A Toolbox of Generative Models and DTA Prediction for In-Silico Molecular Design and Drug Discovery

by

Azamat Bakytzhan

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

April 2023

Author . . . .
Department of Computer Science
2023

Certified by . . . .
Siamac Fazli
Associate Professor, dept. Computer Science
Thesis Supervisor

Accepted by . . . .
Dean, School of Engineering and Digital Sciences

A Toolbox of Generative Models and DTA Prediction for In-Silico Molecular Design and Drug Discovery

by

Azamat Bakytzhan

Submitted to the Department of Computer Science
on 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

The typical drug development process involves multiple stages, including target identification, target validation, lead discovery, lead optimizations, ADMET evaluation, and several phases of clinical trials leading to registration [27]. This standard flow usually spans around 17 years, with the chances of successful drug registration being only 1 out of 5000 [26].

To facilitate in-silico studies, various tools like RDKit [28], Open Babel [29], SWISS-MODEL [32], and AutoDock Vina [30] have been developed. However, the decentralized development of many frameworks has given rise to challenges like version conflicts, platform dependencies, complex installations, and scattered knowledge. This makes it difficult and time-consuming for new researchers to get onboarded in the domain, often requiring several weeks or even half a year to study existing frameworks and workflows.

My framework, DDBox, aims to address these issues by consolidating the most popular tools into a single platform. By doing so, it not only simplifies in-silico studies but also contributes to knowledge sharing in the in-silico drug design field.

Keywords — Toolbox, Framework, Python Package, Platform, In-Silico Drug Discovery, De Novo Molecular Design, Drug-Target Affinity, Benchmark, Docking, System Design, Drug Design.

Thesis Supervisor: Siamac Fazli
Title: Associate Professor, dept. Computer Science

Acknowledgments

I am deeply grateful to Professor Siamac Fazli and Nazarbayev University for providing an excellent research direction and the opportunity to pursue my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Drug discovery

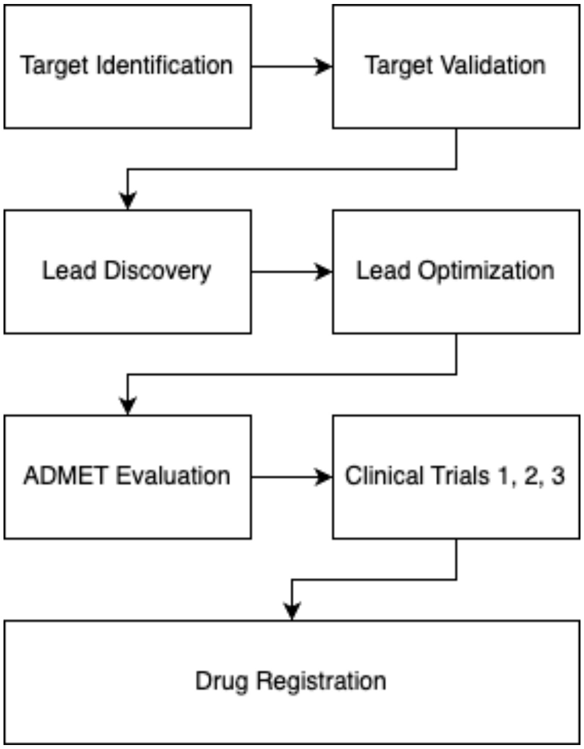


Figure 1-1: Standard Workflow.

In earlier times, researchers discovered medicinal compounds from plants using random screening and trial-and-error techniques [35], which involved conducting experiments on animals and humans. While this approach did yield results, it also posed considerable risks to researchers. Given the limited availability of systematic knowledge back then, this method was practical in exploring potential medicinal properties.

With the advancement of science, society, and human knowledge, drug development techniques have evolved to become safer and less hazardous. Nowadays, there is a standardized process for drug development, which includes several ordered stages: target identification, target validation, lead discovery, lead optimization, ADMET evaluation, and multiple phases of clinical trials leading to drug registration (see figure 1-1) [27].

In recent years, in-depth research on medications and plants has resulted in a gradual reduction in the need for testing drugs in-vivo on animals and humans during clinical trials. This progress stems from a better understanding of medicinal properties and an increased focus on developing safer and more efficient drug development approaches.

1.2 A drug from the molecular perspective

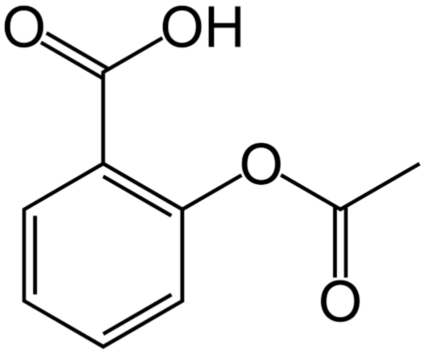


Figure 1-2: Aspirin Molecule. SMILES: O=C(C)Oc1ccccc1C(=O)O.

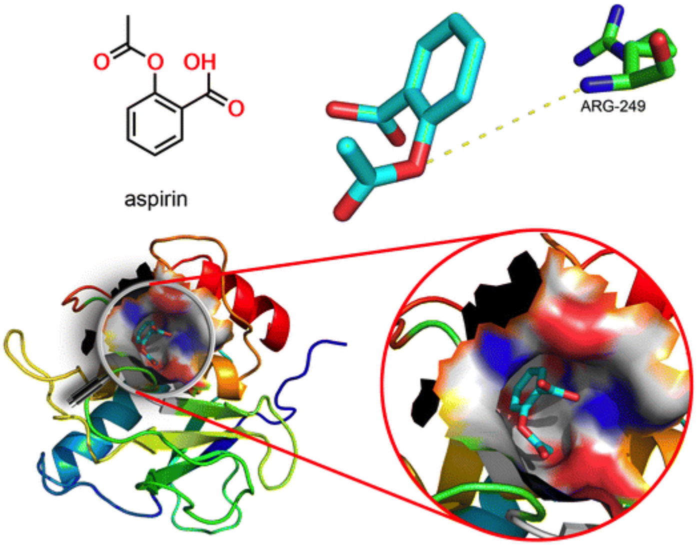


Figure 1-3: Docking of aspirin into the binding pocket of MMP-9. Atoms and bonds were displayed as sticks, protein was displayed as solid ribbon, and the binding pocket was displayed as surface (In Silicon Approach for Discovery of Chemopreventive Agents - Scientific Figure on ResearchGate [92]).

From a molecular perspective, a drug refers to a chemical substance (see figure 1-2 of aspirin molecule) that interacts with specific molecules within the body, predominantly proteins or enzymes, in order to produce a therapeutic effect. Drugs can achieve this by either binding to the target molecules and altering their activity or by inhibiting their function entirely (see figure 1-3 of docking of aspirin molecule). However, it's important to note that in some instances, drugs may also bind to other molecules within the body, leading to potential side effects or unintended interactions [80].

The design of a drug molecule is carefully crafted to enhance its interaction with the target molecules, utilizing specific features such as shape, charge, or other chemical properties of the target site [81]. For instance, some drugs are intentionally shaped to resemble natural ligands that bind to the target protein, while others are designed to attach to a specific region of the protein, effectively blocking its activity . The effectiveness of a drug relies on its ability to bind selectively to the target molecule while minimizing any unintended effects on other molecules and potentially harmful side effects [82].

Drugs can be grouped into different categories, primarily based on their molecular structure and how they work in the body. One classification includes small molecules, which are typically organic compounds with a molecular weight below 1000 Da, making them the most common type of drugs. On the other hand, biologics are large molecules derived from living cells, including proteins, antibodies, and nucleic acids, among others.

The molecule that a drug interacts with is often called its "target protein" or "receptor." This target protein can be found on the surface of a cell, where it serves as a receptor for neurotransmitters or hormones. Alternatively, it may be an intracellular protein that plays a role in specific signaling pathways within the cell.

Drugs can interact with their target molecules in various manners. For instance, certain drugs directly bind to the active site of the protein, while others may attach to a different site, leading to a conformational change in the protein. Additionally, some drugs bind reversibly, while others do so irreversibly. The strength of the binding can influence the drug's potency and the duration of its action.

The specificity of a drug's interaction with its target molecule primarily relies on the molecular complementarity between the drug and the target site. This complementarity involves various factors, including electrostatics, hydrogen bonding, and hydrophobic interactions.

It's crucial to consider that drugs can interact with other molecules in the body, leading to potential side effects or unintended outcomes. For instance, some drugs might bind to off-target proteins, resulting in toxicity, or interact with other drugs to produce undesired interactions [37].

Creating a new drug usually involves a multi-step process that combines molecular design, chemical synthesis, and testing both in laboratory settings (in-vitro) and in living organisms (in-vivo) to evaluate its effectiveness, safety, and how it behaves in the body (pharmacokinetics). Nowadays, computational methods, such as molecular modeling and machine learning, are indispensable tools in drug discovery. These techniques help researchers in designing and improving new drug candidates by predicting their behavior and interactions with target molecules.

1.3 Drug studies

1.3.1 In-silico

In-silico pharmacology, also known as computational therapeutics or computational pharmacology, is an increasingly prominent field that aims to harness the power of software-based methods to gather, analyze, and integrate biological and medical data from diverse sources. This discipline delves into the intricacies of using this data to construct computer models or simulations that can predict outcomes, propose hypotheses, and ultimately drive advancements and improvements in medical treatments and medicine.

The utilization of in-silico techniques has become vital in the quest for discovering new drugs. These methods have the potential to revolutionize the entire drug development process by identifying and exploring new potential treatments more affordably and in less time. Computer-aided drug design (CADD) techniques play a significant role in minimizing the need for animal testing in experiments, promoting the development of safer medications, and repositioning existing drugs for novel applications. By employing these methods, medicinal chemists can effectively design, discover, develop, and optimize drugs. In contrast, traditional drug discovery methods often involve expensive and random screenings of artificial or organic compounds.

