive contribution by electron to ESP which is always positive due to the interaction with proton. The subtraction order is opposite for Fukui function (Figure 4.7e) as shown in equation 4.6. Therefore the red region means a positive value indicating to more contribution to the negativity of the region with reducing potential. Table 4.1 contains surface electronic values for ESP, Fukui Function, and Fukui potential with thiabendazole calculation. As expected, comparing with soft interactions (Fukui function), hard properties (ESP) has dominate effect when present.

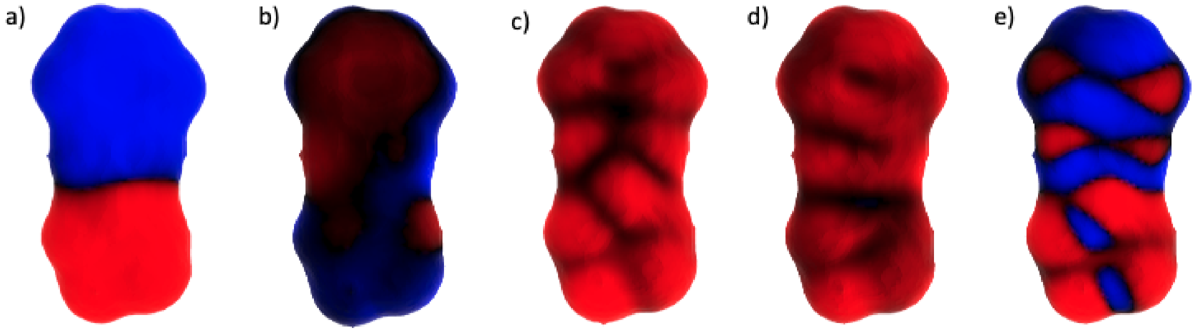


Figure 4.7. 3D surface electronic maps calculated for thiabendazole: (a) 3D Fukui potential (ESP^+ - ESP^-), red: negative; blue: positive; (b) 3D ESP, red: negative; blue: positive; (c) Nucleopholoc Fukui Function, f^+ (marking electrophilic region); (d) Electrophilic Fukui Function, f^- (marking nucleophilic region; (e) Fukui Function f^2, blue: negative; red: positive.

4.3.2 MEMS calculation

Ligand MEMS calculations were performed with NeVR non-linear dimensionality reduction method as mentioned previously. Both closed maps and open cuts were generated. Figure 4.8 demonstrates close map embeddings of electronic properties residing on 3D molecule surface of thiabendazole. The embeddings were applied to minimize the number of misses by squashing the molecule flat from 3D space to 2D area. It yielded few misses with the introduction of false positives. Figure 4.9 demonstrates false positives observed in thiabendazole ESP calculation. The area labeled with red rectangle contains noise neutral data points displaying as white dots. This observation is due to the misplacing of information from the opposite side of the molecule 3D structure. Still, major electronic values were well captured locally. The noise can be smoothed with further calculations.

Figure 4.11 shows 4 random open cut calculations generated to minimize the number of false positives by cutting 3D molecules open. The cutting performed on 3D molecule surface is demonstrated in Figure 4.10. When applied to model training, four cuts generated would be used together to minimize the effect at the tears.

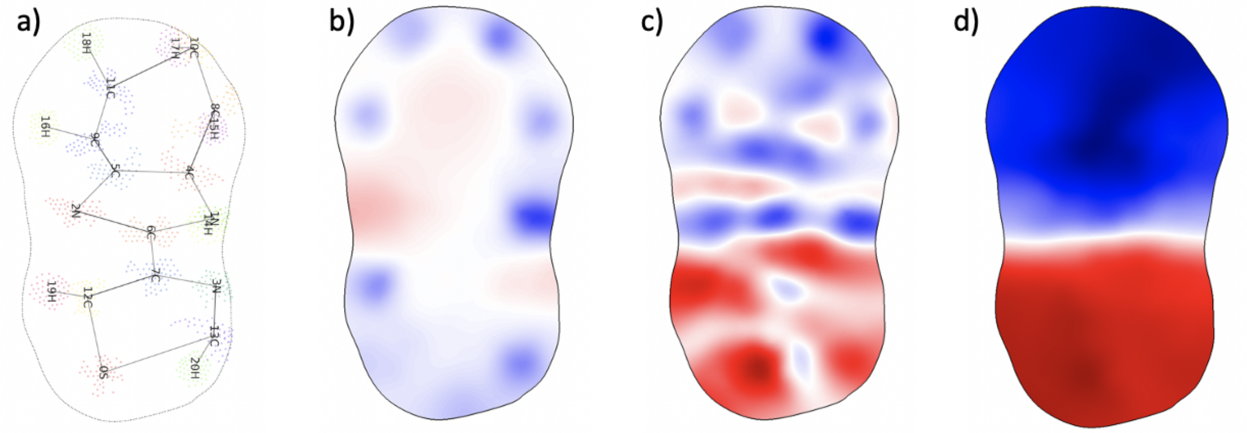


Figure 4.8. Closed map calculation for 2D electronic maps embedded from 3D surface of thiabendazole: (a) shows calculated 2D contour of the molecule maps including surface points closest to atom coordinates; (b)-(d) are 2D molecule maps for ESP, Fukui function, and Fukui potential, respectively.

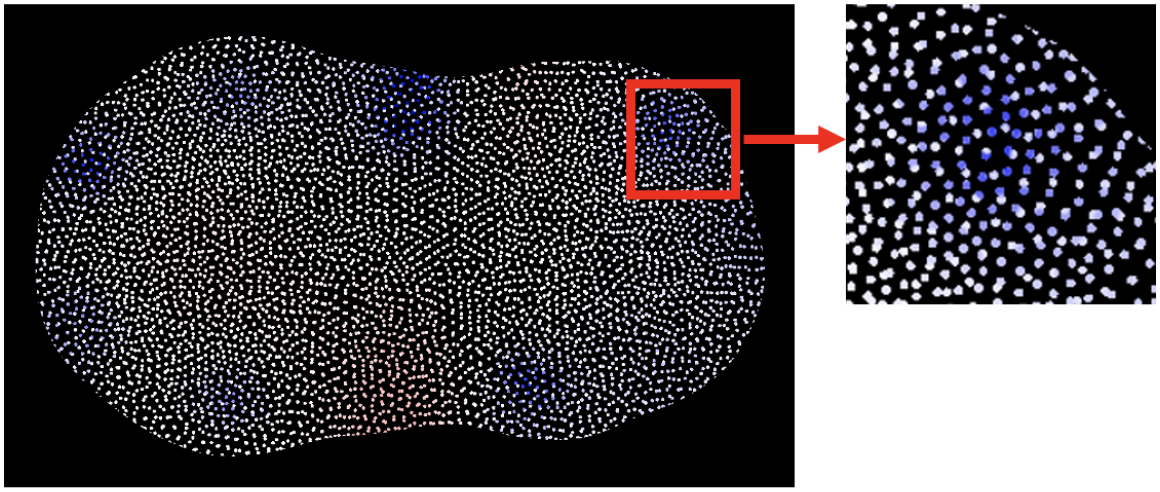


Figure 4.9. ESP results from MEMS close map calculation OF thiabendazole introduce significant false positives.

4.4 Conclusion

With the current MEMS workflow, we are able to generate molecular surfaces with respective to the isolated ligands and calculate 2D maps that is not affected by the orientation of molecules. Fukui functions, Fukui potentials, and ESPs are calculated as local functions bearing electronic characteristic distributions. 3D molecular maps with those properties creates a wide degrees of freedom that can hurdle further calculations. To generate molecular surfaces that map the electronic properties and the molecular surface conformational flexibility in feature learning, electronic properties calculated with respect to 3D surface needs to be mapped in 2D dimension to enable further mathematical evaluation and avoid the "curse of dimensionality". Manifold learning method, NeRV, optimizing the cost function to better balance the information precision and recall from high dimensional space to 2D manifold is applied. Ideally, we would like to avoid misses and, at the same time, eliminate false positives as much as possible. To better evaluate the ability of retaining significant features, two manifolds were calculated to minimize either false positives or misses with closed map calculation and open cuts method, respectively. The current approaches are able to preserve the majority of local features and provide possibilities to characterize intermolecular

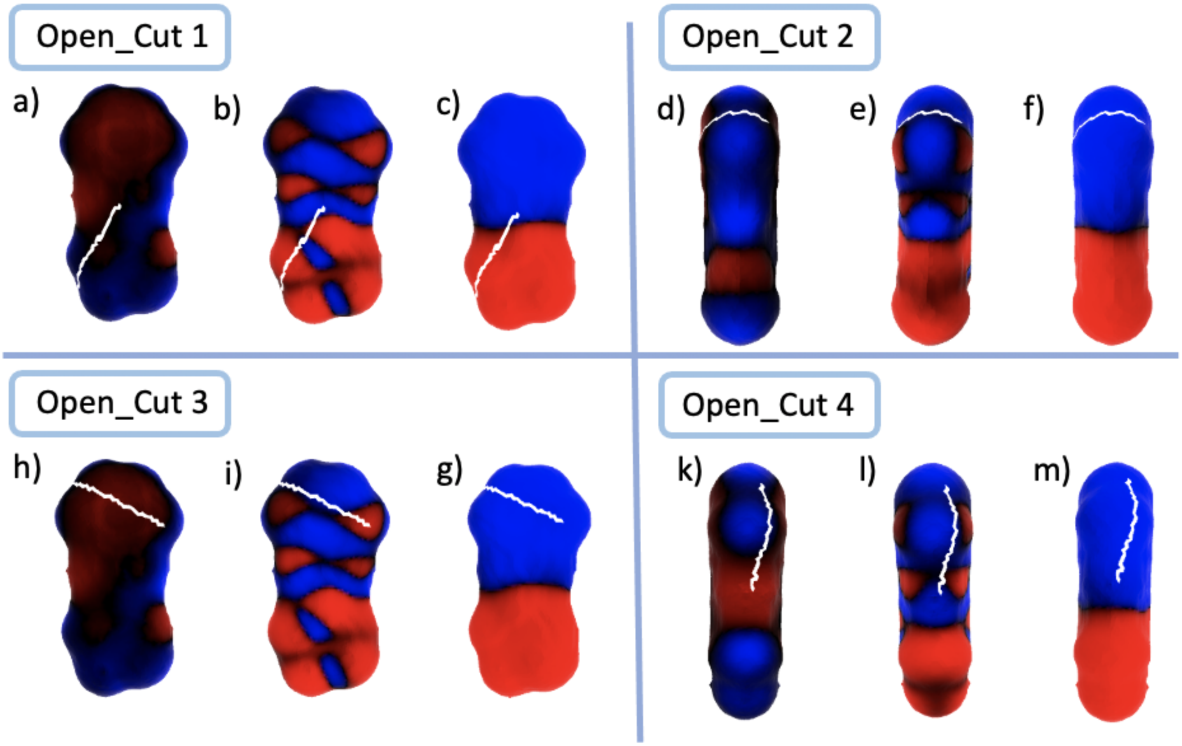


Figure 4.10. Four cuts generated randomly to perform dimensionality reduction with open cut calculation on 3D surface of thiabendazole: (a) - (c) show cut 1 on 3D molecule surfaces for ESP, Fukui function, and Fukui potential, respectively; (d) - (f) show cut 2 on 3D molecule surfaces for ESP, Fukui function, and Fukui potential, respectively; (h) - (g) show cut 3 on 3D molecule surfaces for ESP, Fukui function, and Fukui potential, respectively; (k) - (m) show cut 4 on 3D molecule surfaces for ESP, Fukui function, and Fukui potential, respectively.

interactions. Quantitative definition for molecular "hardness" and "softness" through the framework of electronic softness and hardness are captured in 2D maps. The assumption is atomic and molecular property can be described with electron density and perturbation responses. With the geodesic distance between two data points on a molecular surface applied as neighborhood similarity computing the retrieval probabilities, electronic property values for each data point embedded on the surface are carried directly to corresponding positions on the 2D maps. The generated 2D feature maps can be used to capture different local electronic properties which are further characterized with machine learning and computer

