22CSA699 DISSERTATION

AI BASED VIRTUAL SCREENING APPLICATION DEVELOPMENT FOR DRUG DISCOVERY

Submitted in partial fulfillment of
the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

Submitted by

Roll No	Name of Student
AM.EN.P2MCA22059	ACHYUT M SHARMA

Under the guidance of
Ms. Ani R

DEPARTMENT OF COMPUTER SCIENCE AND APPLICATIONS
AMRITA SCHOOL OF COMPUTING
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI CAMPUS

MAY, 2024

DEPARTMENT OF COMPUTER SCIENCE AND APPLICATIONS
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI CAMPUS

BONAFIDE CERTIFICATE

This is to certify that the project report entitled *AI BASED VIRTUAL SCREENING APPLICATION DEVELOPMENT FOR DRUG DISCOVERY* submitted by ACHYUT M SHARMA (AM.EN.P2MCA22059) in partial fulfillment of the requirements for the award of the Degree Master of Computer Applications is a bonafide record of the work carried out under my guidance and supervision at Amrita School of Computing, Amritapuri.

Signature of Project Guide with Date:
Mrs. Ani R:
Assistant Professor, Department of Computer Science and Applications

Signature of the Project Coordinators with Date:
Dr. S. Subbulakshmi:
Assistant Professors,Department of Computer Science and Applications

Signature of the Chairperson of the Department With date:
Mrs, Ani R:
Chairperson, Department of Computer Science and Applications

This project report was evaluated by us on ..... (24/05/2023)

INTERNAL EXAMINER EXTERNAL EXAMINER

DEPARTMENT OF COMPUTER SCIENCE AND APPLICATIONS
AMRITA VISHWA VIDYAPEETHAM
AMRITAPURI CAMPUS

DECLARATION

We, Deepak Menon, (AM.EN.P2MCA22017) and Achyut M Sharma, (AM.EN.P2MCA22059) hereby declare that this project entitled "AI BASED VIRTUAL SCREENING APPLICATION DEVELOPMENT FOR DRUG DISCOVERY" is a record of the original work done by me under the guidance of Ani R, Dept. of Computer Science and Applications, Amrita Vishwa Vidyapeetham , that this work has not formed the basis for any degree /diploma /associationship /fellowship or similar awards to any candidate in any university to the best of my knowledge.

Signature of Student with Date:
Deepak Menon

Signature of Student with Date:
Achyut M Sharma

Signature of Project Guide with Date:
Ani R

Place: Amritapuri, Kerala
Date: .

Acknowledgment

I offer my prayers and salutations at the Lotus Feet of the divine mother and the Chancellor of the University, Sri. Mata Amritanandamayi Devi.

I would like to express my gratitude to Mrs,Ani.R, Chairperson, Department of Computer Science and Applications, Amrita School of Computing for providing us the opportunity to do the project.

I am also grateful to Ms. Jisha.R.C, Vice Chairperson, Department of Computer Science and Applications for the support she extended throughout.

I want to thank my project guide, Mrs. Ani R, for her constant support and guidance throughout the project. Her valuable suggestions and feedback have been instrumental in shaping this project.

I am also grateful to Dr. S. Subbulakshmi, Assistant Professor, for providing us with the necessary resources and infrastructure to carry out this project. I want to thank the management of Amrita Vishwa Vidyapeetham for allowing us to work on this project.

I also would like to extend my sincere gratitude towards the reviewers of the project whose valuable comments and suggestions were instrumental in each phase of the project.

Many people, especially my classmates and friends, have made valuable comments and suggestions on this proposal, who gave us an inspiration to improve my project. I thank all of them for their valuable comments.

Deepak Menon
Achyut M Sharma

Abstract

Drug discovery, a multifaceted endeavor crucial for identifying therapeutic compounds, has traditionally involved extensive experimental work and significant financial investments, leading to high failure rates and prolonged development timelines. Recent advancements have moved the focus to computational methods and machine learning techniques to speed up and streamline this intricate process.

This study introduces VISMAI, a comprehensive tool designed for virtual screening that includes modules for descriptor calculation, fingerprint generation and for predicting drug-likeness, a key criterion in early drug development.

The descriptor calculation module extracts chemical properties from molecular structures, while the fingerprint generation module uses SMILES strings to produce molecular fingerprints. The GCNN-based model, leverages graph-based techniques to predict drug-likeness with high accuracy.

(Include testing done on the two Graph based models and the Ml model used by our seniors)

By integrating advanced neural network techniques, VISMAI enhances the predictive accuracy and efficiency of drug-likeness assessments. Future work will focus on improving model interpretability and expanding datasets to refine the tool further. This study underscores the transformative potential of computational approaches in pharmaceutical research, promoting innovation and collaboration within the community.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Drug discovery is the process of identifying and developing new therapeutic molecules that can be used as medications to treat various diseases or medical conditions. It involves a series of complex and time-consuming steps, including target identification, lead compound discovery, lead optimization, preclinical testing, and clinical trials. The traditional approach to drug discovery has been predominantly experimental, relying on techniques such as high-throughput screening (HTS) and structure-based drug design.

Computational drug discovery, or in silico drug discovery, refers to using computational methods and techniques to accelerate the drug discovery pipeline, complementing experimental approaches. It enables virtual screening of large chemical compound libraries against potential drug targets to rapidly identify promising lead compounds. Once leads are identified, computational techniques like molecular modeling, QSAR analysis, and molecular dynamics can improve their effectiveness, target specificity, and how they behave in the human body. Additionally, computational methods aid in drug repurposing by identifying existing approved drugs for new therapeutic applications, reducing development time and costs.

Our tool focuses on one major aspect of this drug discovery pipeline: Virtual

Screening. To be more specific, it focuses on Ligand-Based Virtual Screening (LBVS). Virtual screening is a computational technique widely used in drug discovery to identify potential lead compounds from large databases of chemical structures. It is a cost-effective and efficient alternative to traditional high-throughput screening methods. Ligand-based virtual screening (LBVS) , in particular is a crucial step in evaluating the drug-likeness and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties of potential lead compounds identified through structure-based virtual screening. This step involves applying various computational filters and predictive models to assess these properties, which are critical determinants of a molecule's suitability as a drug candidate.

In the pursuit of identifying promising drug candidates, assessing the drug-likeness of molecules is a crucial step in the early stages of drug discovery. Drug-likeness refers to the physicochemical and structural properties that are commonly associated with successful drugs and their ability to interact favorably with biological targets. Evaluating drug-likeness helps to prioritize molecules with a higher probability of exhibiting favorable pharmacokinetic and pharmacodynamic profiles, reducing the risk of attrition in later stages of drug development.

The pioneering rule of five (Ro5) was used to evaluate the drug-likeness of candidates and to reduce the number of entries. The majority (90%) of orally absorbed compounds approved before 1997 meets the Ro5. It consists of four criteria: molecular mass ($< 500Da$) , numbers of hydrogen bond donors (< 5)and acceptors (< 10), and log P (< 5) [1]. However, it was pointed out that many drugs, especially modern approved drugs, did not meet some of the criteria of Ro5. Simply relying on Ro5 may not be enough to screen suitable drugs [2]. Hopkins introduced the quantitative estimate of drug-likeness (QED), which is more comprehensive than Ro5 [3] . The QED value

reflects the underlying distribution of molecular properties, such as molecular weight, log P, topological polar surface area, the number of hydrogen bond donors and acceptors, the numbers of aromatic rings and rotatable bonds and the presence of unwanted chemical functional groups. Even so, researchers found that QED was inappropriate for the drugs distributed in different chemical spaces. This method summarizes the common properties of drug properties but does not consider the properties used to distinguish between drugs.