1.3.2 In-vitro

In the pre-clinical trials stage, researchers have data on the effectiveness of a drug candidate against a specific disease or infection. However, vital information regarding how this chemical behaves in terms of pharmacokinetics, toxicity, safety, and metabolism in humans remains unknown. Therefore, it becomes essential to determine the initial dosage and thoroughly assess all the mentioned criteria. It's worth noting that only one out of every 5000 substances that undergo pre-clinical testing ultimately receives approval as a viable drug [26].

The safety pharmacology stage plays a crucial role in the pre-clinical development of drugs, with the primary goal of detecting any possible harmful effects of a medication on the major physiological systems of the human body. Historically, pharmaceutical safety testing heavily relied on animal studies, but there is a growing trend toward utilizing in-vitro tests with established tissues and cell lines. This shift allows for more precise and controlled assessments of a drug's safety profile before advancing to clinical trials.

In-vitro studies, which take place outside of a living organism, have been traditionally regarded as more cost-effective compared to in-vivo testing. Although they are less expensive, researchers need to carefully consider whether in-vitro tests are the most suitable choice for their specific research objectives. While in-vitro assays can be valuable for detecting potential carcinogenic or hazardous reactions, they may not always fully capture the complexity of interactions that occur within a living system. Hence, researchers must weigh the advantages and limitations of in-vitro studies before deciding on the appropriate approach for their investigations.

In drug efficacy evaluations, researchers focus on determining the concentrations of drugs that lead to specific pharmacological effects, such as affinity, potency, and effectiveness. To achieve this, they place the drug in a pre-defined volume, such as a test tube or a well of a plate, and then analyze its effects at both molecular and physiological levels. This method allows researchers to design compounds that have the potential to efficiently target and combat specific proteins, making it a valuable tool in drug development and design.

1.3.3 In-vivo

The term "in-vivo" pertains to processes occurring within a living organism, and it is logical to perform tests within a living model. In addition to using rats, rabbits, and other higher mammals, animals can also serve as in-vivo models. This stage in the drug discovery process is crucial as it showcases how various chemicals affect complex organisms like animals. Animals are employed in in-vivo testing because of their close genetic resemblance to humans, which preserves most biological pathways and the intricacy of the entire body system, something that cannot be replicated in-vitro studies.

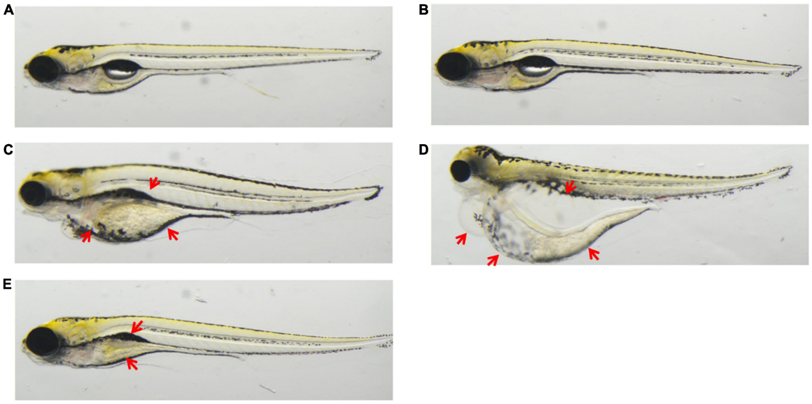


Figure 1-4: Application: triptolide, aconitine, chlorogenic acid, treatment to zebrafish at 5 dpf. (A) Untreated. (B) 0.5% DMSO. (C) $0.1\mu M$ of triptolide treatment led to zebrafish renal edema, pericardial edema, and cardiovascular toxicity. (D) $100\mu M$ of aconitine treatment led to zebrafish having renal edema and pericardial edema. (E) $100\mu M$ of chlorogenic acid treatment led to zebrafish showing swim bladder loss and delayed yolk sac absorption. Author: Song et. al. (Validation, optimization, and application of the zebrafish developmental toxicity assay for pharmaceuticals under the ICH S5 [93]).

Zebrafish have gained significant popularity in preclinical research, particularly in in-vivo testing, due to several advantages they offer over other animals such as rats, rabbits, and mammals. Utilizing zebrafish allows scientists to minimize the use of more complex animals and adhere to the 3Rs principles, which focus on the Replacement, Reduction, and Refinement of animal usage in preclinical research [83]. Zebrafish exemplify the 3R principle as they are highly adaptable to different environments and experimental conditions compared to some rodents, effectively bridging the gap between in-vitro and in-vivo studies. Additionally, they are more cost-effective to house, maintain, and breed. Their genetic similarity to humans, sharing approximately 70% of human DNA [84], facilitates the assessment of drug efficacy and toxicity and enables better extrapolation of results to human use. Overall, zebrafish present a more efficient and cost-effective approach to preclinical research, reducing the reliance on other animal models [38].

1.4 Data-driven drug discovery

Developing new medicines is important for the economy and public health. But it can be tough because many drugs fail during testing, and the ones that make it to the market don't always make enough money. This makes the process long and expensive. We also have a lot of data to handle, so we need better ways to analyze it. That's where data-driven approaches come in handy. They help us discover drugs more effectively and accurately. So, using data-driven methods is becoming more and more important in drug discovery.

Data-driven drug discovery is a process that involves using computational and data science tools to analyze extensive datasets, such as genomic, proteomic, clinical, and chemical data, to discover new drug targets and candidates [85, 86]. The main idea is that the abundance of data in biology and medicine can help identify potential drug targets and candidates more efficiently than traditional methods. By analyzing large datasets, data-driven drug discovery can reveal hidden relationships and patterns that may not be evident through human intuition or conventional laboratory techniques. Successful examples include finding correlations between new and existing drugs, evaluating model performance on subsets of the dataset, and uncovering hidden patterns in molecules.

Data-driven drug discovery offers diverse implementation methods, utilizing various types of data, such as genomic and proteomic data, for target identification. For example, machine learning algorithms can analyze gene expression data to identify genes that show differential expression in disease states, potentially serving as drug targets. In Drug Target Binding Affinity (DTA), data-driven approaches employ machine learning algorithms to predict the properties of new molecules from vast chemical libraries. Deep learning algorithms, for instance, can predict the binding affinity of a molecule to a target protein, leading to the development of more effective and selective drugs. Moreover, data-driven techniques can aid in de novo molecular design [2] by analyzing extensive datasets of known molecules and calculating their latent space to generate novel molecules for screening. Additionally, data-driven approaches can optimize clinical trials by examining comprehensive patient datasets. Machine learning algorithms can identify patient subgroups that are more likely to respond to specific treatments, resulting in more targeted and personalized clinical trials.

Data-driven drug discovery holds immense promise in speeding up and improving the drug discovery process by facilitating the efficient identification of new drug targets and candidates. However, it is essential to acknowledge the limitations and challenges associated with this approach. One crucial factor is the quality and quantity of available data, as insufficient or unreliable data can hinder accurate predictions. Another consideration is the reliability of machine learning algorithms, which need to be carefully assessed and validated to ensure their suitability for the task at hand. Additionally, it is crucial to validate the results obtained through data-driven methods in the laboratory to ensure their accuracy and relevance before proceeding to the next stages of drug development. By approaching data-driven drug discovery with caution and rigorously validating the results, researchers can leverage the full potential of this approach and pave the way for more successful and impactful drug discoveries.

1.5 De-novo molecular design

De novo molecular design [2] is an innovative approach to drug discovery that involves creating entirely new molecules from scratch using computational methods and algorithms (see sample generative models in figure 1-5). Unlike traditional drug design, which focuses on modifying existing compounds, de novo design aims to develop molecules with specific desired properties right from the beginning.

The process of de novo design typically follows a series of steps. It starts with identifying the molecular target, which is the specific biological or chemical process that the molecule is intended to interact with. Next, a library of potential molecular structures is generated using sophisticated computational methods, such as molecular modeling and simulation. These generated structures are then thoroughly evaluated and ranked based on various criteria, such as their predicted properties, including drug-likeness or potential toxicity.

De novo molecular design offers researchers the opportunity to explore a wide array of possibilities and craft novel molecules with customized properties to suit various applications, particularly in drug development. This exciting approach holds the potential for innovative and precise drug discovery, enabling the creation of more effective and targeted therapies.

Following the selection of top candidates, the subsequent crucial phase in de novo drug design involves refining and optimizing these molecules through a series of iterations. This iterative process entails fine-tuning the structure of the molecule, carefully placing specific functional groups, and adjusting other molecular properties to bolster their potential as promising drug candidates. The ultimate objective is to enhance their efficacy and therapeutic potential, paving the way for successful drug development.

Once the molecules are refined, they undergo laboratory synthesis, and their biological activity and therapeutic potential are tested in in-vitro or in-vivo experiments to validate their efficacy.

In recent years, deep learning models have made significant strides, particularly in natural language processing, for example Transformers [87]. These advancements have led to the development of generative machine learning models applicable to de-novo drug design, for example Transmol [88]. These models utilize deep neural networks and are trained to recognize molecular structures. They can then generate entirely new molecular entities without relying on predetermined sets of building blocks and chemical transformations. This approach shows great promise in accelerating the discovery of novel chemical compounds and potentially reducing the time and costs associated with drug development. It opens new avenues for innovative drug discovery and optimization.

The de novo molecular design method is a data-driven approach used in drug discovery to create new drug compounds. This technique involves employing an encoder model to generate points in the latent space and a decoder model to retrieve the latent space. By selecting a specific point in the latent space with desired characteristics, researchers can synthesize new drug candidates. The success of de novo molecular design relies on the statistical analysis between the newly generated molecules and existing ones. However, it's important to note that de novo design doesn't guarantee the efficacy of the new drug, and rigorous pre-clinical and clinical trials are necessary before its production.

