**Ministry of Higher Education and
Scientific Research
University of Diyala
College of Science
Department of Computer Science**

Hybrid Feature Selection and Deep Learning Models for Drug Discovery

A thesis

Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Marwa Abdul Kareem Dawood

Supervised By

Prof. Dr. Jumana Waleed Salih

2025 A.C ***1447 A.H***

Abstract

Drug discovery is a highly intricate and interdisciplinary process that encompasses multiple stages, including target identification, compound screening, optimization, and preclinical testing. Traditional methods, while effective, are often hindered by extensive timelines, high costs, and elevated risks of late-stage failures. These challenges are primarily due to the difficulty of accurately predicting critical molecular properties such as aqueous solubility, bioavailability, and pharmacological efficacy. Consequently, there is a growing demand for advanced computational frameworks capable of accelerating early-stage drug discovery while guaranteeing precision and reducing resource consumption.

With the rapid progress of artificial intelligence (AI), especially deep learning (DL) and machine learning (ML), computational methods have become powerful tools for managing the extensive and complex biomedical data associated with drug discovery. These approaches accelerate compound screening and uncover hidden patterns in biological and chemical data that are otherwise difficult to detect with traditional techniques.

This thesis offers a hybrid approach that combines sophisticated DL models with reliable ML-based feature selection to increase drug discovery's predicted accuracy and efficiency. The process starts with thorough data preprocessing, which includes data partitioning, normalization, and missing value imputation. To find the most important molecular descriptors, a number of feature selection approaches are used, including Random Forest (RF), Lasso Regression, Extreme Gradient Boosting (XGBoost), and Mutual Information (MI). This reduces redundancy and dimensionality while maintaining pertinent information.

A number of DL structures are used to model and forecast molecular properties after feature selection. These include Gated Recurrent Units (GRU) for computationally efficient sequence modeling, Long Short-Term Memory networks (LSTM) and Bidirectional LSTM (Bi-LSTM) for capturing long-range dependencies in molecular representations, and one-dimensional convolutional neural networks (1D-CNN) for feature extraction from sequential data. To combine the benefits of deep temporal learning with ensemble feature selection, a hybrid XGBoost-LSTM model is also proposed.

Two benchmark datasets were used in the experimental evaluation: the Directory of Useful Decoys, Enhanced (DUD-E), which offers a large selection of inactive and active compounds for virtual screening, and the Delaney dataset, which focuses on aqueous solubility. The outcomes show that the suggested XGBoost-LSTM model continuously performed better than the other models that were evaluated. It significantly outperformed baseline models without feature optimization, achieving an MSE of 0.4353 on the Delaney dataset and 0.0149 on the DUD-E dataset. Additional comparisons showed that adding feature selection techniques significantly improved prediction accuracy, reaffirming the significance of dimensionality reduction in drug property modeling.

In conclusion, the findings of this study declare that the proposed hybrid framework constitutes a reliable, scalable and cost-effective solution for modern pharmacological treatment ducts. By bridging the gap between computational predictions and experimental validation, this work contributes to the development of safer, more effective, and timely pharmaceutical therapies.

Chapter One

General Introduction

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Introduction

Drug discovery in the biological sciences is a vital even so complex process that aims to develop unique, effective, and safe therapeutic agents. Commonly, it requires over a decade and billions of dollars, making it both temporally intensive and financially [1]. Conventional methods, which heavily depend on clinical trials and laboratory experiments, face challenges such as long timelines, high costs, high failure rates, and the inability to process vast chemical and biological datasets manually. This emphasizes the urgent need for advanced data analysis tools to handle such complexity efficiently [2].

Recent technological advances and access to large-scale biomedical databases have positioned deep learning (DL) and machine learning (ML) as the main methods for accelerating drug discovery while reducing costs. By exploiting computational power, these methods can predict novel compounds, drug–target interactions, and uncover hidden biological patterns. Their application spans multiple stages, from target identification to clinical trials [3].

Artificial intelligence, particularly deep neural networks (DNNs), enhances major stages of drug discovery, including target identification, binding affinity prediction, toxicity and side-effect estimation, and de novo molecule design. ML methods used include supervised, unsupervised, semi-supervised, and reinforcement learning, all of which improve drug activity prediction by modeling complex nonlinear relationships in biomedical data [4][5].

A critical step in DL-driven drug discovery is molecular representation, achieved through SMILES strings, molecular graphs, and molecular fingerprints, which significantly affect model accuracy and reliability [1]. Beyond molecular property prediction, AI contributes to tasks such as drug repurposing, interaction prediction, toxicity analysis, and the use of knowledge graphs for improved drug–drug interaction modeling, thus enhancing patient safety and treatment outcomes [6][7].