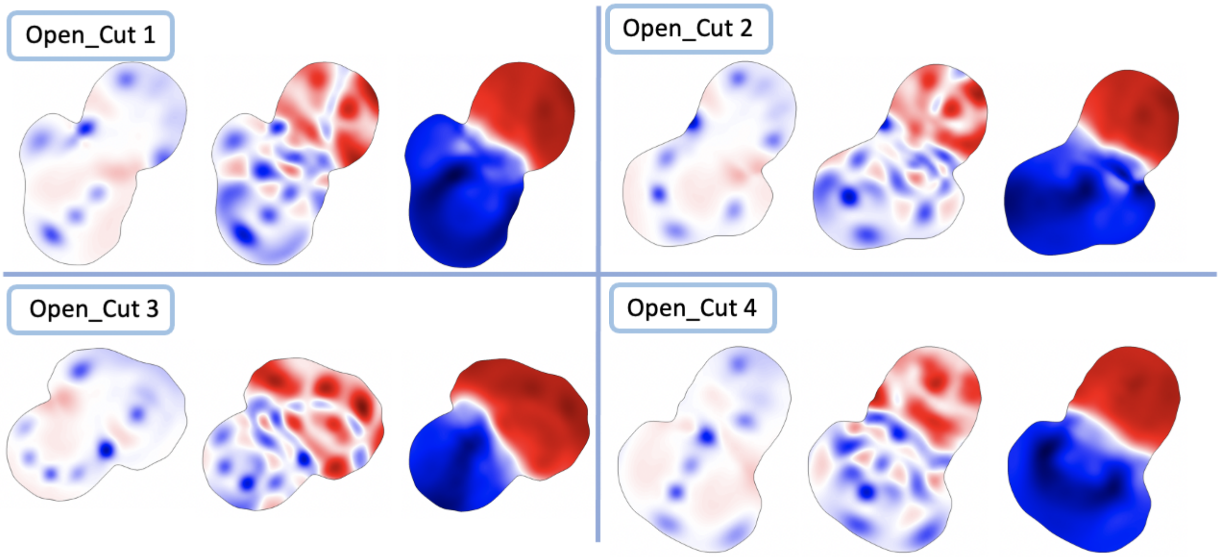


Figure 4.11. Open map calculation for 2D electronic maps embedded from 3D surface of thiabendazole: (a) - (c) show 2D molecule maps for ESP, Fukui function, and Fukui potential, respectively, calculated from open cut 1; (d) - (f) show 2D molecule maps for ESP, Fukui function, and Fukui potential, respectively, calculated from open cut 2; (h) - (g) show 2D molecule maps for ESP, Fukui function, and Fukui potential, respectively, calculated from open cut 3; (k) - (m) show 2D molecule maps for ESP, Fukui function, and Fukui potential, respectively, calculated from open cut 4.

vision algorithms. The strength and locality of intermolecular interactions can be learnt and predicted with such calculation.

5. DEEP LEARNING PREDICTION

5.1 Introduction

Our previous studies with respect to using CDFT concepts to understand intermolecular interactions have demonstrated the indication of the local electronic properties with molecular interactions. The interaction locality and strength can be assessed with local electronic properties casting on molecular surface, providing possibilities to predict inherent propensities of a molecule to interact with another without putting them together. Here we would like to apply the same underlying mechanism to characterize the interactions between a set of ligands and the binding sites of a protein target. The work presenting in this chapter aim to perform image interpretation and model training for lignads activity prediction with previously calculated 2D electronic property maps.

The projected 2D maps from the molecule surfaces for each ligand need to be properly interpreted with mathematical algorithms. In the age of machine learning, many algorithms can be used to evaluate similarities between shapes and to perform object recognition as well as feature extraction. Common molecular descriptors like mentioned in Chapter 1 are untrainable. There is also no protocol or standard process to choose the most reliable combination of descriptors. To develop a orientation insensitive method that can perform ligand evaluation requiring no docking between the ligand and binding site, we adopted shape context approach to generate translational invariant molecular descriptors for each ligand with mathematical representations. Like human can recognize many objects based on their shapes, this method provides a good practice with translation and rotation invariance to detect image similarity with the 2D feature maps generated from MEMS calculations. It captures the distribution over relative positions with a set of points corresponding with other shape points. The points count summarize to describe global shape information with a rich, local descriptor [218].

5.1.1 Shape Context

Shape context method was first developed by Belongie and coworkers in 2000 [219]. The main goal of shape context is to solve the correspondence between two shape items. The method is designed to describing shapes in a way allowing measurement of shape similarity and recover to point correspondences [218].

Briefly, the first step of the calculation starts with capturing the shape of an object by a discrete subset of n points P $=\{p_1, p_2, p_3, ..., p_n\}$, $p_\text{i} \in \mathbb{R}^2$ sampled from the internal and external contour of the shape. For each point, p_i, the relative coordinates can be obtained from all other n - 1 points. A histogram h_i for each points can be generated with respect to the Euclidean distance and angular bins against neighbor points defined as the shape context of responsive point p_i [219]:

$$h_\text{i}(k) = \{q \neq p_\text{i} : (q - p_\text{i} \in bin(k))\} \tag{5.1}$$

The descriptor is generated with polar coordinates to be more sensitive with respect to differences in relatively nearby pixels. The graph is normalized with log space map to count the number of points in each region. Two shapes can then be compared based on the information retained in histograms. More specifically, as shown in Figure 5.1, a set of points P $=\{p_1, p_2, p_3, ..., p_n\}$, $p_\text{i} \in \mathbb{R}^2$ can be obtained from edge elements of Figure 5.1(a) and another set of points Q $= \{q_1, q_2, q_3, ..., q_n\}$, $q_\text{i} \in \mathbb{R}^2$ can be obtained from edge elements of Figure 5.1(b). The Euclidean distance (r) and angular bins (θ) from each point in the set that can be compared against with the other n - 1 points. For each point, the number of points lie in each bin can then be represented as shown in Figure 5.1(d) - (e), where corresponding points tends to have similar descriptors. The algorithm enables global understanding of the shape which make sure the descriptor is insensitive to translation and rotation. Robustness to affine transformations and rotation is proven in [220].

The algorithm capture robust features for assessing shape similarity and can be used to determine the common information encountered between a set of images. For each molecular surface map, each ligand can be represented with a set of discrete points sampled from the

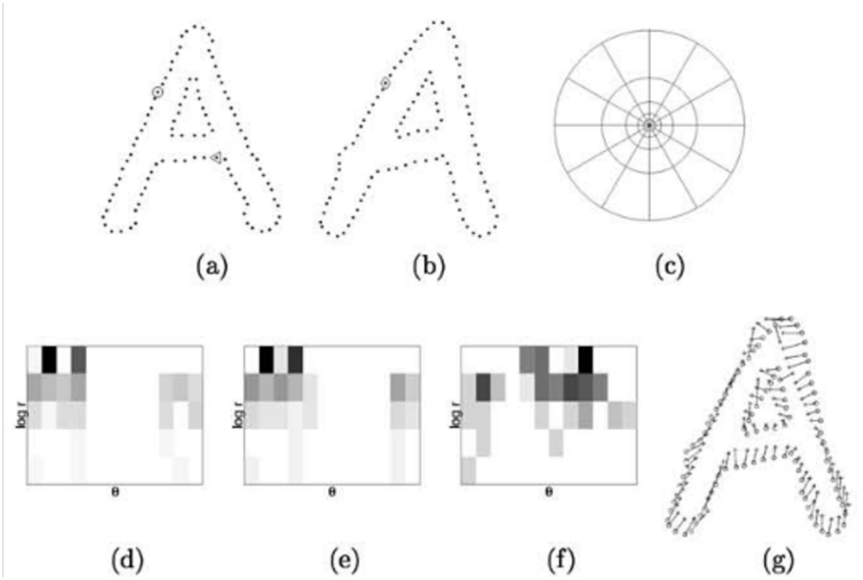


Figure 5.1. This figure is adopted from reference [219]. Shape context computation is demonstrated. (a) and (b) show two sampled edge points of different shapes. (c) is the set up of log-polar histogram bins applied in computing the descriptor. (d) - (f) are examples of corresponding points $\circ, \Diamond, \triangleleft$, mapped in (a) and (b). Gray-scale plot is used with darker color representing larger value. (g) is correspondences found using bipartite matching, with costs defined by chi-squared distance between histogram.

captured external contours. Descriptor bins is calculated for each atom, enable information retrieval and machine learning.

5.1.2 Neural Network

With translational invariance molecular descriptor calculated, a deep learning algorithm is applied to figure out the correlation between molecular surface features and inhibition activities. The fundamental output feature should not alter if the coordinates are translated or permutated when molecule conformation remain unchanged. A deep neural network architecture, DeepSets, developed by Zaheer and co-workers that can tolerate permutation is adopted to operate on the collection of atom set based data to perform feature extraction and and machine learning [221]. DeepSets is a permutation equivariance method developed to deal with sets inputs X $=\{x_1, x_2, x_3, ..., x_M\}$. For any permutation π [221],

$$f([x_{\pi(1)}, ..., x_{\pi(M)}]) = [f_{\pi(1)}(x), ..., f_{\pi(M)}(x)] \tag{5.2}$$

function f(x) acting on the sets are permutation equvariant to the order of objects in the set. For a set of X, the function is computed as equation 5.3:

$$f_{\Theta}(X) = \sigma(\Theta X) \tag{5.3}$$

When assessing each atom as set input, not all information have equal function and effect with respect to other atoms in a molecule. The context information can be important. Attention mechanism is applied to evoke the concept of context information. Inspired by the similar idea with how human brains and visual system dealing with massive inter-correlated sensory inputs, attention mechanism implemented in neural network stimulates direct attention to objects of interests while ignoring other signals. Focusing on task-relavant contexts of a neural network, it allows the model to capture the most relevant information from the inputs and yield a better estimation [222], [223]. Attributing the highest weight values to the most relevant vectors, attention mechanism allows the decoder to treat input sequences using the most relevant information.

A general attention mechanism contains three major components: the queries, the keys, and the values as shown in Figure 5.2. Trainable matrices are applied to each of the factors enabling attention learning. When assessing one query, a set of attention weights will be applied to assess how much attention should be paying to each key. A set of values is applied to perform bias selection and generate final output. The attention process can be calculated with three steps: alignment score, attention weights, and context vector [224]. The alignment score calculation takes the encoder hidden states and previous decoder states to compute the score that maps how well the inputs align with current output around positions. With ”s” be the scores generated for the state ”h”, the alignment model is represented with equation 5.4 implemented by a feedforward neural network:

$$\mathrm{e}_{\mathrm{i,j}} = a(s_{\mathrm{i}-1}, h_{\mathrm{j}}) \tag{5.4}$$

with i and j representing sequence input and output, respectively. $\mathrm{e}_{\mathrm{i,j}}$ represents node j in i neighborhood. s_{i_1} is derived from the query vector and encoder input h_{j} from the key vector. The attention weights α_{ij} are then normalized by a softmax function given by the following equation 5.5:

$$\alpha_{\mathrm{ij}} = \frac{\mathrm{e}xp(\mathrm{e}_{\mathrm{ij}})}{\sum_{k=1}^{T_x} \mathrm{e}xp(\mathrm{e}_{\mathrm{i}k})} \tag{5.5}$$

The context vector z for the output is then generated using the weighted sum of the annotations:

$$\mathbf{z} = \sum_{\mathrm{j}=1}^{T_x} \alpha_{\mathrm{ij}} \mathbf{h}_{\mathrm{j}} \tag{5.6}$$

with keys and values assigned as the same vector.

Given the input layer generated with shape context method and experimental binding affinities extracted as mentioned in Chapter3, deterministic features of ligand enzyme binding are identified with respect to each atom in the ligand. The relationships of other atoms are incorporated in the calculation. Ligand similarities are assessed based on quantum properties and drug enzyme interactions are constructed based on the information about molecule ”hardness” and ”softness”. Higher-dimensional information is determined and embedding interpretations can be generated for each ligand.

5.2 Method

Within the context of our project, relative positional information of atoms coordinates are captured with 2D molecular maps calculated for Fukui functions, Fukui potentials, and ESP. Descriptors with respect to each atom within the ligand is calculated with molecular contours. Molecular atoms were calculated with a coarse grain generator developed with KDTree to remove redundant information. First described in Maneewongvatana and Mount

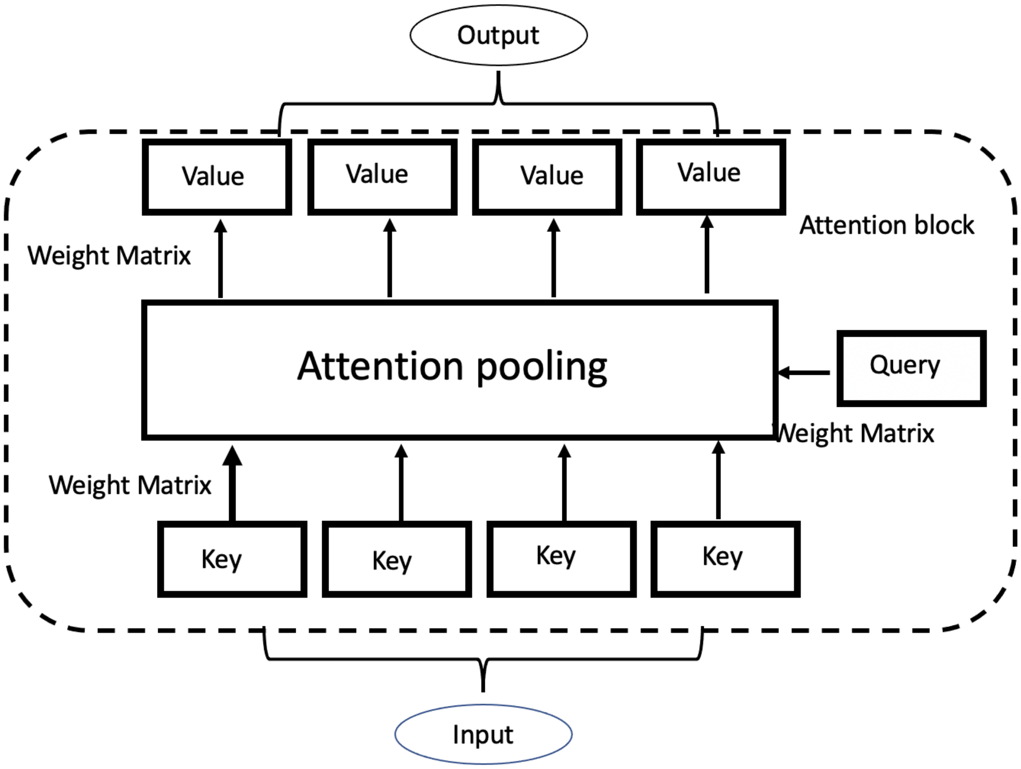


Figure 5.2. Attention mechanisms with queries, keys, and values

.