Traditional machine learning techniques such as Support Vector Machines (SVM)[4], Random Forests (RF) [5], and Artificial Neural Networks (ANN)[6] have historically been employed to predict drug-likeness by learning from approved drugs. However, recent progress in deep learning algorithms has outperformed these conventional methods. For example, Pei et.al [7] introduced a state-of-the-art model based on deep autoencoder neural networks (AE) with molecular descriptors, showcasing the potential of deep learning in enhancing drug-likeness prediction.However , more efficient models are still in need to improve the accuracy of drug development.

Graph-based neural networks, particularly Graph Convolutional Neural Networks (GCNNs), have shown significant promise in the field of drug-likeness prediction due to their ability to effectively model and analyze molecular structures [8]. Our tool integrates a drug-likeness prediction module that uses this approach. Here , the SMILES strings of molecules were converted into molecular graphs that served as the inputs for the model . Then, the model was applied to learn the real-valued molecular representation and predict the drug-likeness of that given compound or set of compounds.

Our Virtual Screening tool also focuses on the calculation of molecular descriptors, which provide valuable insights into the chemical structure, prop-

erties, and behavior of molecules. These descriptors are quantifiable representations of a compound's physicochemical properties, such as molecular weight, hydrophobicity, and hydrogen bond donors/acceptors, among others [9]. By analyzing these descriptors, researchers can gain critical insights into a molecule's reactivity, biological activity, solubility, stability, and other essential properties that influence its drug-likeness and ADMET characteristics.

Finally, our tool also includes a module for fingerprint generation that provides a unique representation of molecular structures, aiding in the identification of similar compounds. Fingerprints are bit strings or binary arrays where each bit represents the presence or absence of particular substructures or features within a molecule. They serve as a compact representation of a molecule's structure, capturing essential information about its chemical characteristics and facilitating rapid similarity searches and comparisons.

The tool employs a robust computational framework built on Python and the Django web framework, utilizing scientific libraries such as RDKit for molecular descriptor calculation and PyTorch for deep learning model development. The platform allows users to input SMILES (Simplified Molecular Input Line Entry System) strings, representing the chemical structures of molecules, and subsequently generates detailed outputs in the form of molecular descriptors, fingerprint and the drug-likeness value.

Chapter 2

Problem Definition

Drug discovery is a complex and resource-intensive process, with a high failure rate and prolonged timelines. Traditional experimental methods for identifying promising drug candidates and evaluating their pharmacokinetic properties are time-consuming and costly. There is a pressing need for computational approaches that can streamline the drug discovery pipeline, reduce the reliance on extensive experimental testing, and provide accurate predictions of drug-likeness and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties.

Recent advancements in machine learning and deep learning techniques, particularly Graph Convolutional Neural Networks (GCNNs), have shown promising results in predicting drug-likeness directly from molecular structures. However, existing approaches often suffer from limitations in handling complex molecular representations, capturing intrinsic structural patterns, and providing interpretable predictions.

As part of our work, we aim to develop a comprehensive Virtual Screening tool that leverages the power of Graph CNNs for accurate and interpretable drug-likeness prediction, while integrating RDKit for ADMET property prediction and fingerprint generation.

The proposed tool will address the following key challenges:

- Develop a novel Graph CNN architecture capable of learning from molecular graphs and capturing the topological patterns and functional group relationships within chemical structures, leading to improved drug-likeness prediction accuracy.
- Implement an unsupervised feature extraction mechanism, such as a molecular autoencoder, to learn latent representations of molecules, capturing intrinsic structural features that can enhance the predictive performance of the drug-likeness model.
- Integrate RDKit, a widely-used cheminformatics toolkit, for efficient ADMET property prediction and molecular fingerprint generation, enabling a comprehensive evaluation of drug candidates within the Virtual Screening tool.

By addressing these challenges, the proposed Virtual Screening tool aims to streamline the drug discovery process, reducing the time and resources required for identifying promising drug candidates and evaluating their pharmacokinetic properties. The tool's accurate and interpretable predictions, coupled with its integration of ADMET property prediction and fingerprint generation capabilities, will facilitate the optimization of drug design and lead to more efficient and cost-effective drug development pipelines.

Chapter 3

Related Work

Our study looks at the important work of Eric Gifford and Han van de Waterbeemd[13], which highlights the value of integrating in vitro and in silico techniques and data quality. Key obstacles and opportunities for advancement in the state-of-the-art prediction models are also given in their work. We also focus on the work of Dara S, Dhamercherla S, et al.[14], which highlights the potential of integrating AI technology with medical science to enhance decision-making, prediction power, and deep learning algorithms for improved therapeutic outcomes. Also of deep importance is the groundbreaking work of Hongming Chen, Ola Engkvist, et al.[15], where an integrative analysis of deep learning in drug development is done , addressing future obstacles and prospects.

SMILES (Simplified Molecular Input Line Entry System) strings are used by Adrià Cereto-Massagué, María José Ojeda, et al. in their work [16] because of their discrete nature, which makes comparison and similarity search in virtual screening efficient. Their research shows that virtual screening procedures can be made much more accurate and efficient by using SMILES fingerprints. A benchmark experiment utilizing the TOX 21 dataset showed the efficacy of the suggested methodology, with the SMILES-based CNN created in the paper by Maya Hirohara, Yutaka Saito, et al. [17] outperforming traditional

fingerprint approaches.

Some of the major research done in the field of drug discovery that were of high importance to our work have been analyzed further :

[A] ADMET in silico modelling: Towards prediction paradise? [13]

The paper examines two types of computational techniques for predicting the ADME(Absorption, Distribution, Metabolism, Excretion, and Toxicity) characteristics of small molecules: molecular modeling and data modeling.

- Molecular modeling methodologies utilize the three-dimensional configuration of proteins to predict the manner in which minute molecules will engage with them. This procedure can be performed through homology modeling, wherein a structural model of a protein is generated founded on the configuration of a linked protein. On the other hand, pharmacophore models may be employed, which are constructed by overlapping recognized substrates of the protein to ascertain shared characteristics implicated in binding.

- Data modeling approaches employ statistical analysis to ascertain correlations between a particular characteristic (such as an ADME parameter) and a collection of molecular descriptors. These models are trained on a dataset comprising molecules with established experimental data, enabling projections to be made on novel molecules. Various mathematical approaches can be employed in data modeling, encompassing multiple linear regression and partial least squares (PLS).

The paper also discusses on how machine learning methods are utilized in ADMET prediction to learn patterns and relationships from training

data and make predictions on new data. These methods can capture complex relationships and improve predictive accuracy by leveraging large datasets. Examples of machine learning methods mentioned in this paper include neural networks, self-organizing maps (SOM), recursive partitioning (RP), and support vector machines (SVM).

[B] Convolutional neural network based on SMILES representation of compounds for detecting chemical motif [17]

- The first technique utilized in the study is called SMILES notation, which functions as a kind of specialized language for describing chemical structures. It is employed to produce a sort of "map" for every chemical, using various properties to symbolize the atoms and unique symbols found in the structure of the molecule. These characteristics aid in understanding the numerous structural components of a chemical.

- A Convolutional Neural Network (CNN) was developed to analyze these chemical "maps." This system learns and recognizes patterns inside chemical structures, much like the way our brain recognizes patterns in images. It creates brief summaries, or "fingerprints," by examining these maps and capturing the most important properties of every molecule.

- They trained this system using a dataset of chemicals to teach it how to recognize and understand different compounds. They used specific ways to teach this system and checked how well it learned by looking at a graph that measures how accurate the computer was in understanding different chemicals.