Overall, DL-based approaches represent a promising direction to accelerate drug development, reduce costs, and enable the creation of safer and more effective therapies for complex diseases [4].

1.2 Related Works

The use of sophisticated drug discovery systems is inevitable given the inherent difficulties in improving and streamlining the process of successfully and efficiently identifying possible medications. In the past few years, DL techniques have been widely used for a variety of drug discovery phases, including target identification; lead compound optimization, and toxicity prediction. By revealing encouraging outcomes and an improved repository of methodologies, these advancements have significantly increased the body of knowledge in the field of pharmaceutical sciences[4].Comparisons between several relevant drug discovery predictions are depicted in Table (1.1).

- ***Ramsundar***[8]***, 2018,*** demonstrated DL drug discovery with the Deep Chem library on the Delaney and DUD-E datasets. DUD-E is a large dataset of active and inactive molecules against a wide range of drug targets, whereas Delaney is largely derived from molecular solubility descriptors. Over 200 molecular features were calculated, such as physicochemical properties, i.e., molecular weight, number of bonds, and electronic descriptors, which presented comprehensive information regarding each molecule. (DNNs), defined as having more than one layer, were used for the prediction of solubility and classification of compounds based on their pharmacological activity. The results showed that the models performed satisfactorily, as evidenced by RMSE scores between 0.56 and 0.65 on the Delaney test set AUC scores ranging from 0.80 to 0.85 on the DUD-E test set and. One of its most reliable features might be its ability to handle complex and diverse datasets with notable advancements in molecular representation. However, the deep network's inherent complexity makes result analysis more difficult. For example, it requires large and diverse training data sets to avoid bias and overfitting.

- ***Chen et al.*** [9]***, 2019,*** The issue of intrinsic bias in the popularly utilized DUD-E dataset, utilized to train deep models in drug discovery, has been the focus of numerous research papers. It was found during the work that some biases would lead to biased evaluation of the performance measures of the models, particularly the high similarity between active and inactive compounds. Over 100 physical and chemical molecular properties were employed in the study. The results indicated that AUC values for several models fell from around 0.9 to 0.75 after removing these biases, which could imply that their performance might not generalize over disparate datasets in the future. This framework prioritizes bias estimates and data quality checks at the top for the validity assurance of model estimates. The primary contribution is a caution against excessive reliance on DUD-E without adequate verification; this line of reasoning wills certainly invite further testing and model-building in future studies.
- **Cleves and Jain**[10]**, 2020,** used the DUD-E dataset to create a virtual screening model that combines ligand-based molecular descriptions with structural features of the protein targets. Interestingly, every feature that fell between 150 and 200 was not included, including molecular

descriptors and general binding pocket information. The AUC values, which range from 0.85 to 0.88 for various drug targets under test conditions, demonstrate the model's ability to discriminate between active and inactive molecules. The study demonstrated how predictive performance is greatly improved when ligand characteristics and protein spatial structure analysis are combined, as opposed to when only one information source is used. One of the main advantages of this strategy is that it improves knowledge of molecular interactions, which facilitates the search for more efficient drug candidates. However, the protein binding pocket must be precisely depicted in accordance with the method; any variation in this region may result in a reduction in effectiveness.

- ***Shaikh et al.***[11]***, 2021,*** offered a multifaceted framework for the prediction of small molecule and biologics activity through the integration of structural analysis and ML methods. Irrespective of uncertainty regarding the number of properties covered, the framework utilized the DUD-E and Delaney datasets, which were retrieved as a composite set of molecular and structural features. The models achieved a high level of precision in active compound identification, with rates varying from 85% to 90% and AUC scores varying from 0.88 to 0.92. A key strength of the framework is that it can support multidimensional data since it considers both protein and ligand data, thereby enhancing predictive precision and lowering the incidence of false positives. The expensive computations, which entail high levels of computing resources and large amounts of training data, would restrict its usage in resource-constrained settings.
- ***Bontha et al.***[12]***, 2021,*** also reported reinforcement learning (RL)-based approaches to de novo drug-like molecule design on Binding-DB and the Delaney dataset. To facilitate the design of molecules with desired

properties, they prioritized the ESOL solubility score while calculating over 50 physicochemical descriptors. The RL model exhibited a consistent quantitative solubility prediction with an RMSE value ranging from 0.5 to 0.6. This model facilitates automatic generation of new compounds with optimally preferred properties, thereby speeding up the drug discovery process through the elimination of lengthy laboratory testing. The main challenges are maintaining control over certain molecular properties during the generation process, e.g., effectiveness and producibility, as well as training the reinforcement learning algorithms. Despite these limitations, this study illustrates how the accuracy and productivity of molecular design can be enhanced through the integration of chemical data and reinforcement learning, opening up new possibilities for computational drug discovery.