While de novo design has shown promising results in various studies, it remains a relatively new and challenging field with several obstacles that require careful consideration. One of the primary concerns is the accuracy of the computational models used to generate new molecules, as the success of de novo design hinges on the reliability of these models. Moreover, the chemical space itself is highly intricate and diverse, posing complexities in identifying novel and effective drug candidates. Additionally, current synthesis techniques have limitations in creating complex molecular structures, which further complicates the process. Despite these challenges, researchers are optimistic about the potential of de novo design in accelerating drug discovery and streamlining drug development, making it an area of active exploration and development. Continued research and advancements in this domain will be pivotal in overcoming these obstacles and fully harnessing the capabilities of de novo molecular design.

1.6 Genertive Models

Generative models are computational models used in drug discovery to generate new small molecule candidates with desired properties. These models use machine learning algorithms and deep learning techniques to learn patterns in large datasets of molecules and their properties, and then use that knowledge to generate new molecules with similar properties.

There are various types of generative models used in drug discovery, including variational autoencoders (VAEs), recurrent neural networks (RNNs), and generative adversarial networks (GANs) (see sample generative models in the figure 1-5). VAEs

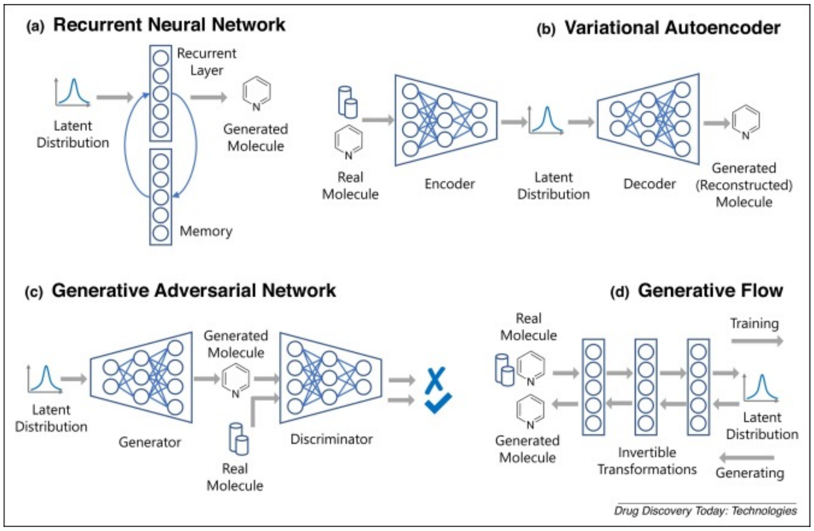


Figure 1-5: Generative Models. Author: Xia et. al. (Graph-based generative models for de Novo drug design [94]).

are used to learn the probability distribution of the training data and generate new samples from that distribution. RNNs are used for sequence modeling, which is useful for generating molecules with specific patterns or structures. GANs are used to generate new molecules by learning the distribution of the training data and generating new samples that are difficult to distinguish from real data.

These models can be trained on a variety of chemical and biological data, such as chemical structure and activity data, gene expression data, and clinical data. By using these generative models, drug discovery researchers can generate new molecules with desired properties, potentially accelerating the drug discovery process and reducing the cost of drug development.

Generative models can be utilized to produce new molecules that have specific properties, such as high potency or solubility, once they have been trained using deep learning algorithms. These new molecules can then be synthesized and tested in a laboratory setting to determine their biological activity and assess their potential for use in drug development.

Generative models can speed up drug discovery by simulating the generation of new compounds without the need for extensive experimental screening and synthesis, potentially saving both time and money. They can also be used to search for novel chemical structures that traditional methods may not easily identify.

However, one of the challenges of using generative models is finding a balance between the novelty and quality of the generated molecules. The models must produce diverse molecules with desirable properties that are chemically feasible, while ensuring that the generated molecules are safe and effective. To achieve this balance, the models must be thoroughly validated and tested to ensure that the generated molecules meet the required standards.

Despite these challenges, generative models have the potential to revolutionize the drug discovery process and lead to the development of new and more effective treatments for a variety of diseases.

1.7 DTA Models

Drug-target affinity (DTA) models are computational models used in drug discovery to predict how well small molecules bind to specific targets, such as proteins or enzymes. The strength of this interaction, known as the binding affinity, is a crucial factor in determining the biological activity and potential efficacy of a drug candidate.

DTA models use a variety of computational methods, such as molecular docking, molecular dynamics simulations, and machine learning algorithms to predict the binding affinity of small molecules to their targets. To train these models, large datasets of small molecules and their corresponding binding affinities are used. Once trained, the models can be used to predict the binding affinity of new small molecules, potentially accelerating the drug discovery process by helping to identify promising drug candidates more efficiently.

There are two main types of DTA models [91]: ligand-based [89] and structure-based [90].

Ligand-based models use the properties of known ligands to predict the affinity of new molecules. They assume that molecules with similar structures have similar binding affinities. These models use molecular descriptors to encode the structural and physicochemical properties of small molecules and their ligands.

Structure-based models, on the other hand, use the 3D structure of the target and the small molecule to predict the binding affinity. These models take into account protein conformational changes, solvent, and ligand flexibility, which can affect the binding affinity.

DTA models can be used in virtual screening, where large libraries of small molecules are screened against a target to identify potential drug candidates. DTA models help to prioritize the molecules that are most likely to bind to the target and have the desired biological effect.

Chapter 2

Literature Review

2.1 Molecular Representations

2.1.1 BoW

The Bag-of-words (BoW) molecule representation is a method for representing molecules in a format that can be processed by a computer. This approach is similar to the BoW model used in natural language processing. To use this method, a molecule is first converted into a string of characters or a sequence of atom types, which is then processed using a tokenization algorithm to create a set of features, or "words". These features are typically individual atoms or pairs of neighboring atoms. Each molecule is then assigned a frequency vector, which represents the frequency of each feature in the molecule. This vector is mostly empty, with only a few non-zero values [44].

One limitation of the BoW molecule representation is that it does not consider the spatial arrangement of atoms in the molecule, which can be important in determining its behavior and properties. Other methods, such as graph-based or 3D conformations, can be used to account for this information.

2.1.2 Graph Representation

In cheminformatics, graph representation is a popular way to represent molecules, which captures both structural and topological information. This approach represents the atoms and bonds of a molecule as nodes and edges in a graph, respectively. Each atom is a node, and each bond is an edge, with the atom and bond type encoded as node and edge labels. Additional attributes such as charges and hybridization can be included as node properties.

There are different types of graph representations for encoding molecular structures, including adjacency matrix-based and edge list-based representations. Adjacency matrix-based representations use a matrix to represent the molecule, with rows and columns corresponding to atoms and matrix elements corresponding to the edges. Edge list-based representations use a list of edges, where each edge is represented as a tuple of two atoms and a bond type.

Graph representations are advantageous because they can capture spatial and topological information, which can be used to identify substructures and molecular patterns to aid in drug discovery. Additionally, graph representations can be easily analyzed using graph theory algorithms to extract meaningful features and patterns from the molecular structure [45].

2.1.3 SMILES

SMILES (Simplified Molecular Input Line Entry System) is a text-based notation system that represents the structure of molecules in a simple and compact way.

In SMILES notation, each atom in the molecule is represented by its atomic symbol, and the bonds between the atoms are represented by different symbols, such as "-" for a single bond, "=" for a double bond, and "#" for a triple bond. For example, the SMILES notation for water (H2O) is "O-H-H", where "O" represents the oxygen atom and "H" represents the hydrogen atoms.

SMILES notation can represent more complex structures, such as cyclic and branched molecules. Rings in a molecule can be represented by enclosing the atoms in the ring with a set of numbers, and branches can be represented by using parentheses and numbers to indicate the position of the branches. For example, the SMILES notation for benzene (C6H6) is "c1ccccc1", where the lowercase "c" indicates an aromatic ring, and the SMILES notation for isobutane (C4H10) is "CC(C)C", where the parentheses indicate a branch [43].

2.1.4 InChi

InChI (International Chemical Identifier) is a standard text-based notation system that represents the structure of molecules in a unique and machine-readable format [46].

InChI notation contains several layers of information, including the connectivity of the atoms, the stereochemistry of the molecule, and the tautomeric and protonation states of the molecule. It also includes a fixed-length identifier called the InChIKey that uniquely identifies the molecule.

The InChI notation is designed to be both human-readable and machine-readable, making it useful in various applications, such as chemical informatics, chemical database management, and chemical property prediction. InChI notation can be generated from a molecular structure using software tools that implement the InChI algorithm, which generates a unique and canonical representation of the molecular structure.

InChI notation provides a standardized way of representing molecules that can be easily shared and communicated between different software tools and databases. It is also designed to be robust and flexible, allowing for the representation of a wide range of chemical structures. Many scientific journals and publications recognize InChI as the standard format for reporting chemical structures.

2.1.5 Molecule descriptors

Molecule descriptors play a vital role in cheminformatics and computational chemistry by providing numerical representations of chemical compounds. These descriptors encode a wide range of properties and characteristics of the molecules, enabling researchers to perform various tasks such as similarity analysis, virtual screening, and predictive modeling. RDKit, a popular cheminformatics toolkit, offers over 200 descriptors (see table 4.1) for any given molecule, providing a rich set of information for analysis [28].

Many descriptors in RDKit's second category often contain zero values for certain molecules (see table 4.1). This occurs because some molecular properties may not be applicable to all compounds, resulting in zero values for those specific descriptors. Despite these zero values, RDKit's extensive collection of descriptors still offers valuable insights into the chemical nature and behavior of molecules, facilitating a deeper understanding of their potential applications in drug discovery and other areas of research.

2.2 Databases and Datasets

- *Zinc*: The ZINC dataset is a vast collection of over 1 billion commercially available small molecules for drug discovery [47]. It includes diverse compounds and is regularly updated by the University of California, San Francisco. Researchers can search, filter, and predict molecular properties using this valuable resource. Specialized subsets for specific applications are also available.