- Additionally, they figured out a way to pick out important parts, or "chemical motifs," from these chemical fingerprints to help explain why certain chemicals might do specific things. These parts are like the essential features in a jigsaw puzzle that help understand what the final picture looks like.

[C] Prediction of Drug-Likeness Using Deep Autoencoder Neural Networks [7]

- The paper describes the process of data curation, including the removal of unsuitable compounds, duplicates, and the transformation of SMILES strings into canonical form. This step ensures the quality and consistency of the dataset used for model training.

- The study employs a variety of evaluation metrics, such as accuracy, specificity, sensitivity, AUC, MCC, R2, MAE, and RMSE, to assess the performance of both classification and regression models. The use of repeated training with random data splitting helps ensure robust model evaluation.

- The implementation of the MGA framework for multitask learning is a significant methodology employed in the paper. This approach allows shared learning across multiple endpoints, improving model generalization and prediction capabilities.

[D] DrugRep: an automatic virtual screening server for drug repurposing [18]

In the DrugRep tool, several methodologies are applied to enable drug repurposing efficiently. Here's a summary of the key methodologies used in DrugRep:

- **Receptor-based screening (RBS)**: DrugRep uses receptor-based screening, where it automatically determines the docking box for molecular docking. This method uses a novel cavity detection method to identify potential binding pockets of protein receptors. The tool then runs a docking simulation of each drug with the receptor using the AutoDock Vina docking software and predicts the binding affinity. This process enables the screening of potential drug compounds against specific protein targets.

- **Ligand-Based Screening (LBS)**: DrugRep performs ligand-based screening to measure the similarity of submitted ligands to compounds in drug libraries. It uses a variety of similarity metrics including Morgan Fingerprint, LSalign-Rigid, LSalign-Flex, FP2 and FP4. The Ligand-Screen score ranges from 0 to 1.00, where 1.00 means a perfect match and 0 means no match. This approach helps identify compounds with similar structures or properties to known drugs.

- **Benchmarking**: DrugRep validates its performance using established benchmark datasets such as DUD, DUD-E and MUV. These datasets consist of known active compounds and decoys for specific protein targets. DrugRep evaluates its screening results using metrics such as area under the curve (AUC) and enrichment factor (EF) to assess the accuracy and effectiveness of its predictions.

[E] Molecular fingerprint similarity search in virtual screening [19]

- The paper discusses the extensive use of virtual screening in drug discovery facilitated by two decades of computational advancements. Molecular fingerprints, known for their ease of use and efficiency in

substructure and similarity searches, play a crucial role in this context. The review focuses on widely used fingerprint algorithms, emphasizing the prevalence and advantages of 2D fingerprint-based methods over 3D alternatives in virtual screening for diverse targets

- The paper discusses the extensive use of virtual screening in drug discovery facilitated by two decades of computational advancements. Molecular fingerprints, known for their ease of use and efficiency in substructure and similarity searches, play a crucial role.

- Beyond virtual screening and drug discovery, the research introduces substructure keys-based and pharmacophoric fingerprints, detailing their encoding mechanisms. It provides insights into the software landscape for fingerprint-based virtual screening, highlighting notable packages and tools.

- In advocating for data fusion by combining different fingerprint approaches, the paper aligns with the current trend in molecular fingerprint-based similarity searching. This concise overview underscores the significance of fingerprints in drug discovery, emphasizing their versatility and efficiency in various applications.

[F] Graph Convolutional Neural Network-Based Virtual Screening of Phytochemicals and In-Silico Docking Studies of Drug Compounds for Hemochromatosis [20]

- The paper implements virtual screening , to identify potential drug candidates from large libraries of chemical compounds. Virtual screening involves two major approaches: Ligand-Based Virtual Screening (LBVS) and Structure-Based Virtual Screening (SBVS). LBVS relies

on the similarity of the chemical structures of potential drug candidates to known active compounds, while SBVS involves docking the candidate molecules into the three-dimensional structure of the target protein and evaluating their binding affinity.

- For LBVS, the paper employs various machine learning algorithms to predict the drug-likeness of compounds based on their molecular descriptors (physicochemical properties) and molecular fingerprints (binary representations of structural features).

- The paper proposes the use of Graph Convolutional Neural Networks (GCNs) for drug-likeness prediction. GCNs are a type of deep learning algorithm that can learn features directly from the molecular graph representation of compounds, where atoms are represented as nodes and bonds as edges. The GCN model takes SMILES (Simplified Molecular Input Line Entry System) strings as input and generates graph-based molecular fingerprints, which are then used to predict drug-likeness.

[G] Mordred: a molecular descriptor calculator[21]

- The paper describes the development of Mordred, a software application for calculating molecular descriptors. The performance of Mordred is benchmarked against well-known descriptor-calculation software, demonstrating its speed and ability to calculate descriptors for large molecules. The results show that Mordred is at least twice as efficient as the PaDEL-Descriptor and can handle descriptors for large molecules that other tools cannot.

- Mordred is available as a web application, enabling users to easily access and use the software. The web interface allows users to upload structure files. The web interface also allows users to preview com-

pound conformations, which can be useful for understanding the three-dimensional structure of molecules and their potential binding sites for ligands or other molecules.

- Mordred uses a two-part system to calculate molecular descriptors. The "Descriptor" class contains the algorithms for calculating the descriptors, while the "Calculator" class is used to register and compute the descriptors. This approach allows users to easily calculate over 1800 two- and three-dimensional molecular descriptors.

[H] Deep learning in drug discovery: an integrative review and future challenges [15]

- The research offers a comprehensive examination of drug discovery, highlighting the crucial analysis of drug interactions within the body to achieve therapeutic effects. Major pharmaceutical companies are increasingly turning to artificial intelligence (AI) and deep learning (DL) techniques, departing from outdated methods with the goal of improving patient outcomes and corporate profitability.

- This review breaks new ground by incorporating recent DL models and applications in various aspects of drug discovery, including drug-target interactions, drug-drug similarity interactions, drug sensitivity, response predictions, and predictions of drug side effects. It also introduces emerging concepts such as explainable AI (XAI) and digital twinning (DT), demonstrating their role in advancing drug discovery challenges.

- The study aims to conduct a systematic literature review (SLR) that integrates recent DL technologies for a range of drug discovery problems, including drug-target interactions, drug-drug similarity interac-

tions, drug sensitivity, and predictions of drug side effects. While DL has demonstrated effectiveness in addressing these challenges, the research emphasizes its ongoing potential as an exciting and open field for further exploration by interested researchers.

Chapter 4

Background Knowledge

There are some key aspects of that hold significant importance to our study. We will go through some of them here :

4.1 Virtual Screening

Virtual screening is a computational technique widely used in the drug discovery process to identify potential lead compounds from large chemical libraries. It serves as a cost-effective and efficient alternative to traditional high-throughput screening methods, enabling researchers to rapidly evaluate and prioritize a vast number of molecules based on their likelihood of exhibiting favorable interactions with biological targets.

The importance of virtual screening in drug discovery is multifaceted, contributing to several key areas:

- **Accelerating Early-Stage Discovery** : Virtual screening fast-tracks early drug discovery by sifting through millions of candidates to a select few with high potential. Unlike traditional, time-consuming HTS methods, virtual screening preselects promising molecules based on their predicted activity, allowing researchers to focus on a smaller pool for further testing.

- **Cost-Effectiveness and Resource Efficiency** : Virtual screening offers a cost-effective alternative to traditional experimental methods. By employing computational techniques, researchers can screen large libraries of compounds without the need for expensive reagents and labor-intensive experimental setups. This efficiency translates into significant cost savings, enabling smaller research teams and academic institutions to participate in drug discovery efforts.