1999, KDTree algorithm was applied to look up the nearest neighbors of each atoms with an index provided with the set of 2-dimensional points containing atom coordinates [225]. Each node specifies an atom and splits the set of atoms based on whether their coordinate along that axis is greater than the threshold set as 0.25. Coarse grained molecule maps were generated with necessary atom information.

Different combinations of angular and bin numbers were calculated as 1 bins for log r with 32 bins for θ and 4 bins for log r with 16 bins for θ, respectively. Shape context feature calculation is performed per atom basis. More specifically, these resulting "matrices" are used to calculate the electronic properties of a key point in the molecule, in particular how the electronic properties and binding of these atoms are affected by the surrounding atoms. This is done by initially centering the matrix over an atom within the molecule. At each of the 32 or 16 bins, the surface electronic properties are calculated that describe what the electronic density of the key atom is with respect to binding. Interactions with the other features of the molecule is housed in that particular bin. The value of a particular bin is calculated as the summation of all values that each bin covers. Normalization was applied by bin area. This process is done for each of the 32 or 16 bins. Once this process is completed for a single atom, it is repeated by re-centering the matrix around the next atom in the molecule. This process is repeated until calculations for each of the 32 or 16 bins are completed for each atom in the molecule. Figure 5.3shows an example of the calculation design.

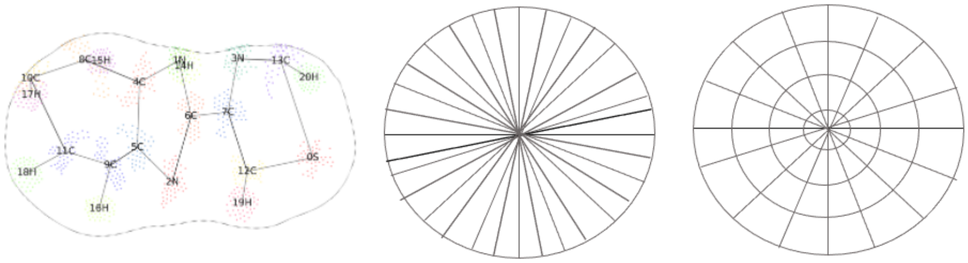


Figure 5.3. Molecular descriptors calculated with Shape Context method with a CYP1A2 inhibitor, thiabendazole. (a) Molecule 2D maps derived from 3D structure containing all atoms. (b) - (c) Diagrams of log-polar histogram bins applied in computing the shape contexts with 1 bins for log r with 32 bins for θ, and 4 bins for log r with 16 bins for θ.

Our permutation equivariance deep neural network model were constructed using DeepSets. Figure 5.4 demonstrate the architecture of our models. We implemented self-attention mechanism with the element-wise non-linearity tanh as activation function to include both negative and positive values for proper gradient flow in DeepSets layers as equation 5.7:

$$Attention(X) = softmax(\frac{\tanh f_1(X) \cdot f_2(X)^T}{\sqrt{d_small}}) \cdot X \tag{5.7}$$

where f_1 and f_2 are two single layer MLPs with width $d_small = \max[\frac{d_x}{32}, 4]$. The attention layer is applied to automatically extract important high-level deterministic properties and capture relationships among ligands. It also leverage both the features of drugs and the interaction behavior with enzymes. Softmax function is applied to perform normalization across different nodes and Lekey Relu nonlinearity is used to allow slope for the negative part:

$$a_{\mathrm{i,j}} = LeakeyReLU(softmax_{\mathrm{i}}(\mathrm{e_{i,j}}))) = LeakeyReLU(\frac{exp(\mathrm{e_{ij}})}{\sum_{k \in \mathbf{N}_{\mathrm{i}}} exp(\mathrm{e_{i}}_k)}) \tag{5.8}$$

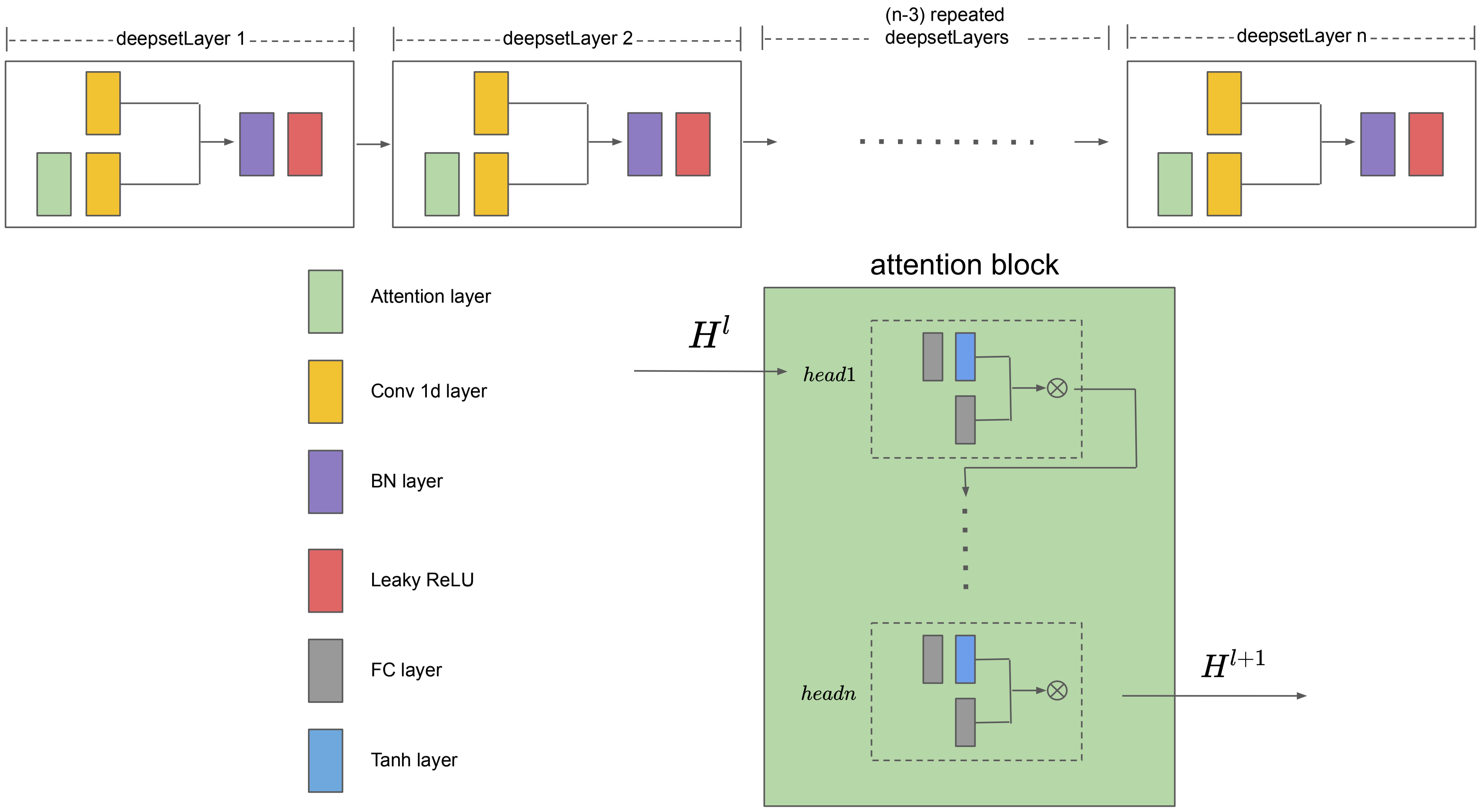


Figure 5.4. Architecture of Deepsets neural network with attention mechanism.

The points (atoms) are summed over to remove permutation variance. Activity score is used as quantitative predicting label.

5.3 Results and Discussion

5.3.1 Feature representation

Figure 5.5 shows an example of 2D molecular mapping with atom coordinates and coarse grain calculation for thiabendazole.

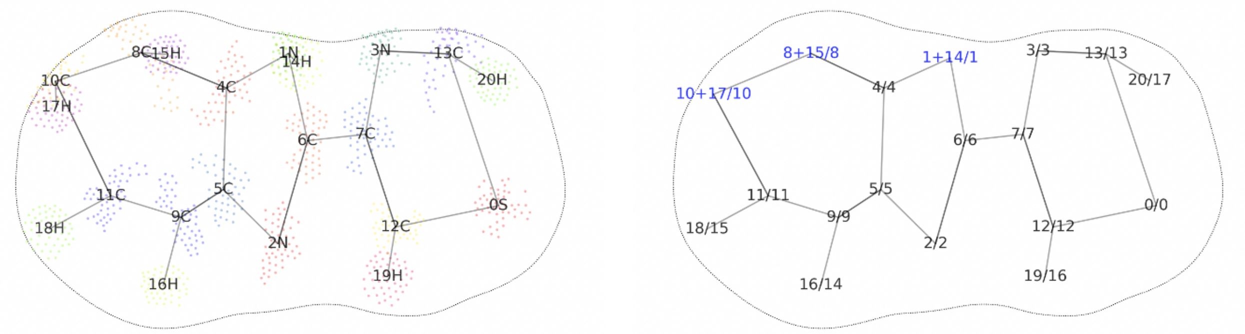


Figure 5.5. 2D molecular maps calculated with atom coordinates for all atoms (a) and coarse grained atom map (b) for thiabendazole.

Different log-polar histograms of the electronic area measured using each atom as the origin is generated as Figure 5.6 and Figure 5.7 for 1 bin for log r with 32 bins for θ and 4 bins for log r and 16 bins for θ.

The results for shape context computed with 4 bins for log r with 16 bins for θ appear to carrying more global information than the result from 1 bins for log r with 32 bins for θ, emphasizing local features and context. The strength of matching and training the image features for ligand suggest matching the hotspots and formation of intermolecular interactions as demonstrated in our earlier benzoic acid study. Because the 2D maps are generated by the dimensionality reduction method from 3D molecular surface, the same 2D images can be used to map different electronic properties calculated for the same molecule. This allow us to study distinct matching among the local polarizability with electrophilic and nucleophilic Fukui functions, study local hardness with the positive and negative ESP, and evaluate the distribution of ”hardness” with Fukui potential.