- *BindingDB*: BindingDB is a comprehensive and publicly accessible database of experimentally determined binding data for protein-ligand complexes [48]. It contains over 1 million protein-ligand interactions involving 9,000 proteins and 360,000 small molecules. The data is manually curated from scientific literature, ensuring reliability and quality. Researchers can utilize various web-based tools and interfaces for searching, filtering, and analyzing the binding data, as well as predicting the binding affinity of new molecules using computational methods.

- *PubChem*: PubChem is a freely accessible public database maintained by the NCBI that contains information on millions of chemical substances and their biological activities [49]. It provides comprehensive data on various molecules, including chemical structures, molecular weights, physical and chemical properties, and pharmacology and toxicity information. Researchers can use PubChem to search, filter, and predict molecular properties, making it a valuable resource for drug discovery and development research. The database also offers specialized subsets for specific applications, catering to different research needs.

- *ChEMBL*: ChEMBL is a freely accessible database maintained by the EBI that contains information on over 2 million bioactive molecules and their targets, along with associated biological activity data [50]. The dataset includes detailed information on molecular structures, physicochemical properties, pharmacology, and toxicity. ChEMBL is a valuable resource for drug discovery research, as it allows researchers to search, filter, and predict molecular properties using computational methods. Additionally, the database offers specialized subsets for specific research needs, making it a user-friendly tool for drug development studies.

- *Protein Data Bank*: The Protein Data Bank (PDB) is a freely accessible database managed by the wwPDB that provides detailed 3D structural information on complex molecules like proteins and nucleic acids [51]. With over 170,000 structures, the PDB offers a comprehensive dataset of atomic coordinates, experimental data, and metadata. Researchers in structural biology and related fields can access this valuable resource to explore various molecular structures, ranging from small peptides to large protein complexes and viruses. The PDB facilitates easy search, filtering, and analysis of structural data, enabling researchers to study structure-activity relationships and identify potential drug targets.

2.3 Docking

Molecular docking is an important computational method utilized in drug discovery and molecular biology to predict and study the interactions between two molecules, typically a protein (receptor) and a small molecule (ligand) [52, 53]. Through simulations, it examines how these molecules bind together, providing insights into their structural characteristics and the strength of their interactions.

In drug discovery, molecular docking plays a vital role in understanding how potential drug candidates interact with target proteins at a molecular level. The primary objective is to find molecules that can selectively bind to specific target proteins and modify their function, either by inhibiting or enhancing their activity. This process is crucial in identifying promising drug candidates for further development and testing in experimental studies and clinical trials.

2.3.1 Searching functions

Molecular docking utilizes search functions to explore small molecule conformations for the best binding to a target protein. The goal is to find the most favorable orientation that minimizes energy and maximizes interactions with the protein.

Methods for exploring conformational space in molecular docking include:

- *Exhaustive search*: Evaluates the entire conformational space on a grid, but it's computationally intensive and time-consuming.

- *Monte Carlo search*: Randomly samples the conformational space and accepts or rejects moves based on energy, often used with simulated annealing to improve efficiency [56].

- *Genetic algorithm*: Evolves a population of conformations, selecting the best-performing ones for efficiency and focusing on promising conformations [55].

- *Fragment-based search*: Divides the small molecule into fragments, exploring their conformational space separately, and combining the best-performing conformations for the final pose [57].

2.3.2 Scoring functions

Molecular docking employs scoring functions to assess the suitability of small molecule binding to a target protein. These functions predict the binding affinity and identify the binding mode with the lowest energy. Various types of scoring functions are used in molecular docking:

- *Empirical Scoring Functions*: Based on experimental observations and use simple mathematical equations to calculate binding affinity [58].

- *Physics-Based Scoring Functions*: Grounded on physics principles, they use molecular mechanics to calculate interaction energy. [59]

- *Knowledge-Based Scoring Functions*: Founded on statistical analysis of known protein-ligand complexes, they use descriptors to predict binding affinity [60].

- *Machine Learning Scoring Functions*: Utilize machine learning algorithms like neural networks and support vector machines to learn the relationship between physicochemical properties and interaction energy [61].

2.4 Metrics and evaluation

2.4.1 Generative models' metrics

Generative models like GANs and VAEs are used in drug discovery to create new molecules, but assessing their usefulness is challenging. Researchers evaluate the generated molecules using various measures:

- *Chemical validity*: Ensuring the molecules adhere to organic chemistry rules and are stable, using tools like RDKit.

- *Diversity*: Checking that the generated molecules are distinct from each other, using clustering or similarity metrics like Tanimoto similarity.

- *Property optimization*: Assessing if the molecules have desired properties like binding affinity or solubility through virtual screening or experimental assays.

- *Novelty*: Verifying that the generated molecules are new and not already known in databases like ChEMBL, PubChem, or ZINC.

- *Scalability*: Ensuring the generative models can efficiently produce large numbers of molecules for practical applications like virtual screening.

2.4.2 DTA models' metric

DTA models predict how a drug molecule interacts with a target protein. Various measurements assess their effectiveness:

- *Binding affinity*: Indicates the strength of drug-target interaction, measured through methods like isothermal titration calorimetry or molecular dynamics simulations.

- *Cross-validation*: Evaluates model performance on different data subsets to assess generalizability and avoid overfitting.

- *ROC curve*: Plots sensitivity against 1-specificity for different binding affinities, assessing the model's ability to distinguish true positive and false positive bindings.

- *Precision-Recall curve*: Shows precision against recall for different cutoffs, useful when dealing with imbalanced data in positive and negative samples.

- *AUC-ROC*: Measures the model's ability to differentiate between interacting and non-interacting drug-target pairs.

- *AUC-PRC*: Measures the model's ability to retrieve positive interactions with high precision.

- *Accuracy*: Calculates the percentage of correctly predicted interactions out of the total interactions in the test set.

- *F1 score*: Balances precision and recall, providing a single metric for model performance ranging from 0 to 1, with higher values indicating better performance.

2.4.3 ADMET

ADMET is crucial in drug development, encompassing absorption, distribution, metabolism, excretion, and toxicity. Common metrics to evaluate ADMET properties of small molecules include:

- *Solubility*: Measured using techniques like shake-flask or HPLC methods, it influences drug bioavailability and dosage.

- *Permeability*: Assessed through in-vitro models like Caco-2 cells or PAMPA, it determines a molecule's ability to cross biological membranes.

- *Metabolic stability*: Evaluated through in-vitro or in-vivo models such as liver microsomes or rat pharmacokinetics, it gauges a molecule's durability in biological systems.

- *Plasma protein binding*: Assessed using in-vitro methods like equilibrium dialysis or ultrafiltration, it examines a molecule's attachment to plasma proteins affecting distribution and bioavailability.

- *CYP inhibition*: Determined by in-vitro assays like IC50 or Ki values, it assesses the inhibition of cytochrome P450 enzymes, which can cause drug-drug interactions.

- *Toxicity evaluations*: Measure potential harm using in vitro or in vivo models like cell viability assays or animal toxicology studies, focusing on hepatotoxicity, cardiotoxicity, and genotoxicity as common endpoints.

2.4.4 Workflow

In drug development, in-silico studies play a crucial role in narrowing down potential candidate drugs for costly and ethically complex in-vivo studies. To conduct these in-silico studies, researchers have access to various widely used libraries and tools, each serving specific purposes.

For tasks involving individual molecules and visualization, RDKit and OpenBabel are commonly employed by researchers. RDKit offers a rich set of functions and molecule descriptors, making it valuable for analyzing and characterizing molecules. Additionally, it allows easy conversion of 2D representations to other 2D formats, further enhancing its utility. Due to its effectiveness in performing molecular-level tasks and computing descriptors, RDKit has become popular among researchers for tasks like converting molecules and calculating molecule descriptors.

In in-silico studies, data-driven approaches extend beyond the functionalities of RDKit, leveraging libraries like PyTorch, pandas, NumPy [78], and TensorFlow [79] to implement machine Learning and deep Learning methodologies. De novo drug design, for instance, involves creating neural networks to learn patterns and compute a latent space from the trained model. This space is then used to navigate and generate new molecules, enabling the discovery of promising drug candidates.

Another application is in Drug-Target Interaction (DTA) prediction models. By training these models on datasets containing successful drug-target pairs, researchers can predict the binding affinity of existing drugs to new targets. This approach eliminates the need for time-consuming and costly clinical studies, streamlining the drug development process.

Once successful candidates are identified from the generated models, researchers conduct docking experiments to understand how these candidates interact with specific targets. This step is critical in studying the binding poses of ligands and calculating their binding affinities with receptors. Various tools, such as OpenBabel, PyMol, AutoDock Vina, and SWISS-MODEL, come into play during this process. OpenBabel is essential for converting ligand 2D shapes into 3D structures, while SWISS-MODEL generates consistent 3D protein models. AutoDock Vina handles the molecular docking, and PyMol aids in the visualization of docking results.

By following this workflow, researchers streamline the drug candidate selection process, leading to improved ADMET evaluation and increasing the likelihood of successful drug development and registration. These data-driven approaches offer valuable insights and advancements in drug discovery, potentially revolutionizing the pharmaceutical industry.

2.4.5 Benchmarks

Benchmarks play a crucial role in evaluating the effectiveness of generative models by providing standardized datasets and evaluation protocols. In de novo drug design, two widely recognized and publicly accessible benchmarks are MOSES and Guacamole. Among them, MOSES is more popular due to its user-friendly interface and available metric computing functions. On the other hand, Guacamole is less favored mainly because of its lack of clear documentation and user-friendliness.

Within the MOSES benchmark [64], several metrics are utilized to assess generative models. These include distribution difference metrics, internal diversity metrics, novelty, validity, uniqueness, filters, FCD (Frechet ChemNet Distance) [63], Tanimoto distance, scaffold similarity, and Fréchet ChemNet Distance. Each of these metrics provides valuable insights into the model's performance and the quality of generated molecules.