- **Enhanced Predictive Accuracy** : The integration of machine learning algorithms into virtual screening workflows significantly enhances the accuracy of drug-likeness assessments. By leveraging extensive datasets of known drug properties, these models can identify subtle patterns and make informed predictions about the suitability of novel compounds. By analyzing potential drugs with in-silico analysis, we can improve our chances of picking good candidates early on. This reduces the risk of failures in later stages of testing, making the whole drug discovery process more efficient.

Virtual screening can be broadly classified into two main approaches: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS).

We will exclusively direct our attention to ligand-based virtual screening, as our study focuses on that.

4.1.1 Ligand Based Virtual Screening

The term "ligand" refers to a small molecule that binds to a biological target, such as a protein or receptor, and modulates its activity. In drug discovery, researchers often start with known active ligands that have demonstrated biological activity against a particular target of interest. The screening process in Ligand Based Virtual Screening (LBVS) relies on using these known

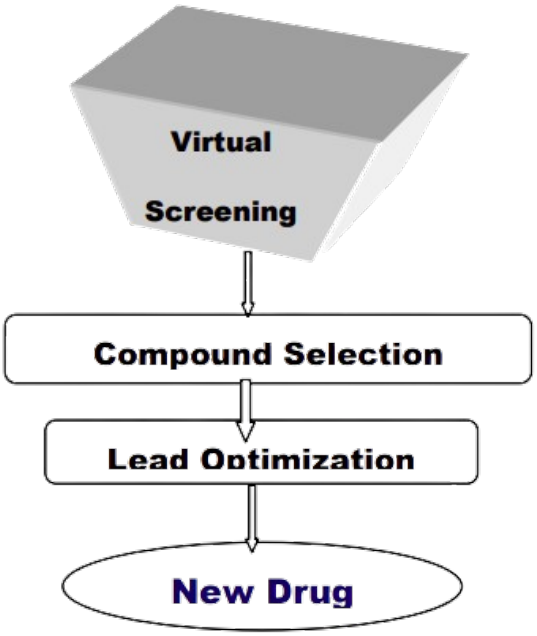

Figure 4.1: A schematic representation of the virtual screening process

active ligands as a reference point to search for and identify new, structurally similar compounds that may also exhibit similar biological activities. The underlying assumption is that molecules with similar structures are likely to share similar properties and interactions with the biological target.

Here, fingerprints of molecules are utilized to prioritize candidate molecules based on their similarity to known active ligands, aiding in the identification of potential hits for further testing.

The process involves the following key steps:

- **Molecular Representation** : Each candidate molecule is represented using fingerprints, which can be 2D or 3D descriptors capturing structural features.
- **Similarity Evaluation** : The fingerprints are used to assess the simi-

larity between candidate molecules and known active compounds. This comparison helps in ranking the candidates based on their resemblance to known actives.

- **Hit Selection** : Candidate molecules are then ranked according to their similarity scores, with the aim of identifying a small number of top-ranked compounds as potential hits from a large library of molecules.

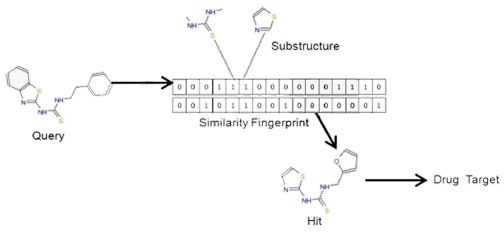

Figure 4.2: Similarity search using fingerprints

4.2 SMILES Notation

The Simplified Molecular Input Line Entry System (SMILES) is a widely used notation for representing chemical structures using ASCII strings. This compact and human-readable format encodes molecular structures by specifying the atoms, bonds, and connectivity within a molecule. SMILES notation has become an indispensable tool in cheminformatics, significantly contributing to drug discovery and virtual screening processes. We use SMILES strings in our work for the following reasons :

- **Descriptor Calculation** : SMILES can be used to generate molecular descriptors, which are numerical values representing various chemical properties. Descriptors play a crucial role in the efficient and effective identification of potential drug candidates from large chemical libraries.

- **Compatibility with Algorithms** : Many virtual screening algorithms and cheminformatics tools are designed to work directly with SMILES strings. This compatibility allows for quick computation of molecular properties, similarity searches, and docking simulations.

- **Data Integration** : SMILES allows for seamless integration of chemical data across various databases, software tools, and research studies. This uniformity ensures that compounds can be easily compared, searched, and analyzed regardless of the source of the data.

Canonical SMILES are a specific version of the SMILES notation that provides a unique representation for a given molecule.When using canonical SMILES, the same molecular structure will always result in the same SMILES string, making it ideal for database searches, indexing, and molecular comparisons.

Compound Name	**Chemical Formula**	**Canonical SMILES**
Amiloxate	C15H20O3	COc1ccc(=C(=O)OCCC(C)C)cc1
Ibuprofen	C13H18O2	CC(C)Cc1ccc(cc1)C(C)C(O)=O
Minoxidil	C9H15N5O	NC1=CC(=NC(=N)N1O)N2CCCCC2

Table 4.1: Example compounds along with their chemical formula and corresponding SMILES notations

4.3 ADMET & Physicochemical properties

The identification and optimization of promising drug candidates heavily rely on their ability to exhibit favorable absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. These properties play a crucial role in determining the efficacy, safety, and overall success of a potential drug molecule. Evaluating ADMET characteristics early in the drug discovery

process is essential to reduce the risk of late-stage failures and improve the chances of clinical success. Let's now delve into this topic in more detail :

4.3.1 Absorption

Effective absorption of a drug molecule into the systemic circulation is essential for its therapeutic efficacy. Various factors, including solubility, permeability, and gastrointestinal stability, play crucial roles in determining a compound's absorption. Inadequate absorption can result in low bioavailability, restricting the drug's therapeutic effectiveness. Computational models and in vitro assays are frequently utilized to anticipate and evaluate the absorption characteristics of potential drug candidates.

4.3.2 Distribution

Once absorbed, the drug molecule must be distributed effectively throughout the body to reach its intended target site. Factors such as plasma protein binding, tissue partitioning, and the ability to cross biological barriers (e.g., blood-brain barrier) can impact a drug's distribution profile.

4.3.3 Metabolism

Metabolism refers to the chemical alterations that a drug undergoes within the body, primarily in the liver, although other organs can also play a role.A key factor to consider in drug metabolism is metabolic stability, which refers to how resistant a drug molecule is to being broken down by enzymes in the body. Drug molecules that have high metabolic stability are less likely to be quickly broken apart by these enzymes. This means that more of the drug can stay intact and available in the body for a longer period of time.

4.3.4 Excretion

The elimination of drug molecules and their metabolites from the body is essential to maintain therapeutic concentrations and prevent unwanted accumulation.Inadequate clearance can lead to drug accumulation and potential toxicity. Computational models, along with in vitro and in vivo studies, can help predict and optimize the excretion properties of drug candidates.

4.3.5 Toxicity

Toxicity refers to the potential of a drug or its metabolites to cause harm or adverse effects to living organisms, typically at therapeutic doses or concentrations. In-silico methods for predicting toxicity leverage computational models and databases to analyze the chemical structure of the drug and its metabolites and predict their potential for causing harm. Toxicity assessment spans various aspects, including acute toxicity, which analyzes immediate harm from single exposures, chronic toxicity, which investigates prolonged risks from repeated use and organ toxicity scrutinizes the drug's impact on specific organs like the liver, kidneys, heart, or nervous system.