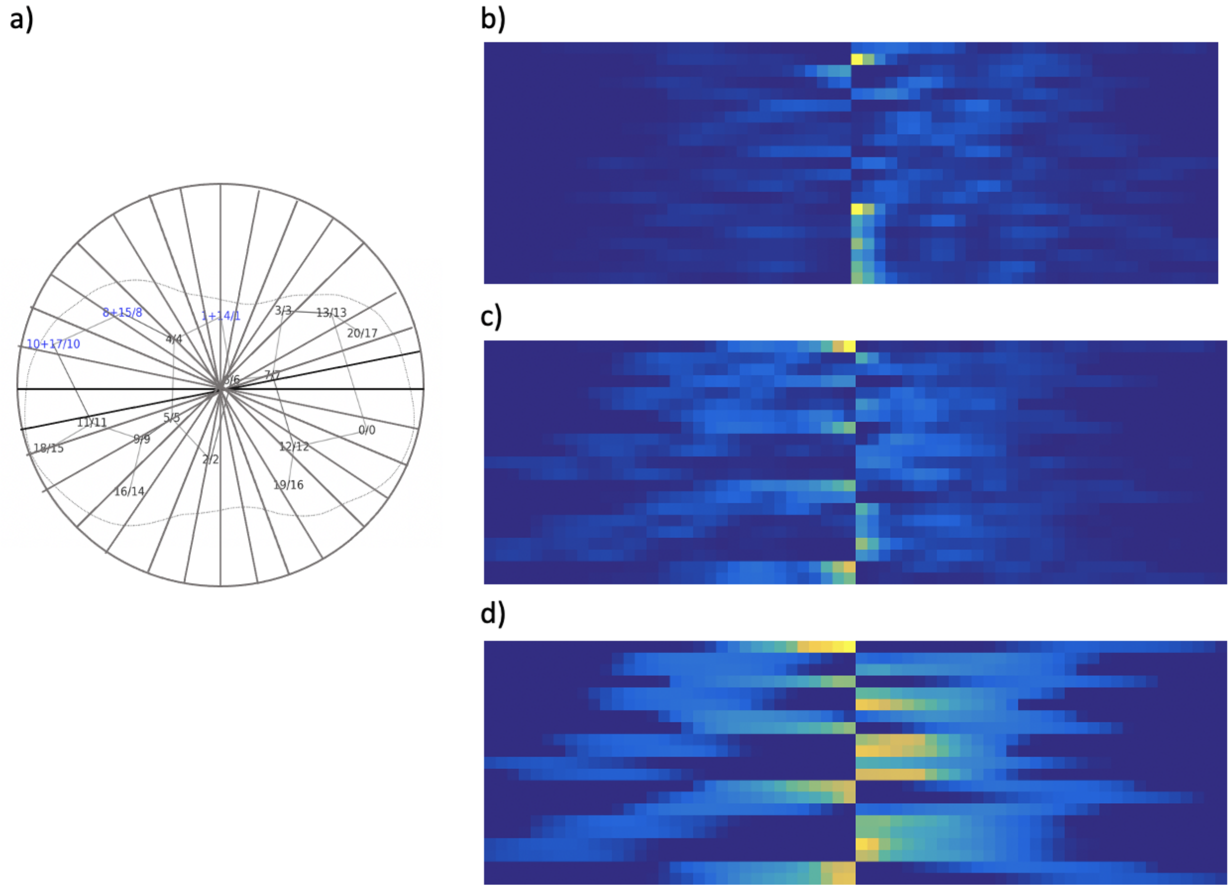


Figure 5.6. Log-polar histograms of the electronic area generated using each atom as the origin. (a) Diagram for shape context computation with 1 bin for log r with 32 bins for θ. (b) - (d) Log-polar histograms for ESP, Fukui functions, and Fukui potential, respectively.

5.3.2 Model Development

A neural network model is developed for each enzyme target. All developed models need to be validated. Here, we apply confusion matrix with each of the prediction. Figure 5.8 demonstrate results for data testing matrix with model trained for CYP1A2 closed map calculation with the highest F-score as 0.63 for testing dataset. Prediction performed poorly comparing with results in open map calculation shown in Figure 5.9 which achieved highest F-score 0.87. Prediction accuracy also improved to 0.79 with open map calculation comparing with results from close map calculation as 0.71 for testing dataset. Calculation with open cut maps for all other targets were able to yield significant better predictions than close

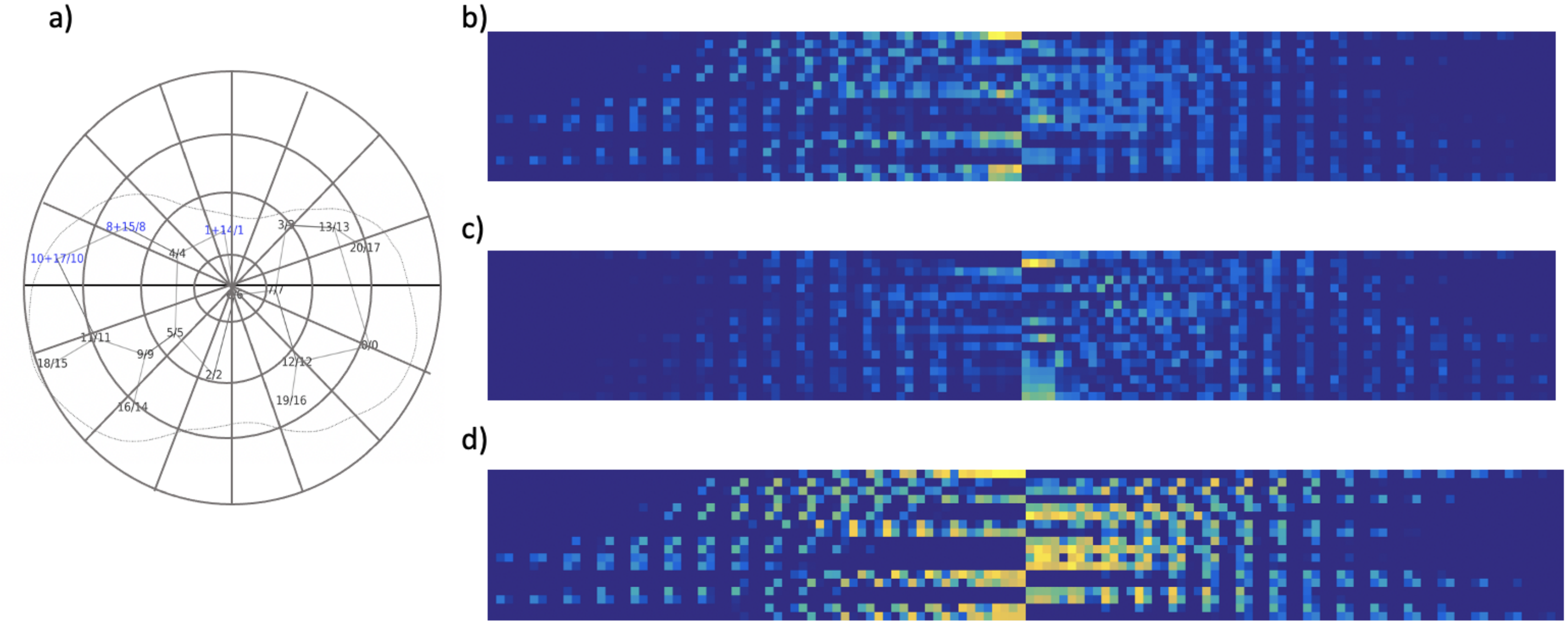


Figure 5.7. Log-polar histograms of the electronic area generated using each atom as the origin. (a) Diagram for shape context computation with 4 bins for log r with 16 bins for θ. (b) - (d) Log-polar histograms for ESP, Fukui functions, and Fukui potential, respectively.

map calculations illustrating the effect of false positives captured from the current close map calculation is unignorable. Table 5.1 summarized best trained NN models with respect to each CYP target with their F-score and accuracy. Besides for CYP2D6, strong model learning activity is observed early on during the iteration process. More complicated models were needed for CYP1A2 and CYP3A4. Further, CYP3A4 required the largest iteration with 81 epochs to pick up the features possibly due the flexibility of its intrinsic structure flexibility as described in literature [226], [227]. Nevertheless, open cut predictions seem to outperform closed map calculations. Significant information loss is observed with the noisy signals created. With open cut method, rational cuts could be achieved considering principal geodesic components that can normalize data points to better transform the information.

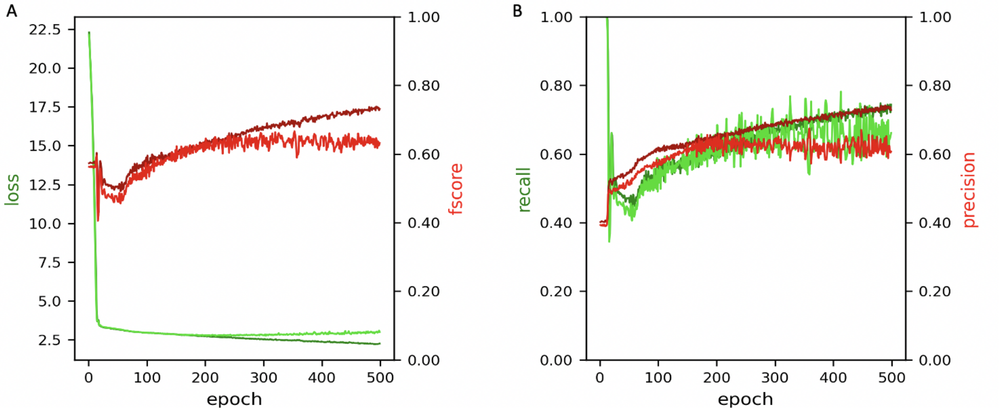


Figure 5.8. Results for data testing matrix with model trained for CYP1A2 closed map calculation. Dark red: F-score for training; Red: F-score for testing; Dark green: loss for training; green: loss for testing.

Table 5.1. NN models trained for each CYP target with best F-score and accuracy obtained with testing dataset. Models are listed with their layer features.

	F-score	Accuracy	Epoch	Model features
CYP1A2	0.87	0.79	40	[1024, 512, 256, 128, 64, 32, 16, 8, 1]
CYP2C19	0.80	0.73	32	[512, 256, 128, 64, 32, 16, 8, 1]
CYP2C9	0.72	0.72	54	[512, 256, 128, 64, 32, 16, 8, 1]
CYP2D6	0.48	0.57	196	[1960, 1024, 512, 256, 128, 64, 32, 16, 8, 1]
CYP3A4	0.78	0.74	81	[1024, 512, 256, 128, 64, 32, 16, 8, 1]

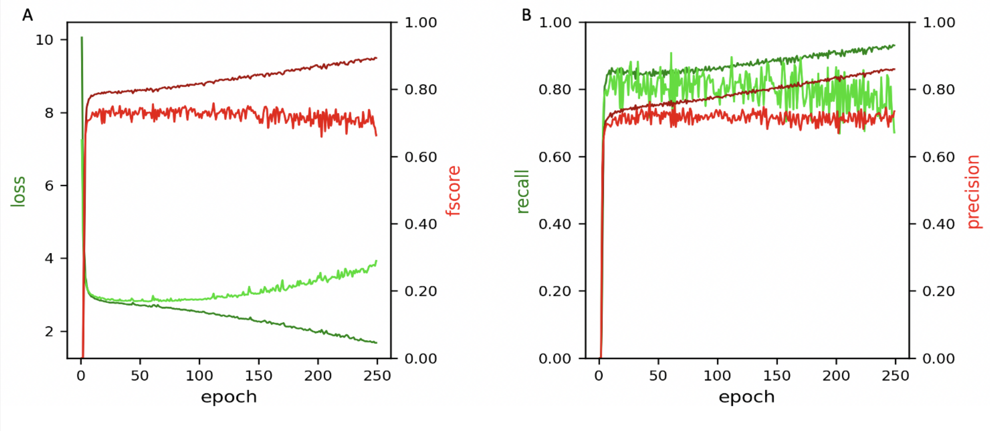


Figure 5.9. Results for data testing matrix with model trained for CYP1A2 closed map calculation.Dark red: F-score for training; Red: F-score for testing; Dark green: loss for training; green: loss for testing.

Figure 5.10 is the training results with all ligand included for CYP2D6. Poor learning is observed with current features extracted. A training with smaller dataset restricting ligand only considering the ones with atoms counts between 25 to 35 improved model learning behavior with reduced feature required as [1024, 512, 256, 128, 64, 32, 16, 8, 1]. Best F-score and accuracy were observed as 0.83 and 0.75 for testing dataset. Interestingly we found that restricing the input data with 10 atom counts differentiation is able to yield about 5% higher accuracy for all targets suggesting the need of a reasonable featurization method, such as Gaussian-based or radial based kernels, that may be applied to achieve analytical representations to better scale the matrix representation.

Comparing with current models predicting binary labels for CYP450 related DDIs with no interaction strength information, our model prediction is based on quantitative labels with Activity Score which is generated based on assay measurement with IC50. With the prediction error for Activity Score ranging from 0 to 100, our best prediction error is 11.5 with CYP1A2, followed with 12.4, 17.5, 21.4, 16.8 respectively for CYP2C19, CYP2C9, CYP2D6 and CYP3A4. The model developed provide better guidance and prioritize compounds based on biological activity strength. With quantitative predictions, the model have flexibility to perform classification with 3 groups corresponding to the classification indicators with

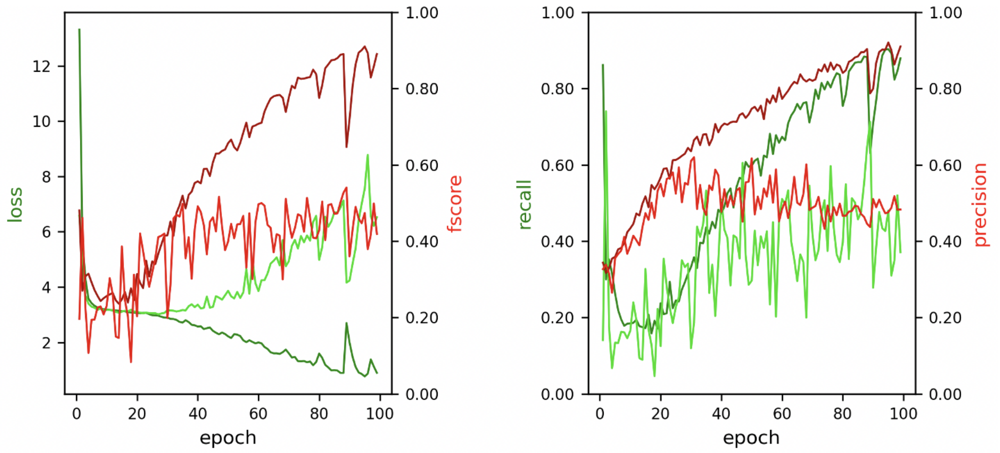


Figure 5.10. Results for data testing matrix with model trained for CYP1A2 closed map calculation. Dark red: F-score for training; Red: F-score for testing; Dark green: loss for training; green: loss for testing.