- ***Chen and Tseng*** [13]***, 2021,*** investigated the use of diverse molecular representations, in this instance via SMILES strings, to improve the performance of DL models for the prediction of compound solubility. The authors used a unique Convolutional Sequence-to-Sequence (ConvS2S) model with an attention mechanism to help the model recognize important patterns in the various chemical representations. Tens of thousands of text-based representations were intended to be produced for each molecule to give the model a diverse set of viewpoints from which to learn about chemical characteristics. As a result, when the model was given a whole collection of SMILES strings—which contain 30,000 representations for each molecule—it trained noticeably better. The mean squared error (MSE) decreased from 2.606 when there were individual representations to 0.310. The initial model R2 was 0.87; it is currently 0.96. Diversity of molecular representations is what this study is about, which can help deep models learn chemical compounds more effectively. However, since there are a lot of entries, there are computational issues as well as data size expansion.

- ***Wang et al.***[14]***, 2023,*** explored the application of graph-based molecular representations, in which atoms are nodes and bonds are edges, for the prediction of molecular properties using Graph Neural Networks (GNNs). The model learned varying numbers of features based on the complexity of graphs in the Delaney and DUD-E datasets. On the DUD-E dataset, this study demonstrated state-of-the-art AUCs of 0.88 to 0.92, and on the Delaney dataset, RMSE of around 0.5 for solubility prediction. The most significant benefit of GNNs is the capability to capture intricate structural and relational information in molecules not captured by traditional descriptors. The biggest drawback is the high computational cost brought on by the significant hardware and training time requirements. This work recognizes the potential of graph-based deep learning methods for boosting chemical structure representation and prediction model precision in drug development operations.
- ***Atasever***[15]***, 2024,*** presented a general review of ML-based drug discovery applications with a specific focus on studies that used the DUD-E dataset. The research encompassed a broad area of ML, from simple algorithms up to deep multi-layered DL structures, assessing over 100 automatically calculated or manually designed molecular features. Findings showed that system performances typically reported AUC between 0.8 and 0.9, with significant heterogeneity based on feature extraction and algorithm choice. The essential problem of dataset bias was also seen in this study, with an adverse impact on system reliability and generalizability .The study is also substantial in the sense that it is far-reaching, encompassing a wide synthesis of rising trends and challenges in AI-assisted drug discovery. But it also brought to light the

pressing need for improved data curation and debasing strategies to enhance system robustness and real-world applicability.

- ***Katsamakas et al.***[16]***, 2024,*** applied AI to the virtual screening process to discover encouraging inhibitors against the COVID-19 virus with the DUD-E dataset. The system was trained on more than 180 molecular features, ranging from chemical, physical, to geometric properties. Their AI-based solution demonstrated excellent performance with AUC values above 0.9. Although the system was effective and efficient, it is presently optimized for compounds with COVID-19, thus restricting its direct use in other therapeutic areas. Yet, the research demonstrates the potential of AI to largely streamline drug discovery processes under stringent conditions, with an ability to create a platform for real-time, rapid pharmaceutical innovation in the future.
- ***Lu et al.*** [17]***, 2024,*** proposed a multimodal DL system to fuse heterogeneous molecular representations to enhance drug property prediction accuracy. The proposed system includes three main processing channels; a Transformer-Encoder for investigating textual SMILES representations, a Bidirectional Gated Recurrent Unit network for processing chemical Extended Connectivity Fingerprints, and a Graph Convolutional Network (GCN) for comprehending intricate graph-based molecular structures . These data from these modalities were fused through statistical approaches like LASSO and Random Forest to maximize ultimate prediction accuracy. Several datasets like Delaney, Lipophilicity, and SAMPL were used for validation. The findings demonstrated a clear performance enhancement where the model realized at least an RMSE of 0.620 and a Pearson correlation coefficient of up to 0.96, demonstrating good predictive capabilities. This research recognizes the advantages of integrating textual and structural representations, which complement each other in the pursuit of a more

profound comprehension of chemical compounds. The main challenges are the model's complexity and the extensive computational resources involved, so a balance between accuracy and efficiency must be carefully struck.

Table (1.1): Comparison of several drug discovery prediction systems.

