Major Project Synopsis

on

DRUG TARGET INTERACTION PREDICTION

COMPUTER ENGINEERING

Submitted by

CHAITREYA SHRAWANKAR

DEEP KHAUT

OMKAR INGOLE

SHREEYASH GAIKI

Under the guidance of

DR. DIPAK W. WAJGI
Associate Professor

Academic Year 2024-25

Department of Computer Engineering

ARISE & SHINE

ST. VINCENT PALLOTTI COLLEGE OF ENGINEERING AND TECHNOLOGY

Wardha Road, Gavsi Manapur, Nagpur

(An Autonomous Institute affiliated to RTMNU, Nagpur)

CONTENTS

ABSTRACT

The identification of drug-target interactions is a crucial step in drug discovery and development [2]. Experimental methods for identifying these interactions are often costly, time-consuming, and labor-intensive. Computational approaches, such as the one implemented in this project, provide a faster and more efficient alternative to experimental techniques. By predicting potential drug-target interactions, such models can support the early stages of drug discovery and reduce the reliance on extensive biological testing [2].

This project focuses on building a deep learning-based model to predict drug-target interaction (DTI) binding affinities. The model is designed using a Multilayer Perceptron (MLP) and uses 1D sequence data from proteins and drugs as input. Drugs are represented using their SMILES strings, which describe their molecular structures, while proteins are represented using their amino acid sequences. The goal is to predict the binding affinity between a given drug and target protein based on these inputs.

The MLP model developed in this project captures patterns in drug and protein sequence data and predicts their interaction strength. This project uses pre-processed numerical representations of drug and protein sequences to train the model. The aim is to contribute to the growing field of computational drug discovery by providing a tool to assist in identifying promising drug candidates for specific targets.

PROBLEM STATEMENT AND OBJECTIVES

The project aims to develop a machine learning model to predict drug-target binding affinity using their 1D representations. This involves utilizing SMILES strings for drugs and amino acid sequences for proteins as inputs to train a predictive model, enabling efficient and accurate identification of potential drug-target interactions.

OBJECTIVES

- Develop a machine learning model to predict potential interactions between drugs and human protein targets using molecular fingerprints and sequence embeddings.
- Collect and integrate datasets, including drug properties, protein sequences, and experimental binding affinity data, to train and evaluate the model.
- Accelerate drug discovery by identifying approved drugs with potential efficacy against emerging or re-emerging diseases.
- Minimize research costs by reusing existing preclinical and clinical trial data, reducing reliance on expensive initial phases.
- Enhance efficiency in resource allocation by focusing on high-potential compounds, streamlining the drug discovery process.

LITERATURE REVIEW

The prediction of drug-target binding affinity is a critical challenge in drug discovery, as it helps identify how strongly a drug interacts with a protein, which directly impacts its therapeutic effectiveness. Computational methods for drug-target interaction (DTI) prediction have evolved significantly over the years, transitioning from traditional statistical approaches to modern deep learning-based models [2].

DeepDTA: Hakime Öztürk et al. (2018) proposed DeepDTA, a pioneering model in DTI prediction. It utilized convolutional neural networks (CNNs) to process the 1D sequence information of drugs (SMILES strings) and proteins (amino acid sequences) [1]. This approach moved away from reliance on 3D structural data, leveraging simpler representations to predict continuous binding affinity values. DeepDTA was evaluated on the **Davis** and **KIBA** datasets, achieving Mean Squared Error (MSE) scores of 0.262 and 0.194, respectively [1]. Its success highlighted the potential of deep learning for DTI prediction, outperforming classical methods like KronRLS and SimBoost.

The **Davis dataset** comprises 68 drugs and 442 proteins, with 30,056 drug-target interaction pairs measured using dissociation constants (Kd).
Other approaches, such as **WideDTA** and **GraphDTA**, have extended the principles introduced by DeepDTA. WideDTA integrates text-based protein representations with SMILES features, while GraphDTA employs graph convolutional networks for more detailed molecular modeling. These studies emphasize the importance of data representation and encoding in DTI prediction [3].

In this project, we leverage the findings of DeepDTA but adapt the architecture to use a Multilayer Perceptron (MLP) model. This approach simplifies computational requirements while maintaining robust predictive performance. By integrating molecular fingerprints for drugs and sequence embeddings for proteins, our work aims to make DTI prediction more accessible and efficient.

SOFTWARE TECHNOLOGY

Python is a versatile programming language with extensive libraries for machine learning, data manipulation, and visualization.