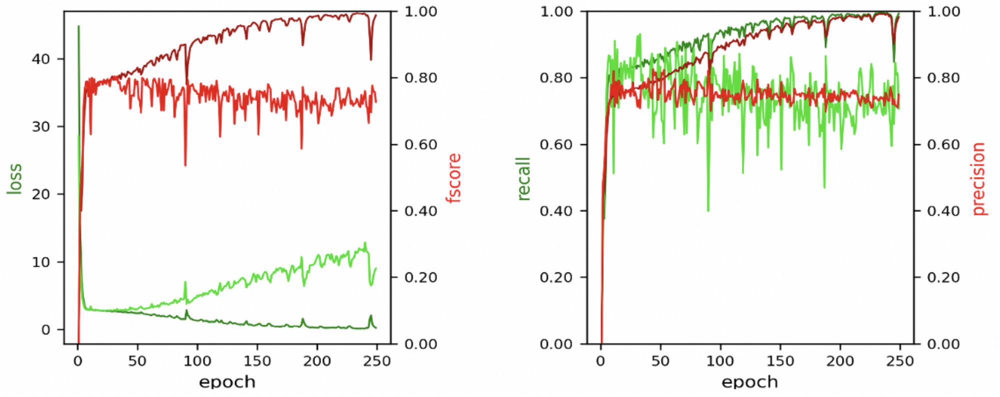


Figure 5.11. Results for data testing matrix with model trained for CYP1A2 closed map calculation.Dark red: F-score for training; Red: F-score for testing; Dark green: loss for training; green: loss for testing.

PubChem database. Further, regression predictions for active compounds were performed and yielded minimum prediction error 7.8 for CYP1A2 and below 15 for all other targets.

5.4 Conclusion

Model training was aimed to predict CYP450 related DDIs and distinguish active ligands from inactive ones with a diversified compound libraries. Strong model learning ability was demonstrated with specificity and strength for all five targets. Not only performing ligand classification and predicting potential DDIs caused by CYP450 enzymes, current learning scheme with MEMS features properly interpreted can be applied to rank the liangds based on quantitative biological activity strength. Open cuts manifold learning yield significant better predictions for all targets. The model is observed to be sensitive with atom counts within the molecules. To expand model applicability, kernel learning methods need to be implemented to improve size scaling.

6. SUMMARY AND OUTLOOK

Countless descriptors and computational methods have been developed in order to capture ”similarity”, where obvious limitations have been observed with the current approaches. The center of the problem is the definitive guideline for extracting accountable and trainable similarity features. It is becoming increasingly more apparent that structurally similarity alone does not equate to imparting biological activity similarity. Challenges still remain, and prediction performance can be further improved. Herein, a novel MEMS descriptor method was developed based on CDFT and manifold learning. This method features molecules based on electron density with its response to perturbation for analyzing chemical properties. Together with neural network, our method have demonstrated promising potentials for CYP450 related DDIs prediction, particularly for CYP450 inhibitions.

Instead of focusing on aspects of chemical properties and atomic structure like the current descriptors characterize, MEMS explore intermolecular interactions, both strength and locality, which are based on local electronic properties derived from CDFT on molecular surface supported by HSAB. The molecular interactions between a molecule and the binding site of a protein is assessed based on the first principles that define the origin of molecular interactions, both thermodynamically and kinetically, using quantum mechanics. The deep learning model with MEMS demonstrated a strong learning power over different targets with specificity with the best F-measure being 0.87% for CYP1A2 with the testing dataset. Regression predictions showed promising results with a prediction error within 20 for all targets, and the best 7.5 for the CYP2C19 prediction.

Another issue arising with current methods is the lack of neighborhood invariance of chemical space in the current descriptors, which can generate misleading results when dealing with activity cliffs. Neighborhood relationships can cause significant alterations in biological activities. Compounds residing in one chemical space may vary in another. Throughout the advancements in molecular descriptor development, MEMS bears machine trainable features able to capture neighborhood relationships. With the shape context method, neighborhood electronic features with respect to each atom is calculated with respect to ligand shape, carrying invariance chemical features that can be used to represent molecular information during model interpretation. DeepSet algorithms with attention mechanism preserve significant neighborhood invariant information to evaluate intermolecular relationships and can capture ”activity cliffs” with minor structural alterations.

In this study, over 14,000 compounds from an HTS assay measuring the dealkylation of pre-luciferin substrates to luciferin were applied for model training and testing. Additional studies on various assays should be explored to evaluate the learnability of the current scheme and specificity of the descriptor. It is worth noting that different molecular conformations which refer to the spatial arrangement alterations of the atoms remains unexplored in this study. With the current representation, it is possible to design a learning layer with rational assumptions to incorporate conformational changes with respect to atom features extracted. With the majority of information projected from 3D molecular surfaces to lower manifold, rational reduction could be achieved using principal geodesic components. It is crucial to apply the scheme using with molecules with a wide size range, and featurization methods with analytical representation kernels needs to be investigated to mitigate size effects and expand model applicability.

REFERENCES

[1] E. R. Hajjar, A. C. Cafiero, and J. T. Hanlon, “Polypharmacy in elderly patients,” *The American journal of geriatric pharmacotherapy*, vol. 5, no. 4, pp. 345–351, 2007.

[2] R. L. Maher, J. Hanlon, and E. R. Hajjar, “Clinical consequences of polypharmacy in elderly,” *Expert opinion on drug safety*, vol. 13, no. 1, pp. 57–65, 2014.

[3] R. B. Correia, L. P. de Araújo Kohler, M. M. Mattos, and L. M. Rocha, “City-wide electronic health records reveal gender and age biases in administration of known drug–drug interactions,” *NPJ digital medicine*, vol. 2, no. 1, pp. 1–13, 2019.

[4] D. M. Qato, J. Wilder, L. P. Schumm, V. Gillet, and G. C. Alexander, “Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the united states, 2005 vs 2011,” *JAMA internal medicine*, vol. 176, no. 4, pp. 473–482, 2016.

[5] P. Du Souich, “In human therapy, is the drug-drug interaction or the adverse drug reaction the issue?” *The Canadian journal of clinical pharmacology= Journal canadien de pharmacologie clinique*, vol. 8, no. 3, pp. 153–161, 2001.

[6] N. Masnoon, S. Shakib, L. Kalisch-Ellett, and G. E. Caughey, “What is polypharmacy? a systematic review of definitions,” *BMC geriatrics*, vol. 17, no. 1, pp. 1–10, 2017.

[7] C. Palleria, A. Di Paolo, C. Giofrè, C. Caglioti, G. Leuzzi, A. Siniscalchi, G. De Sarro, and L. Gallelli, “Pharmacokinetic drug-drug interaction and their implication in clinical management,” *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, vol. 18, no. 7, p. 601, 2013.

[8] T. Prueksaritanont, X. Chu, C. Gibson, D. Cui, K. L. Yee, J. Ballard, T. Cabalu, and J. Hochman, “Drug–drug interaction studies: Regulatory guidance and an industry perspective,” *The AAPS journal*, vol. 15, no. 3, pp. 629–645, 2013.

[9] J. Hochman, C. Tang, and T. Prueksaritanont, “Drug–drug interactions related to altered absorption and plasma protein binding: Theoretical and regulatory considerations, and an industry perspective,” *Journal of pharmaceutical sciences*, vol. 104, no. 3, pp. 916–929, 2015.

[10] N. Ai, X. Fan, and S. Ekins, “In silico methods for predicting drug–drug interactions with cytochrome p-450s, transporters and beyond,” *Advanced drug delivery reviews*, vol. 86, pp. 46–60, 2015.

[11] R. E. Schwartz, *Isolation and differentiation of adult and embryonic stem cells into hepatocytes in vitro*. University of Minnesota, 2006.

[12] H. Van De Waterbeemd and E. Gifford, “Admet in silico modelling: Towards prediction paradise?” *Nature reviews Drug discovery*, vol. 2, no. 3, pp. 192–204, 2003.

[13] S. E. O’Brien and M. J. de Groot, “Greater than the sum of its parts: Combining models for useful admet prediction,” *Journal of medicinal chemistry*, vol. 48, no. 4, pp. 1287–1291, 2005.

[14] L. C. Wienkers and T. G. Heath, “Predicting in vivo drug interactions from in vitro drug discovery data,” *Nature reviews Drug discovery*, vol. 4, no. 10, pp. 825–833, 2005.

[15] S. Sudsakorn, P. Bahadduri, J. Fretland, and C. Lu, “2020 fda drug-drug interaction guidance: A comparison analysis and action plan by pharmaceutical industrial scientists,” *Current Drug Metabolism*, vol. 21, no. 6, pp. 403–426, 2020.

[16] T. Iwatsubo, “Evaluation of drug–drug interactions in drug metabolism: Differences and harmonization in guidance/guidelines,” *Drug metabolism and pharmacokinetics*, vol. 35, no. 1, pp. 71–75, 2020.

[17] H. Kato, “Computational prediction of cytochrome p450 inhibition and induction,” *Drug metabolism and pharmacokinetics*, vol. 35, no. 1, pp. 30–44, 2020.

[18] L. Zhang, Y. D. Zhang, P. Zhao, and S.-M. Huang, “Predicting drug–drug interactions: An fda perspective,” *The AAPS journal*, vol. 11, no. 2, pp. 300–306, 2009.

[19] M. Dickins, “Induction of cytochromes p450,” *Current topics in medicinal chemistry*, vol. 4, no. 16, pp. 1745–1766, 2004.

[20] M. DAUJAT, P. CLAIR, C. ASTIER, I. FABRE, T. PINEAU, M. YERLE, J. GELLIN, and P. MAUREL, “Induction, regulation and messenger half-life of cytochromes p450 ia1, ia2 and iiia6 in primary cultures of rabbit hepatocytes: Cyp 1a1, 1a2 and 3a6 chromosome location in the rabbit and evidence that post-transcriptional control of gene ia2 does not involve mrna stabilization,” *European journal of biochemistry*, vol. 200, no. 2, pp. 501–510, 1991.

[21] O. Pelkonen, J. Mäeenpäeä, P. Taavitsainen, A. Rautio, and H. Raunio, “Inhibition and induction of human cytochrome p450 (cyp) enzymes,” *Xenobiotica*, vol. 28, no. 12, pp. 1203–1253, 1998.

[22] P. F. Hollenberg, “Characteristics and common properties of inhibitors, inducers, and activators of cyp enzymes,” *Drug metabolism reviews*, vol. 34, no. 1-2, pp. 17–35, 2002.

[23] M. Deodhar, S. B. Al Rihani, M. J. Arwood, L. Darakjian, P. Dow, J. Turgeon, and V. Michaud, "Mechanisms of cyp450 inhibition: Understanding drug-drug interactions due to mechanism-based inhibition in clinical practice," *Pharmaceutics*, vol. 12, no. 9, p. 846, 2020.

[24] O. A. Fahmi, T. S. Maurer, M. Kish, E. Cardenas, S. Boldt, and D. Nettleton, "A combined model for predicting cyp3a4 clinical net drug-drug interaction based on cyp3a4 inhibition, inactivation, and induction determined in vitro," *Drug Metabolism and Disposition*, vol. 36, no. 8, pp. 1698–1708, 2008.

[25] M. Jiménez, S. Chazarra, J. Escribano, J. Cabanes, and F. García-Carmona, "Competitive inhibition of mushroom tyrosinase by 4-substituted benzaldehydes," *Journal of Agricultural and Food Chemistry*, vol. 49, no. 8, pp. 4060–4063, 2001.

[26] O. K. F. Thu, O. Spigset, and B. Hellum, "Noncompetitive inhibition of human cyp 2c9 in vitro by a commercial rhodiola rosea product," *Pharmacology research & perspectives*, vol. 5, no. 4, e00324, 2017.