In addition to ADMET properties, physicochemical properties play a crucial role in understanding the behavior and characteristics of chemical compounds. These properties provide insights into a molecule's structure, composition, and interactions with other substances. Among the key physicochemical properties discussed in our work are molecular formula, which delineates the precise composition of a compound's molecule in terms of its constituent elements. Molecular weight offers information about a molecule's mass and size, influencing its physical properties and interactions with other chemicals. Additionally, the presence of radical electrons, which contain unpaired electrons, and valence electrons, located in an atom's outermost energy level, contribute to the molecule's physicochemical attributes.

Moreover, properties like heavy atom, molecular weight and exact molecular weight provide a comprehensive understanding of a molecule's atomic composition by considering both heavy and hydrogen atoms. LogP, also known as the partition coefficient, quantifies a compound's hydrophobicity or lipophilicity, influencing its solubility and distribution in biological systems. Topological polar surface area (TPSA) offers insights into a molecule's surface area when exposed to solvent molecules, indicating its potential interactions with biological targets. The count of rotatable bonds in a molecule reflects its torsional freedom or flexibility, affecting its conformational dynamics and biological activity. Additionally, hydrogen bond acceptors and donors identify functional groups within a compound capable of participating in hydrogen bonding interactions, which play crucial roles in molecular recognition and binding affinity. These physicochemical properties collectively contribute to the characterization and prediction of a compound's behavior, aiding in drug discovery and development processes.

4.4 RDKit

RDKit is an open-source cheminformatics software package designed for chemists, bioinformaticians, and computational scientists. It provides a comprehensive suite of functions and algorithms for molecular modeling, analysis, and visualization. One of its key features is its ability to generate molecular descriptors and fingerprints, which are essential for characterizing and comparing chemical compounds based on their structural and physicochemical properties..

The utilization of RDKit in virtual screening and drug discovery workflows offers several advantages :

- **Open-Source Accessibility**: RDKit's open-source nature fosters collaboration and customization within the scientific community, enhanc-

ing accessibility and encouraging shared development efforts.

- **User-Friendly Integration**: RDKit offers a user-friendly interface and extensive documentation, simplifying its integration into existing computational pipelines and streamlining workflow processes.

- **Multi-Language Support**: With support for multiple programming languages such as Python and C++, RDKit enables seamless integration with various software libraries and platforms, ensuring versatility and compatibility across diverse research environments.

For the above mentioned reasons, our study extensively leverages RDKit and another similar package called Mordred for our descriptor and molecular fingerprint calculation tasks. RDKit prioritizes descriptors that are particularly relevant for drug discovery applications, such as Lipinski's Rule of Five descriptors or fragment-based descriptor. Mordred on the other hand specializes in certain types of descriptors, such as those related to molecular shape, size, flexibility, or electronic properties.

4.5 Drug-likeness

Drug-likeness refers to the concept of assessing whether a chemical compound possesses properties that are characteristic of known drugs. In the context of in-silico drug discovery, drug-likeness serves as a criterion for evaluating the potential of a compound to be developed into a therapeutic agent. It involves the identification of molecular features and physicochemical properties that are commonly observed in drugs with desirable pharmacological profiles.

Simple rule-based filters like Lipinski's Rule of Five, discussed earlier [1], Ghose, Veber, Egan and Muegge methods are commonly used to assess

oral drug-likeness based on molecular weight, lipophilicity, hydrogen bond donors/acceptors, polar surface area and rotatable bonds. Compounds that violate these rules are more likely to have poor ADMET profiles [10].

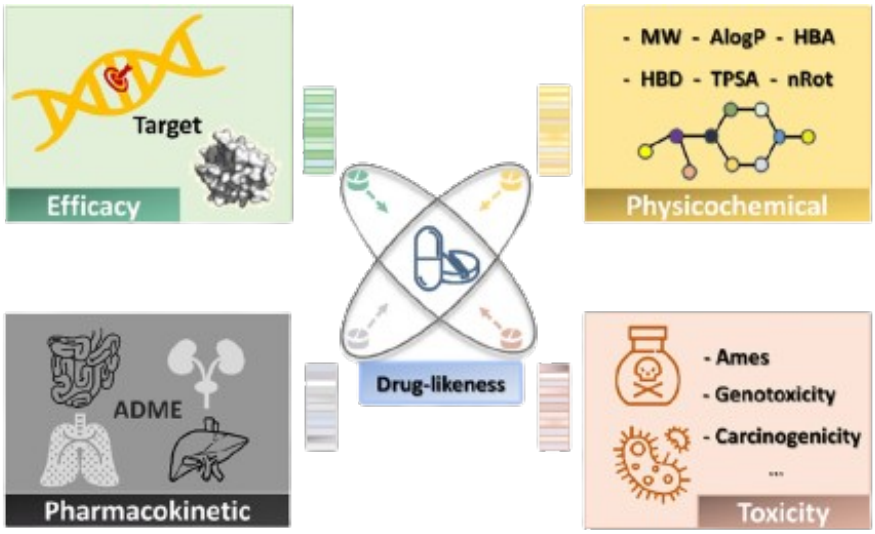

Figure 4.3: Overview of drug-likeness as a result of the various ADMET and Physiochemical properties of a molecule

Quantitative prediction models using advanced machine learning techniques have emerged as powerful tools for assessing drug-likeness in silico. These models can capture complex relationships between molecular structure and drug-like properties, enabling more accurate predictions compared to traditional rule-based filters. Machine learning algorithms such as artificial neural networks (ANNs), support vector machines (SVMs), and decision trees [11] have been trained on large datasets of known drugs and non-drug-like compounds. By learning the patterns and associations between molecular descriptors and drug-likeness, these models can rapidly evaluate the drug-like potential of vast chemical libraries.

As the field of in silico drug discovery continues to evolve, the integration

of cutting-edge machine learning models will play an increasingly important role in the efficient development of new therapeutic agents. Graph Convolutional Neural Networks (GCNs) have emerged as powerful tools in various fields, including drug discovery, due to their ability to effectively model non-Euclidean structured data. We will be discussing about that in the coming section, as those implementations are what we will be using for our drug likeness prediction module in our work.

4.6 Graph Convolutional Neural Networks (GCNs)

Graph Convolutional Neural Networks (GCNs) have their roots in the work of Kipf and Welling [12], who introduced an effective graph model for semi-supervised learning. GCNs can learn representations of data that are inherently structured as graphs, such as social networks, citation networks, and most importantly for our study, molecular structures.

Here's a breakdown of how GCNs work and how they can be applied to molecular data :

- **Graph Representation** : In molecular data, atoms are represented as nodes, and chemical bonds between atoms are represented as edges in a graph. Each node typically contains features representing the properties of the corresponding atom (e.g., atom type, atomic number, charge), and each edge may contain features representing the type and strength of the bond between atoms.

- **Convolutional Operations on Graphs** : GCNs adapt the concept of convolutional operations from traditional neural networks to operate on graphs. In traditional convolutional neural networks (CNNs), convolutions are applied over regular grid structures (e.g., images) to

detect patterns. In GCNs, convolutions are performed over the graph structure to aggregate information from neighboring nodes.

- **Message Passing** : The key operation in GCNs is message passing, where each node aggregates information from its neighboring nodes. This aggregation typically involves a weighted sum of the features of neighboring nodes, where the weights are determined by the structure of the graph (i.e., the adjacency matrix).

- **Multiple Graph Convolutional Layers** : GCNs consist of multiple graph convolutional layers, each performing message passing and feature transformation. Each layer refines the node representations by aggregating information from neighboring nodes and updating the node features based on the aggregated information.