In Drug-Target Interaction (DTA) prediction, researchers can access diverse datasets like DREAM [65] and KIBA [66]. Additionally, publicly available cheminformatics datasets can be used to construct suitable datasets for DTA models. A crucial metric in DTA prediction is the binding affinity, which quantifies the strength of interaction between a drug and a target protein. Accurate binding affinity predictions are vital for identifying potential drug candidates with strong binding capabilities to specific target proteins. Furthermore, binary classifications are often applied in DTA prediction models to categorize drug-target interactions as either positive or negative.

Chapter 3

Related Works

DeepChem [67], ChemML [68], OpenChem [69], Chainer Chemistry, and TorchDrug [70] are all powerful tools and libraries in the field of cheminformatics and drug discovery, each with its unique features and capabilities.

- *DeepChem*: DeepChem [67] stands out as a versatile Python library specifically designed for deep learning in chemistry and materials science. It offers a wide range of tools and models to predict molecular properties, conduct virtual screening of small molecule libraries, and handle molecular data efficiently.

- *ChemML*: ChemML [68], another Python library, specializes in machine learning applications for chemistry and materials science. With a focus on molecular and materials property prediction, it provides robust data handling tools and pre-trained machine learning models, enabling accurate predictions for various chemical and material properties.

- *OpenChem*: OpenChem [69] is a comprehensive platform offering free and open-source access to deep learning tools tailored for chemistry and cheminformatics. With molecular data handling capabilities for fingerprints, SMILES strings, and molecular graphs, OpenChem's deep learning models deliver accurate predictions of molecular properties like solubility, toxicity, and binding affinity. Additionally, it excels in molecular generation and design, utilizing generative models, reinforcement learning, genetic algorithms, and Bayesian optimization.

- *Chainer Chemistry*: Chainer Chemistry is a powerful framework built upon Chainer, a popular deep learning library. Chainer Chemistry extends Chainer's capabilities with a focus on chemical informatics tasks, including molecular property prediction, molecular generation, and virtual screening. Its seamless integration with Chainer allows users to leverage the extensive deep learning functionalities of both libraries for cheminformatics applications.

- *TorchDrug*: TorchDrug [70] is a specialized library built on top of PyTorch, a widely used deep learning framework. It offers tailored tools and models for drug discovery and cheminformatics, empowering users to predict molecular properties, generate new molecules, and optimize molecules for specific properties. With PyTorch as its foundation, TorchDrug provides an accessible and efficient platform for deep learning in chemical research.

Chapter 4

Framework

4.1 Problem definition

The staggered development of software tools has led to various challenges in onboarding new members to de novo drug design. As a researcher, newcomers encounter difficulties with software installations, inadequate access to computational resources, slow network connections, software version conflicts, and other issues that are inherent to decentralized management. Additionally, a lack of knowledge about existing approaches and solutions further hinders their ability to freely embark on development.

The main problem is the absence of a straightforward and comprehensive framework that covers all aspects of in-silico studies in drug design. Such a framework should explicitly outline the fundamental principles and systematic approaches for developing drug design workflows. It should consist of modular components that guide users through the drug development process, ensuring a clear understanding of their objectives. While some researchers may excel in computer science or data science, they might lack expertise in chemistry or biology, which are not crucial skills in data-driven drug development. As a result, the framework must be user-friendly and accessible to individuals without a background in chemistry or biology, as de novo drug design primarily involves discovering new molecules through data-driven techniques.

4.2 Design

I have created a framework that offers essential and valuable functionalities for drug design. My solution revolves around the development of a centralized platform and a helpful tool to assist in crucial tasks, including computing metrics, docking ligands to receptors, downloading datasets, and calculating molecule descriptors for various drug development studies. The framework comprises three key components: a server, a library for local computations, and a library for remote computations. Utilizing remote computations and a centralized platform addresses several gaps in drug design, proving to be promising solutions. Notably, implementing a leaderboard on the centralized server holds tremendous potential to significantly benefit drug design.

The framework is guided by the following core principles:

- Encompassing all in-silico workflows within a single tool, ensuring comprehensive coverage.

- Eliminating computational limitations for researchers, enabling seamless and efficient work.

- Embracing an open-source approach, making the framework accessible to everyone.

- Allowing customization for individuals who require a private server for team collaboration.

- Simplifying metric computations, providing commonly used metrics for researchers' convenience.

- Facilitating the submission of successful candidate drugs to a centralized server, allowing interested parties to study and contribute reports.

- Providing a diverse range of benchmark metrics and datasets to support robust evaluations.

- Sharing knowledge about the entire drug design workflow through practical examples.
- Prioritizing user-friendliness to ensure ease of use for all individuals utilizing the framework.

The significance of a centralized server lies in its ability to accommodate bi-weekly changing datasets and facilitate the reevaluation of existing approaches and models. By maintaining the latest dataset updates and conducting reevaluations through a leaderboard, the centralized server becomes a valuable resource for studying various models and approaches effectively.

Given time constraints, I have successfully accomplished the following functionalities: local computation of MOSES metrics, remote computation of MOSES metrics, a torch data loader for the MOSES dataset, approximately 200 molecule descriptors (see table 4.1), a centralized server for data management, local docking capability, remote docking, and caching.

4.3 Modules

The framework contains three main modules for fulfilling research needs (see figure 4-1).

- *The DDBox Data Server*: The DDBox Data Server is a centralized server specifically created to store data and handle remote tasks, such as metric computation and docking (see figure 4-2). Within this server, various services are integrated, including PostgreSQL for the storage of molecules and receptors, Redis for optimization functionalities, Celery for background computations like docking and metric calculation, and FastAPI as the primary application server to process user requests. As per its philosophical design, this module is intended to include a leaderboard, support multiple datasets, host various docking software, serve as a platform for enthusiastic researchers, act as a bridge to in-vitro studies, and facilitate the generation of reports.

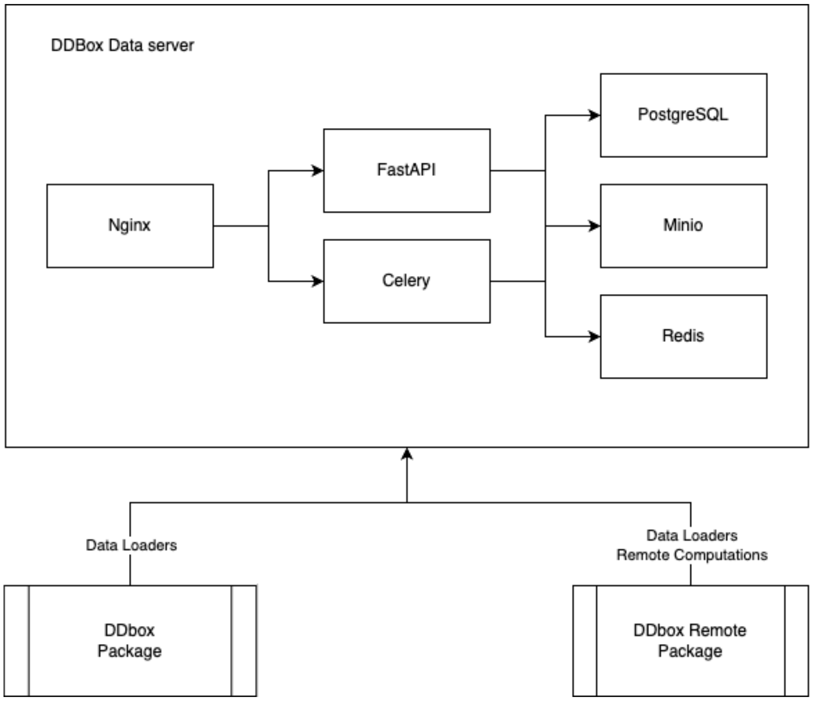
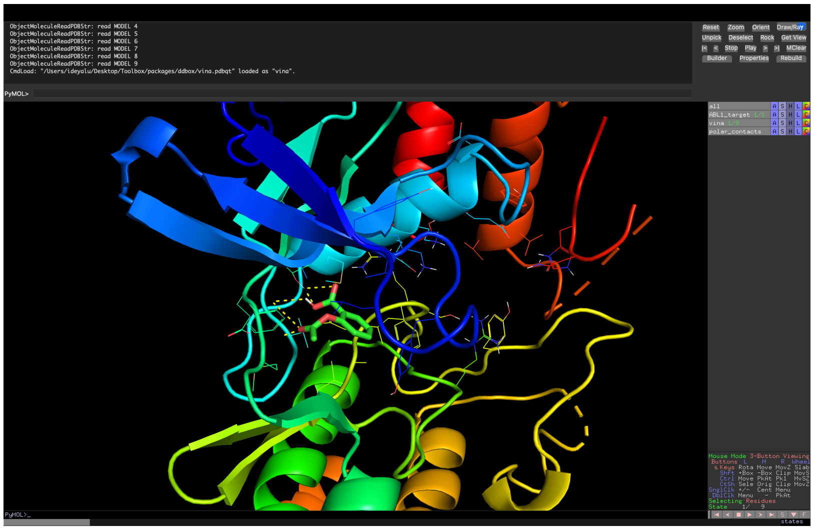


Figure 4-1: DDBox Architecture

- *The DDBox Package*: The DDBox Package is an all-encompassing framework that includes data loaders, metrics, and docking software (see figure 4-2), making it well-suited for conducting local computations. This package is beneficial for local studies, but it might necessitate computational resources that are not readily available at all times. As per its philosophical design, this module is intended to feature visualizations, multiple dataset loaders, information retrievals, various docking software, and a framework for De Novo generative and DTA models.