[27] Z.-Y. Zhang and Y. N. Wong, "Enzyme kinetics for clinically relevant cyp inhibition," *Current drug metabolism*, vol. 6, no. 3, pp. 241–257, 2005.

[28] B. Ring, S. A. Wrighton, and M. Mohutsky, "Reversible mechanisms of enzyme inhibition and resulting clinical significance," *Enzyme Kinetics in Drug Metabolism*, pp. 37–56, 2014.

[29] J. B. Houston and A. Galetin, "Modelling atypical cyp3a4 kinetics: Principles and pragmatism," *Archives of biochemistry and biophysics*, vol. 433, no. 2, pp. 351–360, 2005.

[30] S. T. Orr, S. L. Ripp, T. E. Ballard, J. L. Henderson, D. O. Scott, R. S. Obach, H. Sun, and A. S. Kalgutkar, "Mechanism-based inactivation (mbi) of cytochrome p450 enzymes: Structure–activity relationships and discovery strategies to mitigate drug–drug interaction risks," *Journal of medicinal chemistry*, vol. 55, no. 11, pp. 4896–4933, 2012.

[31] D. Spaggiari, Y. Daali, and S. Rudaz, "An extensive cocktail approach for rapid risk assessment of in vitro cyp450 direct reversible inhibition by xenobiotic exposure," *Toxicology and applied pharmacology*, vol. 302, pp. 41–51, 2016.

[32] Y. Naritomi, Y. Teramura, S. Terashita, and A. Kagayama, "Utility of microtiter plate assays for human cytochrome p450 inhibition studies in drug discovery: Application of simple method for detecting quasi-irreversible and irreversible inhibitors," *Drug metabolism and pharmacokinetics*, vol. 19, no. 1, pp. 55–61, 2004.

[33] J.-Y. Lee, S. Y. Lee, S. J. Oh, K. H. Lee, Y. S. Jung, and S. K. Kim, "Assessment of drug–drug interactions caused by metabolism-dependent cytochrome p450 inhibition," *Chemico-biological interactions*, vol. 198, no. 1-3, pp. 49–56, 2012.

[34] N. E. Thomford, K. Dzobo, D. Chopera, A. Wonkam, A. Maroyi, D. Blackhurst, and C. Dandara, "In vitro reversible and time-dependent cyp450 inhibition profiles of medicinal herbal plant extracts newbouldia laevis and cassia abbreviata: Implications for herb-drug interactions," *Molecules*, vol. 21, no. 7, p. 891, 2016.

[35] L. M. Berry and Z. Zhao, "An examination of ic50 and ic50-shift experiments in assessing time-dependent inhibition of cyp3a4, cyp2d6 and cyp2c9 in human liver microsomes.," *Drug metabolism letters*, vol. 2, no. 1, pp. 51–59, 2008.

[36] F. Alhaj, D. Qutishat, H. A. Harahsheh, N. Obeid, and B. Hammo, "Detecting ddi using ontology: Drug mechanism of action," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, IEEE, 2019, pp. 179–185.

[37] S. Fowler and H. Zhang, "In vitro evaluation of reversible and irreversible cytochrome p450 inhibition: Current status on methodologies and their utility for predicting drug–drug interactions," *The AAPS journal*, vol. 10, no. 2, pp. 410–424, 2008.

[38] Z. Yan, B. Rafferty, G. Caldwell, and J. Masucci, "Rapidly distinguishing reversible and irreversible cyp450 inhibitors by using fluorometric kinetic analyses," *European journal of drug metabolism and pharmacokinetics*, vol. 27, no. 4, pp. 281–287, 2002.

[39] J. E. Sager, S. Tripathy, L. S. Price, A. Nath, J. Chang, A. Stephenson-Famy, and N. Isoherranen, "In vitro to in vivo extrapolation of the complex drug-drug interaction of bupropion and its metabolites with cyp2d6; simultaneous reversible inhibition and cyp2d6 downregulation," *Biochemical pharmacology*, vol. 123, pp. 85–96, 2017.

[40] R. P. Hertzberg and A. J. Pope, "High-throughput screening: New technology for the 21st century," *Current opinion in chemical biology*, vol. 4, no. 4, pp. 445–451, 2000.

[41] J. Bajorath, "Integration of virtual and high-throughput screening," *Nature Reviews Drug Discovery*, vol. 1, no. 11, pp. 882–894, 2002.

[42] W. P. Walters, M. T. Stahl, and M. A. Murcko, "Virtual screening—an overview," *Drug discovery today*, vol. 3, no. 4, pp. 160–178, 1998.

[43] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug discovery today*, vol. 7, no. 20, pp. 1047–1055, 2002.

[44] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, “A geometric approach to macromolecule-ligand interactions,” *Journal of molecular biology*, vol. 161, no. 2, pp. 269–288, 1982.

[45] E. H. B. Maia, L. C. Assis, T. A. De Oliveira, A. M. Da Silva, and A. G. Taranto, “Structure-based virtual screening: From classical to artificial intelligence,” *Frontiers in chemistry*, vol. 8, p. 343, 2020.

[46] E. Lionta, G. Spyrou, D. K Vassilatis, and Z. Cournia, “Structure-based virtual screening for drug discovery: Principles, applications and recent advances,” *Current topics in medicinal chemistry*, vol. 14, no. 16, pp. 1923–1938, 2014.

[47] Q. Li and S. Shah, “Structure-based virtual screening,” in *Protein Bioinformatics*, Springer, 2017, pp. 111–124.

[48] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, “Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases,” *Journal of computer-aided molecular design*, vol. 15, no. 5, pp. 411–428, 2001.

[49] I. Schellhammer and M. Rarey, “Flexx-scan: Fast, structure-based virtual screening,” *PROTEINS: Structure, Function, and Bioinformatics*, vol. 57, no. 3, pp. 504–517, 2004.

[50] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks, “Glide: A new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening,” *Journal of medicinal chemistry*, vol. 47, no. 7, pp. 1750–1759, 2004.

[51] Z. Zsoldos, D. Reid, A. Simon, B. S. Sadjad, and A. Peter Johnson, “Ehits: An innovative approach to the docking and scoring function problems,” *Current Protein and Peptide Science*, vol. 7, no. 5, pp. 421–435, 2006.

[52] S. Vilar, G. Cozza, and S. Moro, “Medicinal chemistry and the molecular operating environment (moe): Application of qsar and molecular docking to drug discovery,” *Current topics in medicinal chemistry*, vol. 8, no. 18, pp. 1555–1572, 2008.

[53] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, “Development and validation of a genetic algorithm for flexible docking,” *Journal of molecular biology*, vol. 267, no. 3, pp. 727–748, 1997.

[54] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, “Autodock4 and autodocktools4: Automated docking with selective receptor flexibility,” *Journal of computational chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.

[55] S. Ghosh, A. Nie, J. An, and Z. Huang, "Structure-based virtual screening of chemical libraries for drug discovery," *Current opinion in chemical biology*, vol. 10, no. 3, pp. 194–202, 2006.

[56] H.-J. Böhm, "Ludi: Rule-based automatic design of new substituents for enzyme inhibitor leads," *Journal of Computer-Aided Molecular Design*, vol. 6, no. 6, pp. 593–606, 1992.

[57] A. J. Knox, M. J. Meegan, G. Carta, and D. G. Lloyd, "Considerations in compound database preparation "hidden" impact on virtual screening results," *Journal of chemical information and modeling*, vol. 45, no. 6, pp. 1908–1919, 2005.

[58] K. C. Parker, "Scoring methods in maldi peptide mass fingerprinting: Chemscore, and the chemapplex program," *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 1, pp. 22–39, 2002.

[59] W. J. Allen, T. E. Balius, S. Mukherjee, S. R. Brozell, D. T. Moustakas, P. T. Lang, D. A. Case, I. D. Kuntz, and R. C. Rizzo, "Dock 6: Impact of new features and current docking performance," *Journal of computational chemistry*, vol. 36, no. 15, pp. 1132–1156, 2015.

[60] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," *Journal of molecular biology*, vol. 295, no. 2, pp. 337–356, 2000.

[61] I. Banitt and H. J. Wolfson, "Paradock: A flexible non-specific dna—rigid protein docking algorithm," *Nucleic Acids Research*, vol. 39, no. 20, e135–e135, 2011.

[62] I. Muegge and Y. C. Martin, "A general and fast scoring function for protein- ligand interactions: A simplified potential approach," *Journal of medicinal chemistry*, vol. 42, no. 5, pp. 791–804, 1999.

[63] H. F. Velec, H. Gohlke, and G. Klebe, "Drugscorecsd knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction," *Journal of medicinal chemistry*, vol. 48, no. 20, pp. 6296–6303, 2005.

[64] A. Munir, S. Elahi, and N. Masood, "Clustering based drug-drug interaction networks for possible repositioning of drugs against egfr mutations: Clustering based ddi networks for egfr mutations," *Computational biology and chemistry*, vol. 75, pp. 24–31, 2018.

[65] C. Hansch and T. Fujita, "P-σ-$\boldsymbol{\pi}$ analysis. a method for the correlation of biological activity and chemical structure," *Journal of the American Chemical Society*, vol. 86, no. 8, pp. 1616–1626, 1964.

[66] D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[67] S. Nembri, F. Grisoni, V. Consonni, and R. Todeschini, “In silico prediction of cytochrome p450-drug interaction: Qsars for cyp3a4 and cyp2c9,” *International journal of molecular sciences*, vol. 17, no. 6, p. 914, 2016.

[68] A. D. Rodrigues and J. H. Lin, “Screening of drug candidates for their drug–drug interaction potential,” *Current opinion in chemical biology*, vol. 5, no. 4, pp. 396–401, 2001.

[69] X. Chen, Y. Lin, M. Liu, and M. K. Gilson, “The binding database: Data management and interface design,” *Bioinformatics*, vol. 18, no. 1, pp. 130–139, 2002.

[70] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “Biogrid: A general repository for interaction datasets,” *Nucleic acids research*, vol. 34, no. suppl_1, pp. D535–D539, 2006.

[71] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, “Chembl: A large-scale bioactivity database for drug discovery,” *Nucleic acids research*, vol. 40, no. D1, pp. D1100–D1107, 2012.

[72] P. V. Desai, A. Patny, Y. Sabnis, B. Tekwani, J. Gut, P. Rosenthal, A. Srivastava, and M. Avery, “Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. the chembridge database,” *Journal of medicinal chemistry*, vol. 47, no. 26, pp. 6609–6615, 2004.

[73] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, “Drugbank 5.0: A major update to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[74] A. Patwardhan, “Trends in the electron microscopy data bank (emdb),” *Acta Crystallographica Section D: Structural Biology*, vol. 73, no. 6, pp. 503–508, 2017.

[75] A. Shivanyuk, S. Ryabukhin, A. Tolmachev, A. Bogolyubsky, D. Mykytenko, A. Chupryna, W. Heilman, and A. Kostyuk, “Enamine real database: Making chemical diversity real,” *Chemistry today*, vol. 25, no. 6, pp. 58–59, 2007.

[76] M. Kanehisa and S. Goto, “Kegg: Kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[77] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, “Online mendelian inheritance in man (omim),” *Human mutation*, vol. 15, no. 1, pp. 57–61, 2000.

[78] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, “The pdbbind database: Methodologies and updates,” *Journal of medicinal chemistry*, vol. 48, no. 12, pp. 4111–4119, 2005.

[79] X. Wang, Y. Shen, S. Wang, S. Li, W. Zhang, X. Liu, L. Lai, J. Pei, and H. Li, “Pharmmapper 2017 update: A web server for potential drug target identification with a comprehensive target pharmacophore database,” *Nucleic acids research*, vol. 45, no. W1, W356–W360, 2017.

[80] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. E. Abola, “Protein data bank (pdb): Database of three-dimensional structural information of biological macromolecules,” *Acta Crystallographica Section D: Biological Crystallography*, vol. 54, no. 6, pp. 1078–1084, 1998.

[81] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant, “Pubchem as a public resource for drug discovery,” *Drug discovery today*, vol. 15, no. 23-24, pp. 1052–1057, 2010.

[82] C. v. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, “String: A database of predicted functional associations between proteins,” *Nucleic acids research*, vol. 31, no. 1, pp. 258–261, 2003.

[83] C. Y.-C. Chen, “Tcm database@ taiwan: The world’s largest traditional chinese medicine database for drug screening in silico,” *PloS one*, vol. 6, no. 1, e15939, 2011.

[84] X. Chen, Z. L. Ji, and Y. Z. Chen, “Ttd: Therapeutic target database,” *Nucleic acids research*, vol. 30, no. 1, pp. 412–415, 2002.

[85] F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, *et al.*, “Update of ttd: Therapeutic target database,” *Nucleic acids research*, vol. 38, no. suppl_1, pp. D787–D791, 2010.