- **Pooling for Graph-level Representations** : Pooling operations in GCNs aggregate information from node-level representations to derive graph-level features. This allows the network to capture holistic characteristics of the molecule in context of our work, beyond the individual atoms and bonds.

- **Readout for Prediction** : Once the graph-level representation is obtained through pooling, a readout operation is applied to map this representation to the task of drug likeness prediction. The readout layer takes the pooled features and transforms them into predictions or scores that indicate the likelihood of the molecule possessing desirable drug-like properties.

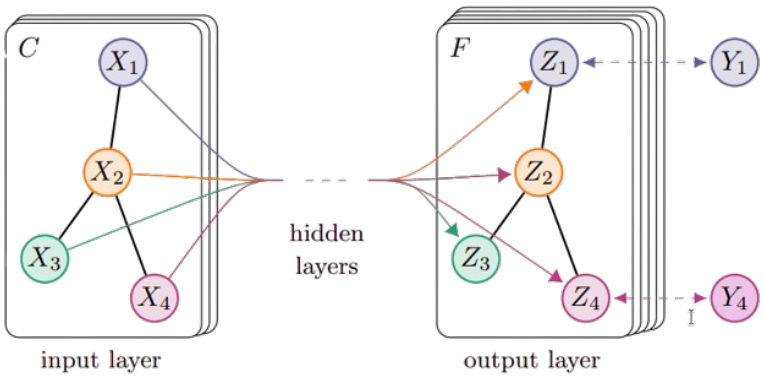

Figure 4.4: Structure of a typical Graph Neural Network

Chapter 5

Design and Methodology

5.1 Proposed System

The proposed Virtual Screening tool is designed to streamline the drug discovery process by leveraging advanced machine learning models and computational techniques. At the core of the system lies a Graph Convolutional Neural Network (GCNN) model, which plays a crucial role in predicting drug-likeness directly from molecular structures.

The system accepts SMILES (Simplified Molecular-Input Line-Entry System) strings as input, representing the molecular structure of potential drug candidates. These SMILES strings are then processed to generate molecular fingerprints, which are compact representations of the structural features and properties of the molecules. The fingerprint generation module, powered by the widely-used RDKit cheminformatics toolkit, ensures efficient processing and representation of molecular structures.

With the molecular fingerprints as input, the GCNN model is able to learn and capture the intricate patterns and relationships within the chemical structures. By leveraging the power of graph convolutional networks, the model can effectively process the molecular graphs and identify the topological patterns, functional group interactions, and other relevant features that

contribute to drug-likeness. The GCNN model's output provides a probability score indicating the likelihood of a given molecule being a potential drug candidate.

In parallel, the RDKit toolkit is integrated into the system to perform ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) property prediction. This module analyzes the molecular structures and provides valuable insights into the pharmacokinetic properties of the drug candidates, such as their bioavailability, metabolism, and potential toxicity. By combining the drug-likeness predictions from the GCNN model with the ADMET property predictions, the Virtual Screening tool offers comprehensive evaluation of the drug candidates, enabling informed decision-making in the drug discovery process.

5.1.1 Architecture Diagram

The user interface (UI) of our web application is crafted using Python Django and Bootstrap, ensuring a responsive and user-friendly design. The integration of Django provides a robust backend while Bootstrap facilitates adaptability across various devices, from desktops to mobile phones. Our application focuses on three main functionalities: molecular descriptor calculation, fingerprint generation, and drug-likeness prediction. Users can easily navigate between these features via dedicated buttons and input SMILES strings to receive results, which are downloadable in CSV format.

Each section—molecular descriptor calculation, fingerprint generation, and drug-likeness prediction—allows users to upload SMILES strings and quickly obtain relevant data. The responsive design ensures optimal performance and usability on any device, catering to the needs of professionals who require access to the application from different locations. This combination of

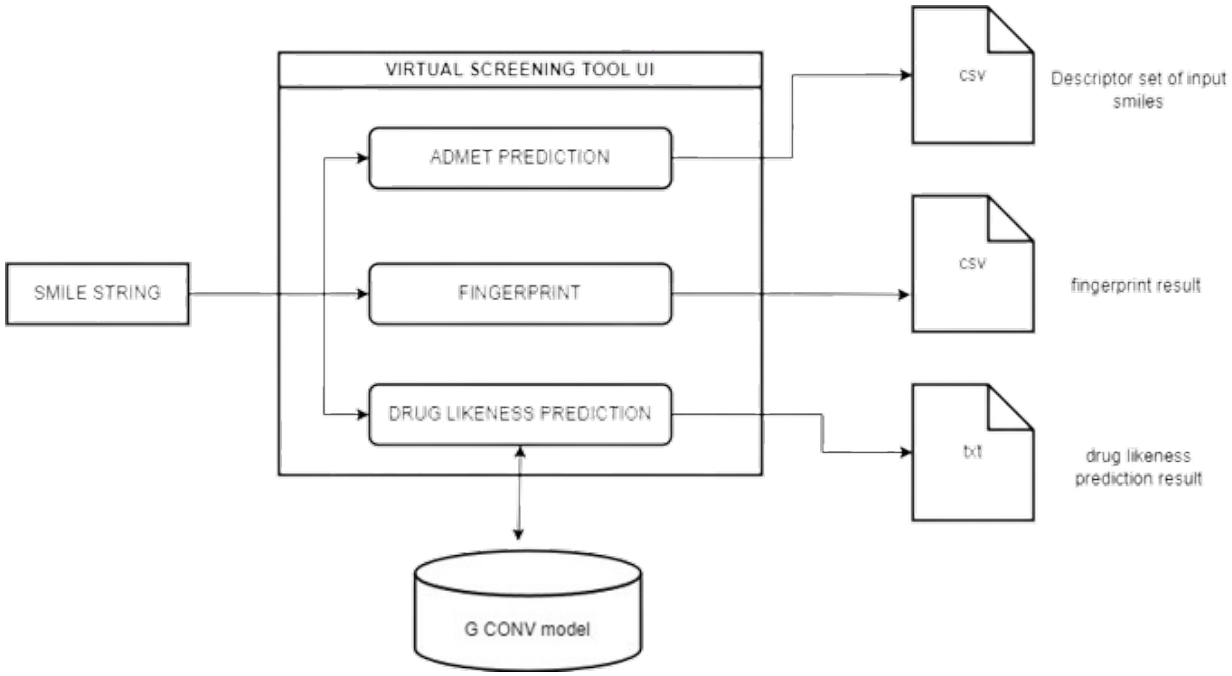

Figure 5.1: Our proposed architecture for the Virtual Screening Tool

Django and Bootstrap delivers a seamless, efficient, and accessible tool for cheminformatics and drug discovery research.

5.1.2 Solution Approach

Considering the drug-likeness prediction module of our work , for the prepration of the Graph Convolutional Neural Network (GCN) models, we used the PyTorch Geometric library. It is a powerful library designed to ease the implementation of GCNs using PyTorch and provides essential functionalities for handling graph-structured data efficiently, making it a go-to choice for researchers and practitioners in the field of graph-based machine learning.

One of the key components of PyTorch Geometric is its collection of graph convolutional layers, which includes GraphConv and SAGEConv. GraphConv is a fundamental building block for GNNs, performing convolution

operations directly on the graph structure. It leverages the graph's adjacency matrix to propagate information across nodes, allowing for effective feature extraction and representation learning in graph data. SAGEConv, on the other hand, stands for Sample and Aggregate Graph Convolutional Network. It extends the concept of graph convolution by employing a neighbor sampling strategy, where it aggregates information from a node's local neighborhood to enhance feature representation while controlling computational complexity. For our study, we have built two models , one using GraphConv layers and the other using SAGEConv layers.