RDKit provides tools for processing chemical information, including generating Morgan fingerprints from SMILES strings. This is critical for representing drug molecules as numerical vectors.

Hugging Face library offers state-of-the-art models like ProtTrans (ESM2), which efficiently encode amino acid sequences into meaningful numerical embeddings. **facebook/esm2_t12_35M_UR50D** a pre-trained model specifically designed for encoding protein sequences, reducing the need for training sequence-specific embeddings from scratch.

Keras offers a high-level API for building and training deep learning models, such as the Multilayer Perceptron (MLP) used in this project.

Pandas used for loading and preprocessing datasets, including drug-target interaction data, mapping IDs, and managing input files.

NumPy provides efficient array operations essential for handling drug and protein feature vectors during model training and prediction.

Matplotlib is used for plotting graphs such as loss curves, mean squared error (MSE), and accuracy trends during model evaluation and training.

Scikit-learn provides tools like mean_squared_error for evaluating the model's performance on testing data.

HARDWARE TECHNOLOGY

CPU intel i5/ ryzen 5 or above for handling preprocessing tasks such as generating molecular fingerprints, tokenizing protein sequences, and loading datasets efficiently.

GPU RTX 3050 accelerates the training of deep learning models and the generation of protein sequence embeddings using transformer-based models like ProtTrans.

RAM min 8GB Necessary for loading and processing datasets, storing embeddings, and running memory-intensive transformer models.

Storage required for storing datasets, pre-trained models, and intermediate files like drug/protein encodings and checkpoints.

OS windows or linux Compatible with deep learning frameworks and essential tools

EXPECTED OUTCOME

Accurate prediction of Drug-Target Binding Affinities

The machine learning model will reliably predict binding affinities for given drug and protein pairs based on their molecular fingerprints and sequence embeddings.

Reduction in Drug Discovery Time

By identifying high-potential drug-target interactions computationally, the project will significantly reduce the time required for experimental screening.

Cost-Effective Drug Discovery

The model will minimize the reliance on expensive laboratory experiments, making drug discovery more economical and accessible.

Enhanced Understanding of Drug-Protein Interactions By analyzing model outputs, researchers can gain insights into how drugs interact with specific protein targets, contributing to better therapeutic design.

Drug Repurposing Opportunities

The project may help identify existing drugs with potential efficacy against new or re-emerging diseases, aiding in rapid response during public health crises.

Efficient Resource Allocation

The model will prioritize high-potential drug candidates, allowing researchers to focus resources on the most promising compounds, reducing waste and inefficiencies in drug development.

REFERENCES

1. Hakime Ozturk, Arzucan Ozgur and Elif Ozkirimli, “DeepDTA: deep drug–target binding affinity prediction,” Department of Computer Engineering and Department of Chemical Engineering, Bogazici University, Istanbul 34342, Turkey. Available at: https://academic.oup.com/bioinformatics/article/34/17/i821/5093245

2. Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, Jimeng Sun, “DeepPurpose: a deep learning library for drug–target interaction prediction Bioinformatics, Volume 36, Issue 22-23, December 2020, Pages 5545–5547. Available at: https://academic.oup.com/bioinformatics/article/36/22-23/5545/6020256

3. Xing Chen, Chenggang Clarence Yan, Xiaotian Zhang, Xu Zhang, Feng Dai, Jian Yin, Yongdong Zhang, “Drug–target interaction prediction: databases, web servers and computational models,”Briefings in Bioinformatics, Volume 17, Issue 4, July 2016, Pages 696–712. Available at: https://academic.oup.com/bib/article/17/4/696/2240330

4. Tudor I. Oprea, Julie E. Bauman, Cristian G. Bologa, Tione Buranda, “Drug repurposing from an academic perspective.” Christopher A. Lipinski – Scientific Advisor, Melior Discovery, Waterford, CT 06385-4122, USA. Availabe at: https://www.sciencedirect.com/science/article/abs/pii/S1740677311000428

5. Davis dataset for drug-target interaction prediction Available at: https://github.com/dingyan20/Davis-Dataset-for-DTA-Prediction

Roll No	Name of the student	Signature
	Chaitreya Shrawankar	
	Deep Khaut	
	Omkar Ingole	
	Shreeyash Gaiki	

Date :

Name of the guide : Dr. Dipak W. Wajgi

Signature with date :

Remarks :

Sign of HOD : Date:

Remarks :