[86] U. Consortium, “Uniprot: A worldwide hub of protein knowledge,” *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[87] J. J. Irwin and B. K. Shoichet, “Zinc- a free database of commercially available compounds for virtual screening,” *Journal of chemical information and modeling*, vol. 45, no. 1, pp. 177–182, 2005.

[88] M. A. Johnson and G. M. Maggiora, *Concepts and applications of molecular similarity.* Wiley, 1990.

[89] R. Cruz, G. Rojas, M. Quintero, and N. López, “Cluster analysis from molecular similarity matrices using a non-linear neural network,” *Journal of Mathematical Chemistry*, vol. 20, no. 2, pp. 385–394, 1996.

[90] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "Pubchem: Integrated platform of small molecules and biological activities," in *Annual reports in computational chemistry*, vol. 4, Elsevier, 2008, pp. 217–241.

[91] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. V. Tetko, and G. Marcou, "Isida-platform for virtual screening based on fragment and pharmacophoric descriptors," *Current Computer-Aided Drug Design*, vol. 4, no. 3, p. 191, 2008.

[92] R. Arimoto, M.-A. Prasad, and E. M. Gifford, "Development of cyp3a4 inhibition models: Comparisons of machine-learning techniques and molecular descriptors," *Journal of biomolecular screening*, vol. 10, no. 3, pp. 197–205, 2005.

[93] H. Briem and U. F. Lessel, "In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes," *Perspectives in Drug Discovery and Design*, vol. 20, no. 1, pp. 231–244, 2000.

[94] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," *Match*, vol. 56, no. 2, pp. 237–248, 2006.

[95] L. B. Kier, L. H. Hall, *et al.*, *Molecular structure description.* Academic, 1999.

[96] P. Franco, N. Porta, J. D. Holliday, and P. Willett, "The use of 2d fingerprint methods to support the assessment of structural similarity in orphan drug legislation," *Journal of cheminformatics*, vol. 6, no. 1, pp. 1–10, 2014.

[97] J. W. Raymond, C. J. Blankley, and P. Willett, "Comparison of chemical clustering methods using graph-and fingerprint-based similarity measures," *Journal of Molecular Graphics and Modelling*, vol. 21, no. 5, pp. 421–433, 2003.

[98] G. Landrum, "Rdkit documentation," *Release*, vol. 1, no. 1-79, p. 4, 2013.

[99] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of cheminformatics*, vol. 3, no. 1, pp. 1–14, 2011.

[100] A. Dalke, "The fps fingerprint format and chemfp toolkit," *Journal of cheminformatics*, vol. 5, no. 1, pp. 1–1, 2013.

[101] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics," *Journal of chemical information and computer sciences*, vol. 43, no. 2, pp. 493–500, 2003.

[102] T. Schablitzki, J. Rogal, and R. Drautz, “Topological fingerprints for intermetallic compounds for the automated classification of atomistic simulation data,” *Modelling and Simulation in Materials Science and Engineering*, vol. 21, no. 7, p. 075 008, 2013.

[103] W. A. Warr, “Scientific workflow systems: Pipeline pilot and knime,” *Journal of computer-aided molecular design*, vol. 26, no. 7, pp. 801–804, 2012.

[104] R. Todeschini and P. Gramatica, “The whim theory: New 3d molecular descriptors for qsar in environmental modelling,” *SAR and QSAR in Environmental Research*, vol. 7, no. 1-4, pp. 89–115, 1997.

[105] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, “Do structurally similar molecules have similar biological activity?” *Journal of medicinal chemistry*, vol. 45, no. 19, pp. 4350–4358, 2002.

[106] Y. Hu, D. Stumpfe, and J. Bajorath, “Advancing the activity cliff concept,” *F1000Research*, vol. 2, 2013.

[107] D. Stumpfe, D. Dimova, and J. Bajorath, “Composition and topology of activity cliff clusters formed by bioactive compounds,” *Journal of Chemical Information and Modeling*, vol. 54, no. 2, pp. 451–461, 2014.

[108] D. Stumpfe and J. Bajorath, “Exploring activity cliffs in medicinal chemistry: Miniperspective,” *Journal of medicinal chemistry*, vol. 55, no. 7, pp. 2932–2942, 2012.

[109] C. S. Leung, S. S. Leung, J. Tirado-Rives, and W. L. Jorgensen, “Methyl effects on protein–ligand binding,” *Journal of medicinal chemistry*, vol. 55, no. 9, pp. 4489–4500, 2012.

[110] J. L. Melville, E. K. Burke, and J. D. Hirst, “Machine learning in virtual screening,” *Combinatorial chemistry & high throughput screening*, vol. 12, no. 4, pp. 332–343, 2009.

[111] K. K. Chohan, S. W. Paine, J. Mistry, P. Barton, and A. M. Davis, “A rapid computational filter for cytochrome p450 1a2 inhibition potential of compound libraries,” *Journal of medicinal chemistry*, vol. 48, no. 16, pp. 5154–5161, 2005.

[112] K. Roy and P. P. Roy, “Comparative qsar studies of cyp1a2 inhibitor flavonoids using 2d and 3d descriptors,” *Chemical biology & drug design*, vol. 72, no. 5, pp. 370–382, 2008.

[113] L. Urra, M. Gonza’lez, and M. Teijeira, “Qsar studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases i and ii: 3d-morse descriptors and statistical considerations about variable selection,” *Bioorg. Med. Chem.*, vol. 14, pp. 7347–7358, 2006.

[114] J. Joyce, “Bayes’ theorem,” 2003.

[115] P. Labute, “Binary qsar: A new method for the determination of quantitative structure activity relationships,” in *Biocomputing'99*, World Scientific, 1999, pp. 444–455.

[116] J. M. Stevenson and P. D. Mulready, *Pipeline pilot 2.1 by scitegic, 9665 chesapeake drive, suite 401, san diego, ca 92123-1365. www. scitegic. com. see web site for pricing information*, 2003.

[117] A. E. Klon, M. Glick, M. Thoma, P. Acklin, and J. W. Davies, “Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results,” *Journal of medicinal chemistry*, vol. 47, no. 11, pp. 2743–2749, 2004.

[118] X. Xia, E. G. Maliski, P. Gallant, and D. Rogers, “Classification of kinase inhibitors using a bayesian model,” *Journal of medicinal chemistry*, vol. 47, no. 18, pp. 4463–4470, 2004.

[119] H. Sun, “A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing,” *Journal of Medicinal Chemistry*, vol. 48, no. 12, pp. 4031–4039, 2005.

[120] A. Bender, H. Y. Mussa, and R. C. Glen, “Screening for dihydrofolate reductase inhibitors using molprint 2d, a fast fragment-based method employing the naive bayesian classifier: Limitations of the descriptor and the importance of balanced chemistry in training and test sets,” *Journal of biomolecular screening*, vol. 10, no. 7, pp. 658–666, 2005.

[121] M. Jacobsson, P. Lidén, E. Stjernschantz, H. Boström, and U. Norinder, “Improving structure-based virtual screening by multivariate analysis of scoring data,” *Journal of medicinal chemistry*, vol. 46, no. 26, pp. 5781–5789, 2003.

[122] U. Fuhr, G. Strobl, F. Manaut, E. Anders, F. Sörgel, E. Lopez-de-Brinas, D. Chu, A. Pernet, G. Mahr, and F. Sanz, “Quinolone antibacterial agents: Relationship between structure and in vitro inhibition of the human cytochrome p450 isoform cyp1a2.,” *Molecular pharmacology*, vol. 43, no. 2, pp. 191–199, 1993.

[123] H. Lee, H. Yeom, Y. G. Kim, C. N. Yoon, C. Jin, J. S. Choi, B.-R. Kim, and D.-H. Kim, “Structure-related inhibition of human hepatic caffeine n3-demethylation by naturally occurring flavonoids,” *Biochemical pharmacology*, vol. 55, no. 9, pp. 1369–1375, 1998.

[124] T. Moon, M. H. Chi, D.-H. Kim, C. N. Yoon, and Y.-S. Choi, “Quantitative structure-activity relationships (qsar) study of flavonoid derivatives for inhibition of cytochrome p450 1a2,” *Quantitative Structure-Activity Relationships*, vol. 19, no. 3, pp. 257–263, 2000.

[125] J. Burton, I. Ijjaali, O. Barberan, F. Petitet, D. P. Vercauteren, and A. Michel, "Recursive partitioning for the prediction of cytochromes p450 2d6 and 1a2 inhibition: Importance of the quality of the dataset," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6231–6240, 2006.

[126] R. E. Bellman, *Adaptive control processes.* Princeton university press, 2015.

[127] D. S. Biovia *et al.*, *Discovery studio modeling environment*, 2017.

[128] C. Ulc, "Molecular operating environment (moe)," *Computing Group ULC*, vol. 1010, 2018.

[129] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[130] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2003, pp. 986–996.

[131] B. F. Jensen, C. Vind, S. B. Padkjær, P. B. Brockhoff, and H. H. Refsgaard, "In silico prediction of cytochrome p450 2d6 and 3a4 inhibition using gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors," *Journal of medicinal chemistry*, vol. 50, no. 3, pp. 501–511, 2007.

[132] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[133] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[134] V. Cherkassky and F. Mulier, "Vapnik-chervonenkis (vc) learning theory and its applications," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 985–987, 1999.

[135] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, 1996.

[136] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: Support vector machines for pharmaceutical data analysis," *Computers & chemistry*, vol. 26, no. 1, pp. 5–14, 2001.

[137] J. Burton, I. Ijjaali, F. Petitet, A. Michel, and D. P. Vercauteren, “Virtual screening for cytochromes p450: Successes of machine learning filters,” *Combinatorial chemistry & high throughput screening*, vol. 12, no. 4, pp. 369–382, 2009.

[138] P. H. Swain and H. Hauska, “The decision tree classifier: Design and potential,” *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[139] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[140] Y. Yuan, L. Wu, and X. Zhang, “Gini-impurity index analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3154–3169, 2021.

[141] S. Nembrini, I. R. König, and M. N. Wright, “The revival of the gini importance?” *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[142] S. Ekins, J. Berbaum, and R. K. Harrison, “Generation and validation of rapid computational filters for cyp2d6 and cyp3a4,” *Drug metabolism and disposition*, vol. 31, no. 9, pp. 1077–1080, 2003.

[143] S. Ekins, *In silico approaches to predicting drug metabolism, toxicology and beyond*, 2003.

[144] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[145] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, “Applications of deep learning and reinforcement learning to biological data,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 6, pp. 2063–2079, 2018.

[146] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, “Deep learning for computational biology,” *Molecular systems biology*, vol. 12, no. 7, p. 878, 2016.

[147] M. Mahmud, M. S. Kaiser, T. M. McGinnity, and A. Hussain, “Deep learning in mining biological data,” *Cognitive computation*, vol. 13, no. 1, pp. 1–33, 2021.

[148] A. M. Persky, “Multi-faceted approach to improve learning in pharmacokinetics,” *American Journal of Pharmaceutical Education*, vol. 72, no. 2, 2008.

[149] R. E. Dupuis and A. M. Persky, “Use of case-based learning in a clinical pharmacokinetics course,” *American journal of pharmaceutical education*, vol. 72, no. 2, 2008.

[150] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, “The rise of deep learning in drug discovery,” *Drug discovery today*, vol. 23, no. 6, pp. 1241–1250, 2018.

[151] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," *Molecular informatics*, vol. 35, no. 1, pp. 3–14, 2016.

[152] A. Lavecchia, "Deep learning in drug discovery: Opportunities, challenges and future prospects," *Drug discovery today*, vol. 24, no. 10, pp. 2017–2032, 2019.

[153] T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martínez-Mayorga, T. Langer, K. Cuanalo-Contreras, and D. K. Agrafiotis, "Recognizing pitfalls in virtual screening: A critical review," *Journal of chemical information and modeling*, vol. 52, no. 4, pp. 867–881, 2012.

[154] C. Tannenbaum and N. L. Sheehan, "Understanding and preventing drug–drug and drug–gene interactions," *Expert review of clinical pharmacology*, vol. 7, no. 4, pp. 533–544, 2014.

[155] F. Cheng, W. Li, G. Liu, and Y. Tang, "In silico admet prediction: Recent advances, current challenges and future trends," *Current topics in medicinal chemistry*, vol. 13, no. 11, pp. 1273–1289, 2013.

[156] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, *et al.*, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[157] *Dragon*, https://chm.kode-solutions.net/pf/dragon-7-0/#1621429726662-c07009e0-fb3b, Accessed: 2010-09-30.

[158] M. Zhang and T. Li, "Intermolecular interactions in organic crystals: Gaining insight from electronic structure analysis by density functional theory," *CrystEngComm*, vol. 16, no. 31, pp. 7162–7171, 2014.