Authors, Year, Ref.	**ML & DL Schemes**	**Datasets**	**Advantages**	**Disadvantages**
Ramsunda, 2018 [8]	Deep CNNs	Delaney, DUD-E	Handles complex d atasets, advanced m olecular representat ion	Difficult result interpretation, requir es large training data
Chen et al., 2019 [9]	DL models with bias analysis	DUD-E	Warns about data bias, improves mod el credibility	Challenge in creating balanced and realistic datasets
Cleves & Jain, 2020 [10]	Virtual screening model	DUD-E	Better accuracy by integrating molecul ar and protein features	Highly dependent on accurate protein binding site representation
Shaikh et al., 2021 [11]	Integration of ML and structural analysis	DUD-E, Delane y	Integrates ligand and protein info for improved accur acy	High computational cost, requires large training sets
Bontha et al., 2021 [12]	(RL) for De novo molecule design	Delaney, Binding DB	Automated new molecule design, speeds drug discovery	Training difficulty, control ling specific molecular pro perties
Chen & Tseng, 2021 [13]	ConvS2S model with attention on SMILES representations	Delaney	Improved predictio n performance, reduced MSE	Increased data size and computation due to multiple Representations

Wang et al., 2023 [14]	GNNs	Delaney, DUD-E	Captures complex molecular structure, improves accuracy	High computational cost, long training time
Atasever, 2024 [15]	ML applications	DUD-E	Identifies bias challenges	Need for improved data curtain and debasing
Katsamakas et al.,2024[16]	AI-based virtual screening	DUD-E	Highly demonstrating AI effectiveness under urgent pandemic scenarios	Limited applicability to non-COVID-19 therapeutic areas
Lu et al., 2024 [17]	Multimodal DL	Delaney, Lipophilicity, SAMPL	Combines different representations, improves accuracy	Model complexity, high computational requirements

Most previous studies relied on a single representation of molecules or deep learning models without effectively integrating feature selection techniques, limiting the accuracy of drug property prediction and their ability to handle complex data. This highlights the need for more efficient hybrid frameworks that combine both approaches.

1.3 Problem Statement

The recognition of the problem in drug discovery utilizing AI methodologies echoes the immense challenges of conventional discovery approaches such as high expenses, lengthy timelines, and high failure rates at later clinical trial stages. The process of discovering a novel pharmacological agent is time-consuming. It requires significant financial resources, heavily relying on expensive laboratory experiments and clinical studies that can span many years, with a high likelihood of failure. This is a setback to the fast release of efficient drugs on the market. Therefore, there is a pressing need to apply contemporary approaches founded on AI, more precisely ML and DL , that leverage large volumes of data and considerable computing power to interpret complicated chemical and biological information with high speed and precision. These methods allow for the prediction of drug interactions with bio-targets, the design of novel drug molecules with enhanced properties, and the initial estimation of toxicity and side effects, thereby minimizing the risks associated with clinical trials and conserving time and resources.

The challenge lies inherently in the creation of a sophisticated and effective framework that is capable of scanning and modeling intricate chemical and biological data sets systematically, identifying the most salient features among numerous molecular descriptors, and subsequently building predictive models that can generalize well for the discovery of potential drugs quickly and effectively , transcending conventional techniques. The difficulty is in combining feature selection methods with DL architectures to enhance predictive accuracy, decrease model complexity, as well as lower the risk of overfitting.

The problem as well entails handling missing data, preprocessing it, scaling measurements, and data transformation to make it acceptable for computational models, along with providing a proper testing and training set the model using proper measures like mean squared error (MSE) and coefficient of determination (R^2) to ensure result quality.

1.4 Aim of Thesis

The general aim of this thesis is to achieve greater predictive accuracy, accelerate and improve drug discovery processes, lower costs, eventually assisting medical research and the pharmaceutical sector in effectively and safely treating various diseases. The main objectives of the suggested drug discovery system are as follows:

1- To create a common framework that employs molecular data to forecast important pharmacological properties, for example, solubility.

2- Conducting extensive experiments over benchmark drug discovery datasets (e.g., the Delaney and DUD-E dataset) for validating the effectiveness of the proposed framework in drug property prediction and accelerating the drug development process.

3- Proposing new approaches for molecular feature selection that enhance prediction accuracy and model interpretability, while implementing and comparing multiple methods to identify the most important molecular descriptors that contribute to improving model performance.

1.5 Thesis Outlines

The first chapter gives a short overview of the drug discovery process and emphasises the urgent need to use ML and DL methods in molecular data analysis. It also talks about recent work on using DL to predict drug properties, the problem statement, and the primary goal of the thesis. The rest of the chapters cover the following:

Chapter Two: This chapter outline the theoretical and technical framework, explaining the basics of machine learning and deep learning, as well as feature selection methods and the most important algorithms used in this field.

Chapter Three: The subject of this study presents the architecture and implementation of the proposed hybrid prediction system. It includes detailed methodology for dataset pre-processing, feature selection strategies, and the design of DL models such as 1D-CNN, LSTM, BI-LSTM and GRU.

Chapter Four: offers system implementation details and experiments conducted on benchmark datasets with model performance evaluation based on precise metrics such as mean squared error and coefficient of determination, and chronicles the effectiveness of the proposed system in improving prediction accuracy and reducing time and cost in drug discovery.

Chapter Five: Condenses this work and depicts its future extent.