- *The DDBox Remote Package*: The DDBox Remote Package functions as an interface framework tailored for remote computations. It provides similar functionalities to the DDBox Package, but the actual computations are performed on the server side, potentially resulting in improved work performance. As per its philosophical design, this module is intended to be similar to the DDBox Package, with the distinction that it should be seamlessly integrated with the DDBox Data Server to carry out remote computations.



```
ObjectMoleculeReadPDBStr: read MODEL 4
ObjectMoleculeReadPDBStr: read MODEL 5
ObjectMoleculeReadPDBStr: read MODEL 6
ObjectMoleculeReadPDBStr: read MODEL 7
ObjectMoleculeReadPDBStr: read MODEL 8
ObjectMoleculeReadPDBStr: read MODEL 9
CmdLoad: "/Users/ideyalu/Desktop/Toolbox/packages/ddbox/vina.pdbqt" loaded as "vina".
PyMOL>
```

```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Wheel
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl Move PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
SnglClk +/-  Cent Menu
 DblClk Menu  -   PkAt
Selecting Residues
State     1/   9
```

Figure 4-2: AutoDock Vina result and PyMol visualization.

4.4 Functionalities

The DDBox data server provides around 200 descriptors for molecules in the MOSES dataset. These descriptors serve various purposes, such as data analysis, training machine learning and deep learning models, and conducting diverse studies. However, it's important to note that many molecules have zero values for the 2nd category descriptors.

You can extract InChi, InChi Key, and SMILES representations from a molecule object using DDBox Package. While both RDKit and DDBox Package have the same descriptors, using only RDKit requires additional time to write conversion logic and leads you to wait for the conversion process to be completed, which typically takes about 4-8 hours. To address this issue, the DDBox data server has already computed these values on the server side. As a result, you can simply download the computed descriptors using the torch dataset class implementation in DDBox Package, saving you time and effort in the computation process.

The DDBox Data server contains the MOSES dataset, which includes molecules with pre-calculated descriptors. Researchers can easily access this dataset by downloading it through the torch dataset class implementation using either DDBox or DDBox Remote. The download process is made efficient by utilizing cached API endpoints on the server side.

In the event that new data or changes are made on the DDBox Data Server, users must delete the local cache and re-initialize the torch dataset instance to ensure they obtain the latest updates. The framework is designed cache API endpoints for some time. When the cache is expired, new data is downloaded.

To enhance the framework's capabilities, it should include information retrieval APIs from various public archives such as PDB, ChEMBL, PubChem, DrugBank, and more. This information retrieval integration will allow researchers to access and utilize additional relevant information for their analyses and studies.

Within the DDBox Package, you can find the MOSES metrics, which are used to evaluate De Novo generative models. I consider the MOSES dataset as a strong foundation for the framework development due to its popularity. Nevertheless, the framework's overarching design philosophy should encompass the incorporation of multiple benchmark metrics.

Remote computation is a highly valuable feature that provides researchers with the freedom to overcome resource limitations. A significant advantage of remote computation is its application during model training. As the evaluation process may require several hours, researchers can avoid wasting time by conducting evaluations asynchronously while continuously training the model. This approach ensures an efficient use of time and computing resources.

4.5 Hardware Requirements

The DDBox data server is built using Docker and Docker Compose. Each container takes up around 250MB of RAM space, and the stack comprises five containers: PostgreSQL, Redis, Minio, Celery, and FastAPI. Except for Redis, each of the four containers uses approximately 250MB of RAM space, and there is potential for further optimization. The DDBox package and DDBox Data Server demand a powerful CPU, such as M1 or CUDA support. On M1 Linux/amd64 architecture, docking typically requires around 2 minutes, while on M1 arm64 architecture, it takes only about 5 seconds. Most metric computations are executed on the CPU, necessitating the use of a robust CPU. Only the FCD distance metric computation requires a GPU instead of a CPU. To achieve optimal performance and faster metric computation, having both a powerful CPU and GPU support is advantageous. However, without a powerful CPU or GPU, computations may take up to 1 hour or even a day. Allocating more than 8GB-16GB of RAM for DDBox Package and DDBox Data Server is advisable. Additionally, you may need 8GB-16GB of hard disk space for DDBox Package caches, which store downloaded datasets with numerous attributes, and 128GB-256GB of hard disk space for DDBox Data Server due to caching optimizations. This can sometimes pose challenges due to resource limitations for some users.

Table 4.1: List of Molecule Descriptors

InChi	InChi Key	SMILES
BalabanJ	BertzCT	Chi0
Chi0n	Chi0v	Chi1
Chi1n	Chi1v	Chi2n
Chi2v	Chi3n	Chi3v
Chi4n	Chi4v	EState_VSA1
EState_VSA10	EState_VSA11	EState_VSA2
EState_VSA3	EState_VSA4	EState_VSA5
EState_VSA6	EState_VSA7	EState_VSA8
EState_VSA9	ExactMolWt	FpDensityMorgan1
FpDensityMorgan2	FpDensityMorgan3	FractionCSP3
HallKierAlpha	HeavyAtomCount	HeavyAtomMolWt
Ipc	Kappa1	Kappa2
Kappa3	LabuteASA	MaxAbsEStateIndex
MaxAbsPartialCharge	MaxEStateIndex	MaxPartialCharge
MinAbsEStateIndex	MinAbsPartialCharge	MinEStateIndex
MinPartialCharge	MolLogP	MolMR
MolWt	NHOHCount	NOCount
NumAliphaticCarbocycles	NumAliphaticHeterocycles	NumAliphaticRings
NumAromaticCarbocycles	NumAromaticHeterocycles	NumAromaticRings
NumHAcceptors	NumHDonors	NumHeteroatoms
NumRadicalElectrons	NumRotatableBonds	NumSaturatedCarbocycles
NumSaturatedHeterocycles	NumSaturatedRings	NumValenceElectrons
PEOE_VSA1	PEOE_VSA10	PEOE_VSA11
PEOE_VSA12	PEOE_VSA13	PEOE_VSA14
PEOE_VSA2	PEOE_VSA3	PEOE_VSA4
PEOE_VSA5	PEOE_VSA6	PEOE_VSA7
PEOE_VSA8	PEOE_VSA9	RingCount
SMR_VSA1	SMR_VSA10	SMR_VSA2
SMR_VSA3	SMR_VSA4	SMR_VSA5
SMR_VSA6	SMR_VSA7	SMR_VSA8
SMR_VSA9	SlogP_VSA1	SlogP_VSA10
SlogP_VSA11	SlogP_VSA12	SlogP_VSA2
SlogP_VSA3	SlogP_VSA4	SlogP_VSA5
SlogP_VSA6	SlogP_VSA7	SlogP_VSA8
SlogP_VSA9	TPSA	VSA_EState1
VSA_EState10	VSA_EState2	VSA_EState3
VSA_EState4	VSA_EState5	VSA_EState6
VSA_EState7	VSA_EState8	VSA_EState9
fr_Al_COO	fr_Al_OH	fr_Al_OH_noTert
fr_ArN	fr_Ar_COO	fr_Ar_N
fr_Ar_NH	fr_Ar_OH	fr_COO
fr_COO2	fr_C_O	fr_C_O_noCOO
fr_C_S	fr_HOCCN	fr_Imine
fr_NH0	fr_NH1	fr_NH2
fr_N_O	fr_Ndealkylation1	fr_Ndealkylation2
fr_Nhpyrrole	fr_SH	fr_aldehyde
fr_alkyl_carbamate	fr_alkyl_halide	fr_allylic_oxid
fr_amide	fr_amidine	fr_aniline
fr_aryl_methyl	fr_azide	fr_azo
fr_barbitur	fr_benzene	fr_benzodiazepine
fr_bicyclic	fr_diazo	fr_dihydropyridine
fr_epoxide	fr_ester	fr_ether
fr_furan	fr_guanido	fr_halogen
fr_hdrzine	fr_hdrzone	fr_imidazole
fr_imide	fr_isocyan	fr_isothiocyan
fr_ketone	fr_ketone_Topliss	fr_lactam
fr_lactone	fr_methoxy	fr_morpholine
fr_nitrile	fr_nitro	fr_nitro_arom
fr_nitro_arom_nonortho	fr_nitroso	fr_oxazole
fr_oxime	fr_para_hydroxylation	fr_phenol
fr_phenol_noOrthoHbond	fr_phos_acid	fr_phos_ester
fr_piperdine	fr_piperzine	fr_priamide
fr_prisulfonamd	fr_pyridine	fr_quatN
fr_sulfide	fr_sulfonamd	fr_sulfone
fr_term_acetylene	fr_tetrazole	fr_thiazole
fr_thiocyan	fr_thiophene	fr_unbrch_alkane
fr_urea	qed	

Chapter 5

Future Works

The framework must possess scalability, capable of accommodating a large number of users. To handle high loads, the implementation involves utilizing solutions such as a database cluster and caching. Achieving scalability is made possible through the use of a storage cluster (Minio), database cluster (PostgreSQL), broker cluster (Redis), and horizontal scaling of the server application. Computational power is harnessed through the M1 processor, and it can be configured to support CUDA, potentially leading to excellent performance.

The incorporation of visualization tools into the framework holds the potential to enhance its utility in In-silico studies, providing researchers with interpretable data. Possible integrations of visualization tools encompass molecule visualization, receptor visualization, docking visualization, and more.

A centralized server is intended to serve as a platform for potential drug candidates, which will be queued for in-vitro studies. Positioned as an intermediary element within the drug design process, researchers will be able to select ligands that have successfully passed multiple in-silico studies for further in-vitro investigations. Moreover, the reports generated by observers will undergo validation by authorized members. Developing a centralized system for drug design represents a significant and highly valuable study, addressing the current lack of a centralized platform and standardized workflow that incorporates an accessible set of tools and resources. Presently, these tools are scattered across various workflows, with many complementing each other, making it advantageous to consolidate them into a single, user-friendly framework.

The inclusion of a leaderboard is a highly promising and essential feature of the framework, significantly increasing its potential to become a standard platform in the field of drug discovery. However, the development of this leaderboard poses challenges in terms of system design and handling special cases, such as dataset updates, recomputation of benchmark metrics, model submission, versioning, and management. A well-designed centralized system holds the potential to add substantial value to drug development. At present, the framework offers a basic set of tools that meet the fundamental requirements for in-silico drug design research and workflow. It is a work in progress, with incremental improvements expected. The success of the framework will be determined by monitoring its usage statistics and popularity among engineers.