[159] R. Bhattacharjee, K. Verma, M. Zhang, and T. Li, "Locality and strength of intermolecular interactions in organic crystals: Using conceptual density functional theory (cdft) to characterize a highly polymorphic system," *Theoretical Chemistry Accounts*, vol. 138, no. 11, pp. 1–14, 2019.

[160] J. Inglese, R. L. Johnson, A. Simeonov, M. Xia, W. Zheng, C. P. Austin, and D. S. Auld, "High-throughput screening assays for the identification of chemical probes," *Nature chemical biology*, vol. 3, no. 8, pp. 466–479, 2007.

[161] U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon, and S. C. Schürer, "Bioassay ontology (bao): A semantic description of bioassays and high-throughput screening results," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–16, 2011.

[162] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "Pubchem: A public information system for analyzing bioactivities of small molecules," *Nucleic acids research*, vol. 37, no. suppl_2, W623–W633, 2009.

[163] H. Veith, N. Southall, R. Huang, T. James, D. Fayne, N. Artemenko, M. Shen, J. Inglese, C. P. Austin, D. G. Lloyd, *et al.*, "Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries," *Nature biotechnology*, vol. 27, no. 11, pp. 1050–1055, 2009.

[164] J. J. Cali, D. Ma, M. Sobol, D. J. Simpson, S. Frackman, T. D. Good, W. J. Daily, and D. Liu, "Luminogenic cytochrome p450 assays," *Expert opinion on drug metabolism & toxicology*, vol. 2, no. 4, pp. 629–645, 2006.

[165] H. Sun, H. Veith, M. Xia, C. P. Austin, and R. Huang, "Predictive models for cytochrome p450 isozymes based on quantitative high throughput screening data," *Journal of chemical information and modeling*, vol. 51, no. 10, pp. 2474–2481, 2011.

[166] *MS Windows NT kernel description*, https://pubchem.ncbi.nlm.nih.gov/, Accessed: 2021-09-30.

[167] P. Hohenberg and W. Kohn, "Density functional theory (dft)," *Phys. Rev*, vol. 136, B864, 1964.

[168] J. M. Seminario, *Modern density functional theory: a tool for chemistry.* Elsevier, 1995.

[169] W. Kohn, A. D. Becke, and R. G. Parr, "Density functional theory of electronic structure," *The Journal of Physical Chemistry*, vol. 100, no. 31, pp. 12 974–12 980, 1996.

[170] R. G. Parr, *Reviews of modern quantum chemistry: a celebration of the contributions of Robert G. Parr.* World Scientific, 2002, vol. 1.

[171] P. Geerlings and F. De Proft, "Chemical reactivity as described by quantum chemical methods," *International Journal of Molecular Sciences*, vol. 3, no. 4, pp. 276–309, 2002.

[172] R. G. Parr *et al.*, "W. yang density functional theory of atoms and molecules," *Oxford University Press*, vol. 1, p. 1989, 1989.

[173] P. Geerlings, F. De Proft, and W. Langenaeker, "Conceptual density functional theory," *Chemical reviews*, vol. 103, no. 5, pp. 1793–1874, 2003.

[174] R. G. Pearson, "Hard and soft acids and bases, hsab, part 1: Fundamental principles," *Journal of Chemical Education*, vol. 45, no. 9, p. 581, 1968.

[175] R. G. Parr, “Density functional theory of atoms and molecules,” in *Horizons of quantum chemistry*, Springer, 1980, pp. 5–15.

[176] M. Berkowitz and R. G. Parr, “Molecular hardness and softness, local hardness and softness, hardness and softness kernels, and relations among these quantities,” *The Journal of chemical physics*, vol. 88, no. 4, pp. 2554–2557, 1988.

[177] Y. Li and J. N. Evans, “The fukui function: A key concept linking frontier molecular orbital theory and the hard-soft-acid-base principle,” *Journal of the American Chemical Society*, vol. 117, no. 29, pp. 7756–7759, 1995.

[178] T. C. Allison and Y. J. Tong, “Application of the condensed fukui function to predict reactivity in core–shell transition metal nanoparticles,” *Electrochimica Acta*, vol. 101, pp. 334–340, 2013.

[179] M. K. Harbola, R. G. Parr, and C. Lee, “Hardnesses from electrostatic potentials,” *The Journal of chemical physics*, vol. 94, no. 9, pp. 6055–6056, 1991.

[180] C. Cárdenas, W. Tiznado, P. W. Ayers, and P. Fuentealba, “The fukui potential and the capacity of charge and the global hardness of atoms,” *The Journal of Physical Chemistry A*, vol. 115, no. 11, pp. 2325–2331, 2011.

[181] J. S. Murray and P. Politzer, “The electrostatic potential: An overview,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 2, pp. 153–163, 2011.

[182] D. Santos-Martins, P. A. Fernandes, and M. J. Ramos, “Calculation of distribution coefficients in the sampl5 challenge from atomic solvation parameters and surface areas,” *Journal of computer-aided molecular design*, vol. 30, no. 11, pp. 1079–1086, 2016.

[183] D. Xu and Y. Zhang, “Generating triangulated macromolecular surfaces by euclidean distance transform,” *PloS one*, vol. 4, no. 12, e8140, 2009.

[184] J. Li, P. Mach, and P. Koehl, “Measuring the shapes of macromolecules–and why it matters,” *Computational and Structural Biotechnology Journal*, vol. 8, no. 12, e201309001, 2013.

[185] B. Lee and F. M. Richards, “The interpretation of protein structures: Estimation of static accessibility,” *Journal of molecular biology*, vol. 55, no. 3, 379–IN4, 1971.

[186] F. M. Richards, “Areas, volumes, packing, and protein structure,” *Annual review of biophysics and bioengineering*, vol. 6, no. 1, pp. 151–176, 1977.

[187] L. Van Der Maaten, E. Postma, J. Van den Herik, *et al.*, “Dimensionality reduction: A comparative,” *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[188] M. Reutlinger and G. Schneider, "Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery," *Journal of Molecular Graphics and Modelling*, vol. 34, pp. 108–117, 2012.

[189] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[190] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[191] R. Bro and A. K. Smilde, "Principal component analysis," *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[192] A. Giuliani, "The application of principal component analysis to drug discovery and biomedical data," *Drug discovery today*, vol. 22, no. 7, pp. 1069–1076, 2017.

[193] J. K. Westwick and S. W. Michnick, "Protein-fragment complementation assays (pca) in small gtpase research and drug discovery," *Methods in enzymology*, vol. 407, pp. 388–401, 2006.

[194] J. Larsson, J. Gottfries, S. Muresan, and A. Backlund, "Chemgps-np: Tuned for navigation in biologically relevant chemical space," *Journal of natural products*, vol. 70, no. 5, pp. 789–794, 2007.

[195] J. Rosén, A. Lövgren, T. Kogej, S. Muresan, J. Gottfries, and A. Backlund, "Chemgps-npweb: Chemical space navigation online," *Journal of computer-aided molecular design*, vol. 23, no. 4, pp. 253–259, 2009.

[196] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger, "Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors," *Journal of medicinal chemistry*, vol. 39, no. 16, pp. 3049–3059, 1996.

[197] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)," *Computer Science Review*, vol. 40, p. 100 378, 2021.

[198] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[199] J. Zupan, M. Novič, and I. Ruisánchez, "Kohonen and counterpropagation artificial neural networks in analytical chemistry," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, no. 1, pp. 1–23, 1997.

[200] D. Kirew, J. Chretien, P. Bernard, and F. Ros, “Application of kohonen neural networks in classification of biologically active compounds,” *SAR and QSAR in Environmental Research*, vol. 8, no. 1-2, pp. 93–107, 1998.

[201] I. I. Baskin, D. Winkler, and I. V. Tetko, “A renaissance of neural networks in drug discovery,” *Expert opinion on drug discovery*, vol. 11, no. 8, pp. 785–795, 2016.

[202] D. M. Maniyar, I. T. Nabney, B. S. Williams, and A. Sewing, “Data visualization during the early stages of drug discovery,” *Journal of chemical information and modeling*, vol. 46, no. 4, pp. 1806–1818, 2006.

[203] A. Lavecchia, “Machine-learning approaches in drug discovery: Methods and applications,” *Drug discovery today*, vol. 20, no. 3, pp. 318–331, 2015.

[204] D. Hristozov, T. I. Oprea, and J. Gasteiger, “Ligand-based virtual screening by novelty detection with self-organizing maps,” *Journal of chemical information and modeling*, vol. 47, no. 6, pp. 2044–2062, 2007.

[205] F. Bonachera, G. Marcou, N. Kireeva, A. Varnek, and D. Horvath, “Using self-organizing maps to accelerate similarity search,” *Bioorganic & medicinal chemistry*, vol. 20, no. 18, pp. 5396–5409, 2012.

[206] A. Ultsch, “Maps for the visualization of high-dimensional data spaces,” in *Proc. Workshop on Self organizing Maps*, 2003, pp. 225–230.

[207] D. K. Agrafiotis, “Stochastic proximity embedding,” *Journal of computational chemistry*, vol. 24, no. 10, pp. 1215–1221, 2003.

[208] D. K. Agrafiotis, H. Xu, F. Zhu, D. Bandyopadhyay, and P. Liu, “Stochastic proximity embedding: Methods and applications,” *Molecular informatics*, vol. 29, no. 11, pp. 758–770, 2010.

[209] S. Izrailev and D. K. Agrafiotis, “A method for quantifying and visualizing the diversity of qsar models,” *Journal of Molecular Graphics and Modelling*, vol. 22, no. 4, pp. 275–284, 2004.

[210] G. Aloor and L. Jacob, “Distributed wireless sensor network localization using stochastic proximity embedding,” *Computer Communications*, vol. 33, no. 6, pp. 745–755, 2010.

[211] Y. A. Ivanenkov, N. P. Savchuk, S. Ekins, and K. V. Balakin, “Computational mapping tools for drug discovery,” *Drug Discovery Today*, vol. 14, no. 15-16, pp. 767–775, 2009.

[212] G. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *NIPS*, Citeseer, vol. 15, 2002, pp. 833–840.

[213] F. Pérez-Cruz, “Kullback-leibler divergence estimation of continuous distributions,” in *2008 IEEE international symposium on information theory*, IEEE, 2008, pp. 1666–1670.

[214] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[215] A. P. Janssen, S. H. Grimm, R. H. Wijdeven, E. B. Lenselink, J. Neefjes, C. A. van Boeckel, G. J. van Westen, and M. van der Stelt, “Drug discovery maps, a machine learning model that visualizes and predicts kinome–inhibitor interaction landscapes,” *Journal of Chemical Information and Modeling*, vol. 59, no. 3, pp. 1221–1229, 2018.

[216] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, “Information retrieval perspective to nonlinear dimensionality reduction for data visualization.,” *Journal of Machine Learning Research*, vol. 11, no. 2, 2010.

[217] J. Wang and T. Hou, “Develop and test a solvent accessible surface area-based model in conformational entropy calculations,” *Journal of chemical information and modeling*, vol. 52, no. 5, pp. 1199–1212, 2012.

[218] A. Thayananthan, B. Stenger, P. H. Torr, and R. Cipolla, “Shape context and chamfer matching in cluttered scenes,” in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, IEEE, vol. 1, 2003, pp. I–I.

[219] S. Belongie, J. Malik, and J. Puzicha, “Shape context: A new descriptor for shape matching and object recognition,” *Advances in neural information processing systems*, vol. 13, pp. 831–837, 2000.

[220] S. Belongie, G. Mori, and J. Malik, “Matching with shape contexts,” in *Statistics and Analysis of Shapes*, Springer, 2006, pp. 81–105.

[221] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, “Deep sets,” *Advances in neural information processing systems*, vol. 30, 2017.

[222] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.

[223] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, *et al.*, “Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism,” *Journal of medicinal chemistry*, vol. 63, no. 16, pp. 8749–8760, 2019.

[224] J. Zhou, S. Li, L. Huang, H. Xiong, F. Wang, T. Xu, H. Xiong, and D. Dou, “Distance-aware molecule graph attention network for drug-target binding affinity prediction,” *arXiv preprint arXiv:2012.09624*, 2020.

[225] S. Maneewongvatana and D. M. Mount, “Analysis of approximate nearest neighbor searching with clustered point sets,” *arXiv preprint cs/9901013*, 1999.

[226] E. Anzenbacherová, N. Bec, P. Anzenbacher, J. Hudeček, P. Souček, C. Jung, A. W. Munro, and R. Lange, “Flexibility and stability of the structure of cytochromes p450 3a4 and bm-3,” *European Journal of Biochemistry*, vol. 267, no. 10, pp. 2916–2920, 2000.

[227] H. Yuki, T. Honma, M. Hata, and T. Hoshino, “Prediction of sites of metabolism in a substrate molecule, instanced by carbamazepine oxidation by cyp3a4,” *Bioorganic & medicinal chemistry*, vol. 20, no. 2, pp. 775–783, 2012.