Utilizing GraphConv layers

In our first model, the GraphConv layers perform convolution operations directly on the graph structure, allowing the model to extract features from the input graph data. Our specific architecture uses three GraphConv layers that are stacked sequentially, each progressively capturing more abstract and higher-level features from the input node attributes and graph topology. These layers enable the model to learn representations of nodes that incorporate information from their local and global neighborhoods, thus facilitating effective feature extraction in graph data.

Batch normalization (BatchNorm) layers are inserted after each GraphConv layer. Batch normalization helps stabilize and accelerate the training process by normalizing the input to each layer, reducing internal covariate shift and accelerating convergence. By normalizing the activations, BatchNorm layers ensure that the model can learn more efficiently and effectively, leading to improved generalization and faster convergence during training.

Dropout layers are employed after each activation function to mitigate overfitting and enhance the model's robustness to noise in the input data. In this

```python
class ImprovedGraphConvNet(torch.nn.Module):
    def __init__(self, num_node_features, num_classes):
        super(ImprovedGraphConvNet, self).__init__()
        self.conv1 = GraphConv(num_node_features, 64)
        # self.conv1 = SAGEConv(num_node_features, 64) //Incase of SAGEConv
        self.bn1 = BatchNorm(64)
        self.conv2 = GraphConv(64, 128)
        # self.conv2 = SAGEConv(64, 128) //Incase of SAGEConv
        self.bn2 = BatchNorm(128)
        self.conv3 = GraphConv(128, 128)
        # self.conv3 = SAGEConv(128, 128) //Incase of SAGEConv
        self.bn3 = BatchNorm(128)
        self.fc1 = torch.nn.Linear(128, 64)
        self.fc2 = torch.nn.Linear(64, num_classes)
        self.dropout = Dropout(0.5)

    def forward(self, data):
        x, edge_index, batch = data.x, data.edge_index, data.batch
        x = self.conv1(x, edge_index)
        x = self.bn1(x)
        x = F.relu(x)
        x = self.dropout(x)
        x = self.conv2(x, edge_index)
        x = self.bn2(x)
        x = F.relu(x)
        x = self.dropout(x)
        x = self.conv3(x, edge_index)
        x = self.bn3(x)
        x = F.relu(x)
        x = self.dropout(x)
        x = global_mean_pool(x, batch)  # Global mean pooling
        x = F.relu(self.fc1(x))
        x = self.dropout(x)
        x = self.fc2(x)
        return F.log_softmax(x, dim=1)
```

Figure 5.2: Code snippet of our GCN using GraphConv and SAGEConv layers as the convolutional layers

architecture, a dropout probability of 0.5 is used, meaning that each neuron has a 50% chance of being dropped out during training, which helps prevent the model from relying too heavily on any individual feature or node.

Global mean pooling is applied to aggregate the node representations across the entire graph and produce a fixed-size representation for the entire graph. This pooling operation ensures that the model can make predictions based on the entire graph structure rather than individual nodes, enabling it to classify entire graphs into different categories. The resulting pooled representation

is then passed through fully connected layers (Linear) for further processing before the final classification decision is made using a softmax activation function.

Utilizing SAGEConv layers

Similar to GraphConv layers, SAGEConv (Sample and Aggregate Graph Convolutional Network) layers are fundamental components responsible for feature extraction from graph-structured data. They operate by sampling and aggregating information from a node's local neighborhood, allowing the model to capture both local and global graph features effectively.

In our architecture, three SAGEConv layers are sequentially stacked to progressively learn hierarchical representations of the input graph data. These layers enable the model to extract informative features from the graph while controlling computational complexity, making them suitable for handling large-scale graphs.

The batch normalization , dropout and pooling layers are similar to how we have implemented them for our GraphConv architecture and as such we will not be discussing them again.

Responsive UI

The user interface (UI) of our web application is crafted using Python Django and Bootstrap, ensuring a responsive and user-friendly design. The integration of Django provides a robust backend while Bootstrap facilitates adaptability across various devices, from desktops to mobile phones. Our application focuses on three main functionalities: molecular descriptor calculation, fingerprint generation, and drug-likeness prediction. Users can easily navigate between these features via dedicated buttons and input SMILES strings

to receive results, which are downloadable in CSV format.

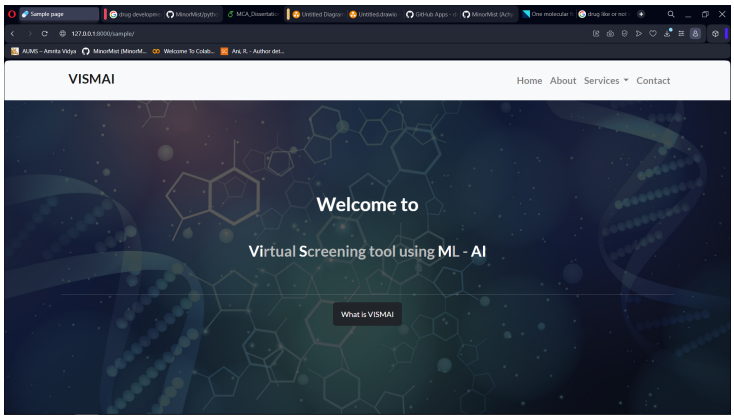

Figure 5.3: Home page of screening tool

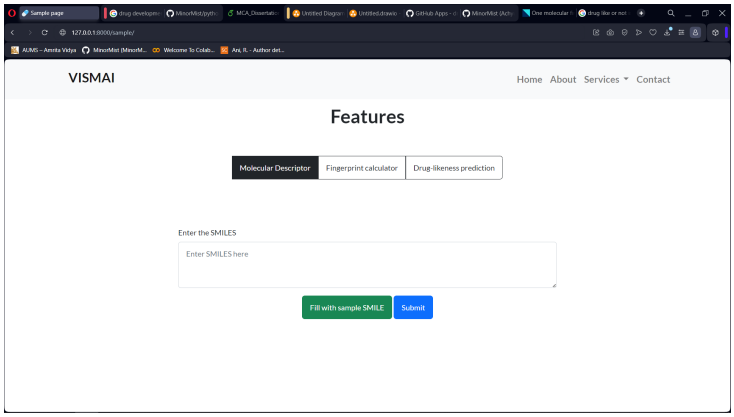

Figure 5.4: Services section

Chapter 6

Experimentation and Result Analysis

6.1 Dataset

Our work includes a comprehensive data extraction and descriptor calculation path, aimed at generating a dataset of molecular descriptors from compounds listed on the DrugBank website. The process begins by iterating through multiple pages of approved drugs on DrugBank and fetching the HTML content using the requests library. We then parse it with BeautifulSoup to extract relevant data.

The script first identifies table rows that contain the compound names and their respective chemical formulas. It uses CSS class selectors to locate these elements accurately. The compound name is then cleaned and processed to extract the SMILES (Simplified Molecular Input Line Entry System) representation using a function that we specified. This function queries the Cactus Chemical Identifier Resolver service to convert chemical names to their corresponding SMILES strings. If the SMILES string is successfully obtained, the script proceeds to calculate molecular descriptors.

Molecular descriptors are calculated using RDKit and Mordred libraries. The

results from both RDKit and Mordred calculations are consolidated into a Pandas DataFrame. This DataFrame includes the compound name, chemical formula, SMILES string, and all computed descriptors. Before saving the data, the script performs several data cleaning steps, such as removing invalid SMILES strings, handling missing values, and dropping non-numeric columns. The final cleaned and combined DataFrame is then appended to a CSV file ensuring that the dataset is incrementally built and stored in a structured format.

