成都理工大學中英合作办学
CDUT Sino-British Collaborative Education

OXFORD
BROOKES
UNIVERSITY

UNDERGRADUATE PROJECT REPORT

Project Title:	**Explainable Deep Learning for Drug-Target Interaction Prediction Using Protein Structures**
Surname:	**Liu**
First Name:	**Kerui**
Student Number:	**202118010317**
Supervisor Name:	**Grace Ugochi Nneji**
Module Code:	**CHC 6096**
Module Name:	**Project**
Date Submitted:	**May 6, 2024**

Chengdu University of Technology Oxford Brookes College

Chengdu University of Technology

BSc (Single Honours) Degree Project

Programme Name: **Computer Science**

Module No.: **CHC 6096**

Surname: Liu

First Name: Kerui

Project Title: Explainable Deep Learning for Drug-Target Interaction Prediction Using Protein Structures

Student No.: 202118010317

Supervisor: Grace Ugochi Nneji

2ND Supervisor (if applicable): **Not Applicable**

Date submitted: **May 6, 2024**

A report submitted as part of the requirements for the degree of BSc (Hons) in Computer Science

At

Chengdu University of Technology Oxford Brookes College

Declaration

Declaration

Student Conduct Regulations:

Please ensure you are familiar with the regulations in relation to Academic Integrity. The University takes this issue very seriously and students have been expelled or had their degrees withheld for cheating in assessment. It is important that students having difficulties with their work should seek help from their tutors rather than be tempted to use unfair means to gain marks. Students should not risk losing their degree and undermining all the work they have done towards it. You are expected to have familiarised yourself with these regulations.

https://www.brookes.ac.uk/regulations/current/appeals-complaints-and-conduct/c1-1/

Guidance on the correct use of references can be found on www.brookes.ac.uk/services/library, and also in a handout in the Library.

The full regulations may be accessed online at https://www.brookes.ac.uk/students/sirt/student-conduct/

If you do not understand what any of these terms mean, you should ask your Project Supervisor to clarify them for you.

I declare that I have read and understood Regulations C1.1.4 of the Regulations governing Academic Misconduct, and that the work I submit is fully in accordance with them.

Signature Liu Kerui Date 4/5/2025

REGULATIONS GOVERNING THE DEPOSIT AND USE OF OXFORD BROOKES UNIVERSITY MODULAR PROGRAMME PROJECTS AND DISSERTATIONS

Copies of projects/dissertations, submitted in fulfilment of Modular Programme requirements and achieving marks of 60% or above, shall normally be kept by the Oxford Brookes University Library.

I agree that this dissertation may be available for reading and photocopying in accordance with the Regulations governing the use of the Oxford Brookes University Library.

Signature Liu Kerui Date 4/5/2025

Acknowledgment

First, I sincerely thank my supervisor, Dr. Grace Ugochi Nneji, for giving me precise guidance and continuous encouragement throughout the entire project process. Her professional knowledge in the field of deep learning is of immeasurable help with the method design and result interpretation of this research.

Meanwhile, thanks are given to Dr. Joojo Walker (the person in charge of the CHC 6096 module), as well as all the instructors and counselors. Their classroom lectures and discussions have provided a solid theoretical basis for this research.

I would also like to express my gratitude to Oxford Brookes University and Chengdu University of Technology for the resources and facilities they provide. The superior academic environment and open collaboration space they offer have created favorable conditions for my research work.

Finally, I would like to express my gratitude to all my classmates and friends for their company, constructive criticism and spiritual support during this academic journey. It is precisely because of their help that this project could be completed successfully.