The framework currently has several internal functions that lack optimization and are not generalized for different platforms, except for Linux and MacOS. Unfortunately, Windows support is not yet included in the framework.

Although certain libraries under the framework have default CUDA support, the framework itself requires explicit support for CUDA. For the MVP, parallel computations have been disabled, but future optimization efforts are planned to include parallel computations and improve computation speeds. I see these optimization tasks as the next development steps for the framework.

Chapter 6

Conslusion

I have dedicated approximately 60-100 hours to working on integrations, exploring implementation methods, and extensively studying drug discovery articles. I am confident that newcomers may have to go through the same learning process or become acquainted with it. The efforts I have invested can potentially save many hours of future research and aid engineers and scientists in swiftly integrating themselves into the Drug Discovery domain.

The tool offers datasets, metrics, docking capabilities, and helper functions to facilitate seamless and efficient work, enabling a rapid start. By introducing this platform and providing the framework, it has the potential to make a substantial contribution to the domain. It effectively addresses various challenges related to dataset updates, dataset downloads, benchmarking platforms, leaderboards, remote computations, and resource limitations. The framework's versatile applications encompass molecular docking, data analysis, machine learning model construction, deep learning model development, statistical analysis, and more.

Adopting an open-source approach for the framework's development could bring significant advantages to both the project's future and the research domain. By embracing an open-source model, the framework can undergo continuous enhancement and incremental improvements, overseen and managed by the community itself. Community-driven management ensures a consistent evolution of the framework, fostering swift advancements that have the potential to make valuable contributions to the entire Drug Design domain. Ultimately, this collaborative effort aims to establish the framework as a standard tool for enthusiastic researchers.

The framework currently incorporates solely MOSES benchmark metrics and datasets. However, there are additional benchmarks like Guacamole, as well as diverse datasets used in Drug Design. In the future, the framework should aim to support all popular datasets and benchmarks. A challenging aspect of the framework lies in handling dataset changes, which necessitate re-evaluating submitted models, incurring potential costs. To address this issue, a viable solution is to design a flow that includes dataset updates as a platform feature. This entails recalculating the leaderboard whenever dataset updates occur. At present, utilizing static data from the MOSES benchmark serves as a suitable starting point due to its popularity and simplicity.

The current version of the framework includes datasets, docking, metric computations, and molecule descriptors, which form the essential components. However, there are additional features that should be developed, such as visualization tools, configuration editing, support for various metrics, benchmarks, and platforms, integration of other docking software, incorporation of diverse datasets, leaderboard functionality, and scalability, among others. The overarching goal of the framework is to encompass and streamline all workflows in Drug Design, serving as an intermediary platform that connects data-driven engineering to in-vitro studies.

Considering my personal objectives, I strongly believe that the framework holds significant value for engineers involved in drug design.

Bibliography

Bibliography

[1] Pirintsos, S., Panagiotopoulos, A., Bariotakis, M., Daskalakis, V., Lionis, C., Sourvinos, G., Karakasiliotis, I., Kampa, M. and Castanas, E., 2022. From Traditional Ethnopharmacology to Modern Natural Drug Discovery: A Methodology Discussion and Specific Examples. Molecules, 27(13), p.4060.

[2] Meyers, J., Fabian, B. and Brown, N., 2021. De novo molecular design and generative models. Drug Discovery Today, 26(11), pp.2707-2715.

[3] Do, Q.T., Renimel, I., Andre, P., Lugnier, C., Muller, C.D. and Bernard, P., 2005. Reverse pharmacognosy: application of Selnergy, a new tool for lead discovery. The example of ϵ-viniferin. Current drug discovery technologies, 2(3), pp.161-167.

[4] Grisoni, F. and Schneider, G., 2019. De novo molecular design with generative long short-term memory. CHIMIA International Journal for Chemistry, 73(12), pp.1006-1011.

[5] Abdelkrim, A., Bouramoul, A. and Zenbout, I., 2021, December. Machine Learning Methods In Drug Discovery: A Selective Review. In 2021 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS) (pp. 1-8). IEEE.

[6] Marcilin, L.J.A., Sheeba, I.R., Sugadev, M., Velan, B. and Chitra, P., 2021, March. Identification of Drug Discovery for Patients Using Machine Learning. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 274-278). IEEE.

[7] Biswas, R., Basu, A., Nandy, A., Deb, A., Haque, K. and Chanda, D., 2020, February. Drug Discovery and Drug Identification using AI. In 2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN) (pp. 49-51). IEEE.

[8] Qian, Y., Xing, Y. and Dong, L., 2021, March. Deep learning for a low-data drug design system. In 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM) (pp. 1-4). IEEE.

[9] Lee, D., 2020, December. Accelerating Drug Discovery with an AI-Based Virtual Human System CODA. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 3-3). IEEE.

[10] Floresta, G., Zagni, C., Gentile, D., Patamia, V. and Rescifina, A., 2022. Artificial Intelligence Technologies for COVID-19 De Novo Drug Design. International Journal of Molecular Sciences, 23(6), p.3261.

[11] Padalkar, G.R., Patil, S.D., Hegadi, M.M. and Jaybhaye, N.K., 2021, January. Drug discovery using generative adversarial network with reinforcement learning. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-3). IEEE.

[12] Lin, E., Lin, C.H. and Lane, H.Y., 2020. Relevant applications of generative adversarial networks in drug design and discovery: molecular de novo design, dimensionality reduction, and de novo peptide and protein design. Molecules, 25(14), p.3250.

[13] Blanchard, A.E., Stanley, C. and Bhowmik, D., 2021. Using GANs with adaptive training data to search for new molecules. Journal of cheminformatics, 13(1), pp.1-8.

[14] Schneider, H.C. and Klabunde, T., 2013. Understanding drugs and diseases by systems biology?. Bioorganic & medicinal chemistry letters, 23(5), pp.1168-1176.

[15] Butcher, E.C., Berg, E.L. and Kunkel, E.J., 2004. Systems biology in drug discovery. Nature biotechnology, 22(10), pp.1253-1259.

[16] Jamshidi, M.B., Talla, J., Lalbakhsh, A., Sharifi-Atashgah, M.S., Sabet, A. and Peroutka, Z., 2021, December. A Conceptual Deep Learning Framework for COVID-19 Drug Discovery. In 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 00030-00034). IEEE.

[17] Fakhraei, S., Huang, B., Raschid, L. and Getoor, L., 2014. Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(5), pp.775-787.

[18] Peng, L., Liao, B., Zhu, W., Li, Z. and Li, K., 2015. Predicting drug–target interactions with multi-information fusion. IEEE journal of biomedical and health informatics, 21(2), pp.561-572.

[19] Ban, T., Ohue, M. and Akiyama, Y., 2017, October. Efficient hyperparameter optimization by using Bayesian optimization for drug-target interaction prediction. In 2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) (pp. 1-6). IEEE.

[20] Rezaei, M.A., Li, Y., Wu, D., Li, X. and Li, C., 2020. Deep learning in drug design: protein-ligand binding affinity prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[21] Pandey, M., Radaeva, M., Mslati, H., Garland, O., Fernandez, M., Ester, M. and Cherkasov, A., 2022. Ligand Binding Prediction using Protein Structure Graphs and Residual Graph Attention Networks. bioRxiv.

[22] Ahmed, A., Mam, B. and Sowdhamini, R., 2021. DEELIG: A deep learning approach to predict protein-ligand binding affinity. Bioinformatics and Biology Insights, 15, p.11779322211030364.

[23] Özçelik, R., Öztürk, H., Özgür, A. and Ozkirimli, E., 2021. Chemboost: A chemical language based approach for protein–ligand binding affinity prediction. Molecular Informatics, 40(5), p.2000212.

[24] Nguyen, T., Le, H. and Venkatesh, S., 2019. GraphDTA: prediction of drug–target binding affinity using graph convolutional networks. BioRxiv, p.684662.

[25] Paggi, J.M., Belk, J.A., Hollingsworth, S.A., Villanueva, N., Powers, A.S., Clark, M.J., Chemparathy, A.G., Tynan, J.E., Lau, T.K., Sunahara, R.K. and Dror, R.O., 2021. Leveraging nonstructural data to predict structures and affinities of protein–ligand complexes. Proceedings of the National Academy of Sciences, 118(51), p.e2112621118.

[26] Kraljevic, S., Stambrook, P.J. and Pavelic, K., 2004. Accelerating drug discovery: Although the evolution of '-omics' methodologies is still in its infancy, both the pharmaceutical industry and patients could benefit from their implementation in the drug development process. EMBO reports, 5(9), pp.837-842.

[27] Hughes, J.P., Rees, S., Kalindjian, S.B. and Philpott, K.L., 2011. Principles of early drug discovery. British journal of pharmacology, 162(6), pp.1239-1249.

[28] Landrum, G., 2006. RDKit: Open-source cheminformatics.

[29] O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. Journal of cheminformatics, 3(1), pp.1-14.

[30] Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2), pp.455-461.

[31] Seeliger, D. and de Groot, B.L., 2010. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. Journal of computer-aided molecular design, 24(5), pp.417-422.

[32] Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C., 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic acids research, 31(13), pp.3381-3385.

[33] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. and Blaschke, T., 2018. The rise of deep learning in drug discovery. Drug discovery today, 23(6), pp.1241-1250..

[34] DeLano, W.L., 2002. Pymol: An open-source molecular graphics tool. CCP4 Newsl. Protein Crystallogr, 40(1), pp.82-92.

[35] Wang, M.W., Hao, X. and Chen, K., 2007. Biological screening of natural products and drug innovation in China. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1482), pp.1093-1105.

[36] Nogrady, T. and Weaver, D.F., 2005. Medicinal chemistry: a molecular and biochemical approach. Oxford University Press.

[37] Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Côté, S. and Shoichet, B.K., 2012. Large-scale prediction and testing of drug activity on side-effect targets. Nature, 486(7403), pp.361-367.