6.2 Experimental Setup

The system was built using the Django web framework, which provided a robust and scalable foundation for the application's backend. HTML and CSS were utilized for creating the user interface, enabling an intuitive and visually appealing front-end experience.

For the development and training of the Graph Convolutional Neural Network (GCNN) models, we utilized the powerful Kaggle platform. Kaggle offers good GPU availability , allowing for efficient model development, experimentation, and evaluation.

The drug-likeness prediction module was implemented using the PyTorch Geometric library, a state-of-the-art framework designed specifically for working with graph-structured data. This library facilitated the construction and training of our GCNN models, enabling us to leverage the power of graph convolutional networks for accurate drug-likeness predictions.

6.3 Results

6.3.1 Evaluation Metrics for Inference

The performance of classification models like the one's proposed above using Graph Convolutional Neural Networks are evaluated using certain metrics. The most prevalent ones among them are :

1. Accuracy : Accuracy is one of the most straightforward and widely used metrics for evaluating the performance of a classification model. It is defined as the proportion of correctly classified instances out of the total number of instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Figure 6.1: Calculation of Accuracy of a model

2. Precision : Precision is the ratio of the number of true positives to the total number of positive predictions.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Figure 6.2: Calculation of Precision of a model

3. Recall : Recall is the ratio of the number of true positives to the total number of actual (relevant) samples.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Figure 6.3: Calculation of Recall of a model

4. F1 Score : The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns. It is particularly useful when you need to take both false positives and false negatives into account.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 6.4: Calculation of F1 Score of a model

6.3.2 Comparison of Evaluation Metrics

We now compare the metrics mentioned above for both the models.

	Accuracy	**Precision**	**Recall**	**F1 Score**	**MSE**
GraphConv	0.5084	0.4964	0.4968	0.4853	0.5000
SAGEConv	0.4581	0.2304	0.4958	0.3124	0.5405

Table 6.1: Comparing Evaluation Metrics of the GraphConv and SAGEConv models

We can see that our model made using GraphConv gives better results compared to SAGEConv. The metrics can be improved and further preprocessing and changing the hyperparameters may improve the performance of the model.

Experimenting with other Graph Convolutional Neural Network implementations might also result in better and accurate results.

.

Chapter 7

Conclusion

The development of our proposed virtual screening tool, represents a significant step towards streamlining the drug discovery process. By leveraging advanced machine learning techniques, particularly Graph Convolutional Neural Networks (GCNNs), this tool offers better prediction of drug-likeness, a crucial criterion in early drug development.The integration of RDKit and Mordred libraries enables robust calculation of molecular descriptors and fingerprint generation, providing valuable insights into the chemical properties and structural features of potential drug candidates.

The experimental results demonstrate the effectiveness of the proposed approach, with the GraphConv model outperforming the SAGEConv model in terms of evaluation metrics such as accuracy, precision, recall, and F1 score. However, there is still room for improvement, and further refinements to the models, such as optimizing hyperparameters and exploring alternative GCNN architectures, could lead to enhanced predictive performance.

Future work will focus on improving model interpretability, enabling researchers to gain insights into the underlying factors influencing drug-likeness predictions. Additionally, expanding the dataset and incorporating diverse chemical spaces could further refine the tool's predictive capabilities, ensuring its applicability in the long term drug discovery domain. .

Bibliography

[1] Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 64, 4–17. doi:10.1016/j.addr.2012.09.019

[2] Shultz MD. Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. J Med Chem. 2019 Feb 28;62(4):1701-1714. doi: 10.1021/acs.jmedchem.8b00686. Epub 2018 Sep 27. PMID: 30212196.‘

[3] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012 Jan 24;4(2):90-8. doi: 10.1038/nchem.1243. PMID: 22270643; PMCID: PMC3524573.

[4] Li,Q. et al. (2007) A large descriptor set and a probabilistic kernel-based clas- sifier significantly improve druglikeness classification. J. Chem. Inf. Model., 47, 1776–1786

[5] Chen,H. et al. (2018) The rise of deep learning in drug discovery. Drug Discov. Today, 23, 1241–1250

[6] Korkmaz,S. (2020) Deep learning-based imbalanced data classification for drug discovery. J. Chem. Inf. Model., 60, 4180–4190.

[7] Hu Q, Feng M, Lai L, Pei J. Prediction of Drug-Likeness Using Deep Autoencoder Neural Networks. Front Genet. 2018 Nov 27;9:585. doi: 10.3389/fgene.2018.00585. PMID: 30538725; PMCID: PMC6277570.

[8] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, Fei Wang, Graph convolutional networks for computational drug development and discovery, Briefings in Bioinformatics, Volume 21, Issue 3, May 2020, Pages 919–935, https://doi.org/10.1093/bib/bbz042

[9] Mauri, Andrea Consonni, Viviana Todeschini, Roberto. (2017). Molecular Descriptors. 10.1007/978-3-319-27282-5$_5$1.

[10] Udugade, S.B., Doijad, R.C., Udugade, B.V. (2019). In silico evaluation of pharmacokinetics, drug-likeness and medicinal chemistry friendliness of Momordicin1: an active chemical constituent of Momordica charantia. Journal of Advanced Scientific Research, 10(03 Suppl 1), 222-229

[11] Tian, Sheng Wang, Junmei Li, Youyong Xu, Xiaojie Hou, Tingjun. (2012). Drug-likeness Analysis of Traditional Chinese Medicines: Prediction of Drug-likeness Using Machine Learning Approaches. Molecular pharmaceutics. 9. 2875-86. 10.1021/mp300198d.

[12] Kipf, Thomas Welling, Max. (2016). Semi-Supervised Classification with Graph Convolutional Networks.

[13] van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov. 2003 Mar;2(3):192-204. doi: 10.1038/nrd1032. PMID: 12612645.

[14] Dara, S., Dhamercherla, S., Jadav, S.S. et al. Machine Learning in Drug Discovery: A Review. Artif Intell Rev 55, 1947–1999 (2022). https://doi.org/10.1007/s10462-021-10058-4

[15] Askr, H., Elgeldawi, E., Aboul Ella, H. et al. Deep learning in drug discovery: an integrative review and future challenges. Artif Intell Rev 56, 5975–6037 (2023). https://doi.org/10.1007/s10462-022-10306-1

[16] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, Gerard Pujadas, Molecular fingerprint similarity search in virtual screening, Methods, Volume 71, 2015, Pages 58-63, ISSN 1046-2023, https://doi.org/10.1016/j.ymeth.2014.08.005.

[17] Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics. 2018 Dec 31;19(Suppl 19):526. doi: 10.1186/s12859-018-2523-5. PMID: 30598075; PMCID: PMC6311897.

[18] Gan, Jh., Liu, Jx., Liu, Y. et al. DrugRep: an automatic virtual screening server for drug repurposing. Acta Pharmacol Sin 44, 888–896 (2023). https://doi.org/10.1038/s41401-022-00996-2

[19] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015 Jan;71:58-63. doi: 10.1016/j.ymeth.2014.08.005. Epub 2014 Aug 15. PMID: 25132639.

[20] Ani, R. and Deepa, O. S., "Graph Convolutional Neural Network-Based Virtual Screening of Phytochemicals and In-Silico Docking Studies of Drug Compounds for Hemochromatosis", ¡i¿IEEE Access¡/i¿, vol. 11, IEEE, pp. 138687–138698, 2023. doi:10.1109/ACCESS.2023.3338735.

[21] Moriwaki, H., Tian, YS., Kawashita, N. et al. Mordred: a molecular descriptor calculator. J Cheminform 10, 4 (2018). https://doi.org/10.1186/s13321-018-0258-y