[38] Cassar, S., Adatto, I., Freeman, J.L., Gamse, J.T., Iturria, I., Lawrence, C., Muriana, A., Peterson, R.T., Van Cruchten, S. and Zon, L.I., 2019. Use of zebrafish in drug discovery toxicology. Chemical research in toxicology, 33(1), pp.95-118.

[39] Brown, N., Fiscato, M., Segler, M.H. and Vaucher, A.C., 2019. GuacaMol: benchmarking models for de novo molecular design. Journal of chemical information and modeling, 59(3), pp.1096-1108.

[40] Brown, N., Fiscato, M., Segler, M.H. and Vaucher, A.C., 2019. GuacaMol: benchmarking models for de novo molecular design. Journal of chemical information and modeling, 59(3), pp.1096-1108.

[41] Ofer, D., Brandes, N. and Linial, M., 2021. The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal, 19, pp.1750-1758.

[42] David, L., Thakkar, A., Mercado, R. and Engkvist, O., 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. Journal of Cheminformatics, 12(1), pp.1-22.

[43] Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 28(1), pp.31-36.

[44] Zhang, Y., Jin, R. and Zhou, Z.H., 2010. Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1, pp.43-52.

[45] Yi, H.C., You, Z.H., Huang, D.S. and Kwoh, C.K., 2022. Graph representation learning in bioinformatics: trends, methods and applications. Briefings in Bioinformatics, 23(1), p.bbab340.

[46] Heller, S.R., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D., 2015. InChI, the IUPAC international chemical identifier. Journal of cheminformatics, 7(1), pp.1-34.

[47] Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S. and Coleman, R.G., 2012. ZINC: a free tool to discover chemistry for biology. Journal of chemical information and modeling, 52(7), pp.1757-1768.

[48] Liu, T., Lin, Y., Wen, X., Jorissen, R.N. and Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic acids research, 35(suppl_1), pp.D198-D201.

[49] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B. and Zaslavsky, L., 2019. PubChem 2019 update: improved access to chemical data. Nucleic acids research, 47(D1), pp.D1102-D1109.

[50] Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J.P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research, 40(D1), pp.D1100-D1107.

[51] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. Nucleic acids research, 28(1), pp.235-242.

[52] Pagadala, N.S., Syed, K. and Tuszynski, J., 2017. Software for molecular docking: a review. Biophysical reviews, 9, pp.91-102.

[53] Dias, R., de Azevedo, J. and Walter, F., 2008. Molecular docking algorithms. Current drug targets, 9(12), pp.1040-1047.

[54] Thomsen, R. and Christensen, M.H., 2006. MolDock: a new technique for high-accuracy molecular docking. Journal of medicinal chemistry, 49(11), pp.3315-3321.

[55] Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., 1997. Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology, 267(3), pp.727-748.

[56] Liu, M. and Wang, S., 1999. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. Journal of computer-aided molecular design, 13, pp.435-451.

[57] Budin, N., Majeux, N. and Caflisch, A., 2001. Fragment-based flexible ligand docking by evolutionary optimization.

[58] Korb, O., Stutzle, T. and Exner, T.E., 2009. Empirical scoring functions for advanced proteinligand docking with PLANTS. Journal of chemical information and modeling, 49(1), pp.84-96.

[59] Li, J., Fu, A. and Zhang, L., 2019. An overview of scoring functions used for protein–ligand interactions in molecular docking. Interdisciplinary Sciences: Computational Life Sciences, 11, pp.320-328.

[60] Huang, S.Y., Grinter, S.Z. and Zou, X., 2010. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Physical Chemistry Chemical Physics, 12(40), pp.12899-12908.

[61] Kinnings, S.L., Liu, N., Tonge, P.J., Jackson, R.M., Xie, L. and Bourne, P.E., 2011. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. Journal of chemical information and modeling, 51(2), pp.408-419.

[62] Van De Waterbeemd, H. and Gifford, E., 2003. ADMET in silico modelling: towards prediction paradise?. Nature reviews Drug discovery, 2(3), pp.192-204.

[63] Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. and Klambauer, G., 2018. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. Journal of chemical information and modeling, 58(9), pp.1736-1741.

[64] Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M. and Kadurin, A., 2020. Molecular sets (MOSES): a benchmarking platform for molecular generation models. Frontiers in pharmacology, 11, p.565644.

[65] Li, S., Li, L., Meng, X., Sun, P., Liu, Y., Song, Y., Zhang, S., Jiang, C., Cai, J. and Zhao, Z., 2021. DREAM: a database of experimentally supported protein-coding RNAs and drug associations in human cancer. Molecular Cancer, 20, pp.1-6.

[66] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K. and Aittokallio, T., 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. Journal of Chemical Information and Modeling, 54(3), pp.735-743.

[67] Ramsundar, B., Eastman, P., Walters, P. and Pande, V., 2019. Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more. " O'Reilly Media, Inc.".

[68] Haghighatlari, M., Vishwakarma, G., Altarawy, D., Subramanian, R., Kota, B.U., Sonpal, A., Setlur, S. and Hachmann, J., 2020. ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(4), p.e1458.

[69] Korshunova, M., Ginsburg, B., Tropsha, A. and Isayev, O., 2021. OpenChem: a deep learning toolkit for computational chemistry and drug design. Journal of Chemical Information and Modeling, 61(1), pp.7-13.

[70] Zhu, Z., Shi, C., Zhang, Z., Liu, S., Xu, M., Yuan, X., Zhang, Y., Chen, J., Cai, H., Lu, J. and Ma, C., 2022. Torchdrug: A powerful and flexible machine learning platform for drug discovery. arXiv preprint arXiv:2202.08320.

[71] Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. and Jensen, K.F., 2022. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science, 12(5), p.e1608.

[72] Yu, W. and MacKerell, A.D., 2017. Computer-aided drug design methods. Antibiotics: methods and protocols, pp.85-106.

[73] Wadood, A., Ahmed, N., Shah, L., Ahmad, A., Hassan, H. and Shams, S., 2013. In-silico drug design: An approach which revolutionarised the drug discovery process. OA Drug Des Deliv, 1(1), p.3.

[74] Csermely, P., Agoston, V. and Pongor, S., 2005. The efficiency of multi-target drugs: the network approach might help drug design. Trends in pharmacological sciences, 26(4), pp.178-182.

[75] Kim, H., Kim, E., Lee, I., Bae, B., Park, M. and Nam, H., 2020. Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. Biotechnology and Bioprocess Engineering, 25, pp.895-930.

[76] Xue, D., Gong, Y., Yang, Z., Chuai, G., Qu, S., Shen, A., Yu, J. and Liu, Q., 2019. Advances and challenges in deep generative models for de novo molecule

generation. Wiley Interdisciplinary Reviews: Computational Molecular Science, 9(3), p.e1395.

[77] Lin, E., Lin, C.H. and Lane, H.Y., 2020. Relevant applications of generative adversarial networks in drug design and discovery: molecular de novo design, dimensionality reduction, and de novo peptide and protein design. Molecules, 25(14), p.3250.

[78] Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J. and Kern, R., 2020. Array programming with NumPy. Nature, 585(7825), pp.357-362.

[79] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., 2016. TensorFlow: a system for Large-Scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283).

[80] Mizutani, S., Pauwels, E., Stoven, V., Goto, S. and Yamanishi, Y., 2012. Relating drug–protein interaction network with drug side effects. Bioinformatics, 28(18), pp.i522-i528.

[81] Lionta, E., Spyrou, G., K Vassilatis, D. and Cournia, Z., 2014. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current topics in medicinal chemistry, 14(16), pp.1923-1938.

[82] Anighoro, A., Bajorath, J. and Rastelli, G., 2014. Polypharmacology: challenges and opportunities in drug discovery: miniperspective. Journal of medicinal chemistry, 57(19), pp.7874-7887.

[83] Törnqvist, E., Annas, A., Granath, B., Jalkesten, E., Cotgreave, I. and Öberg, M., 2014. Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. PloS one, 9(7), p.e101638.

[84] Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L. and McLaren, S.,

2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature, 496(7446), pp.498-503.

[85] Zhao, L., Ciallella, H.L., Aleksunes, L.M. and Zhu, H., 2020. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug discovery today, 25(9), pp.1624-1638.

[86] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K. and Kumar, P., 2021. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Molecular diversity, 25, pp.1315-1360.

[87] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

[88] Zhumagambetov, R., Molnár, F., Peshkov, V.A. and Fazli, S., 2021. Transmol: repurposing a language model for molecular generation. RSC advances, 11(42), pp.25921-25932.

[89] Gregori-Puigjané, E. and Mestres, J., 2008. A ligand-based approach to mining the chemogenomic space of drugs. Combinatorial chemistry & high throughput screening, 11(8), pp.669-676.

[90] Anderson, A.C., 2003. The process of structure-based drug design. Chemistry & biology, 10(9), pp.787-797.

[91] Floresta, G., Zagni, C., Gentile, D., Patamia, V. and Rescifina, A., 2022. Artificial intelligence technologies for COVID-19 de novo drug design. International Journal of Molecular Sciences, 23(6), p.3261.

[92] In Silicon Approach for Discovery of Chemopreventive Agents - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Docking-

of-aspirin-into-the-binding-pocket-of-MMP-9-Atoms-and-bonds-were-displayed-as_fig4_318636040 [accessed 25 Jul, 2023]

[93] Song, Y.S., Dai, M.Z., Zhu, C.X., Huang, Y.F., Liu, J., Zhang, C.D., Xie, F., Peng, Y., Zhang, Y., Li, C.Q. and Zhang, L.J., 2021. Validation, optimization, and application of the zebrafish developmental toxicity assay for pharmaceuticals under the ICH S5 (R3) guideline. Frontiers in Cell and Developmental Biology, 9, p.721130.

[94] Xia, X., Hu, J., Wang, Y., Zhang, L. and Liu, Z., 2019. Graph-based generative models for de Novo drug design. Drug Discovery Today: Technologies, 32, pp.45-53.