Table of Contents

List of Figures

Abstract

Drug–target interaction (DTI) prediction is a vital component of the drug discovery pipeline, yet traditional similarity-based and laboratory methods face challenges in handling high-dimensional features, nonlinear relationships, and data imbalance. To overcome these limitations, this paper introduces a new hybrid deep learning model with an FCS mining module, Transformer-based encoders, and CNNs to automatically learn chemical substructure and protein sequence features and fuse them. To demonstrate that the model can generalize well, the model was trained and tested on three datasets (BIOSNAP, BindingDB, DAVIS) with an 80:10:10 split. On the BIOSNAP test set, the model is better at training with training AUC 0.9952. It achieved a test AUC of 0.8779, AUPRC of 0.8851, and an F1 score of 0.8046. Besides that, explainability methods such as attention heatmaps and SHAP analyses were used to give insights into important substructures and sequence motifs governing predictions, making the model more interpretable. The interactive web interface facilitated exploratory analysis of DTI predictions and visualization explanations. These results echo the capability of hybridizing substructure mining with high-level deep learning techniques for interpretable and resilient DTI prediction, addressing the computational foundation to facilitate drug discovery.

Keywords: Drug-Target Interaction, Deep Learning, Predictive Modeling, Transformer, Convolutional Neural Network, Explainability, SHAP, Bioinformatics

Abbreviations

DTI: Drug-Target Interaction

FCS: Frequent Continuous Substructure

CNN: Convolutional Neural Network

AUC: Area Under the Curve

AUPRC: Area Under the Precision-Recall Curve

F1: F1 Score

SHAP: SHapley Additive exPlanations

SMILES: Simplified Molecular Input Line Entry System

ROC: Receiver Operating Characteristic

BPE: Byte Pair Encoding

GNN: Graph Neural Network

API: Application Programming Interface

Glossary

Drug-Target Interaction (DTI): The drug-target bio interaction is used in drug discovery to predict the activity and side effects of the drug molecule on a target protein.

Frequent Continuous Substructure (FCS): The project employed a module to extract and identify biologically significant substructures from protein and drug data. The substructures enhance the predictive model's ability to predict interactions.

Convolutional Neural Network (CNN): One of the most widely used deep learning models for grid-like data, such as images or sequence data. CNNs were used in this project to learn local features from drug and protein sequences.

Area Under the Curve (AUC): A performance metric is employed to compare binary classification models. It measures the goodness of a model in classifying positive and negative instances. A higher value for AUC implies a better model.

Area Under the Precision-Recall Curve (AUPRC): A measure that may be employed in the performance measurement of a model, particularly if the dataset is imbalanced. It is concerned with the capacity of a model to classify the positive class.

F1 Score: A calculation of the precision-recall trade-off of a model, especially for imbalanced datasets. It is the harmonic meaning between precision and recall.

SHapley Additive exPlanations (SHAP): A machine learning model output interpretation approach where importance is given to every feature. SHAP values offer model prediction interpretation by indicating the contribution of every input feature.

Simplified Molecular Input Line Entry System (SMILES): A notation that describes chemical structures using a text string. SMILES strings are employed widely to depict the molecular forms of drugs in computational systems.

Receiver Operating Characteristic (ROC): A graphical plot of the performance of a classifier, plotting the actual positive rate against the false positive rate for various classification thresholds.

Byte Pair Encoding (BPE): A subword tokenization method used in this project to split drug and protein sequences into smaller units. This encoding technique helps in handling rare or unseen words in sequence data.

Graph Neural Network (GNN): A type of deep learning model designed to work with graph-structured data. GNNs are often used for predicting relationships in molecular or protein data by representing molecules as graphs of atoms and bonds.

Application Programming Interface (API): A set of tools and protocols used for building software applications. In this project, the API sends messages to the model and obtains its predictions and visualizations.

Chapter 1 Introduction

1.1 Background

Drug-target interaction (DTI) prediction is an important aspect of the drug discovery process. Precise prediction of drug-target interactions is essential to speed up drug development, increase therapeutic effectiveness, and lower side effects [1]. However, before the advent of deep learning algorithms, traditional approaches to drug-target interaction prediction were plagued by many challenges and shortcomings.

1.1.1 Risk and Factor

Even though drug-target interaction prediction is critical during drug discovery, traditional wet-lab-based methodologies are affixed with drastic risks and limitations. The traditional wet-lab experimental techniques such as biochemical assays and high-throughput screening are substantially capitalized on specialized equipment, chemical reagents, and training personnel, in which case validation costs of a single candidate drug can readily exceed tens of thousands of dollars.

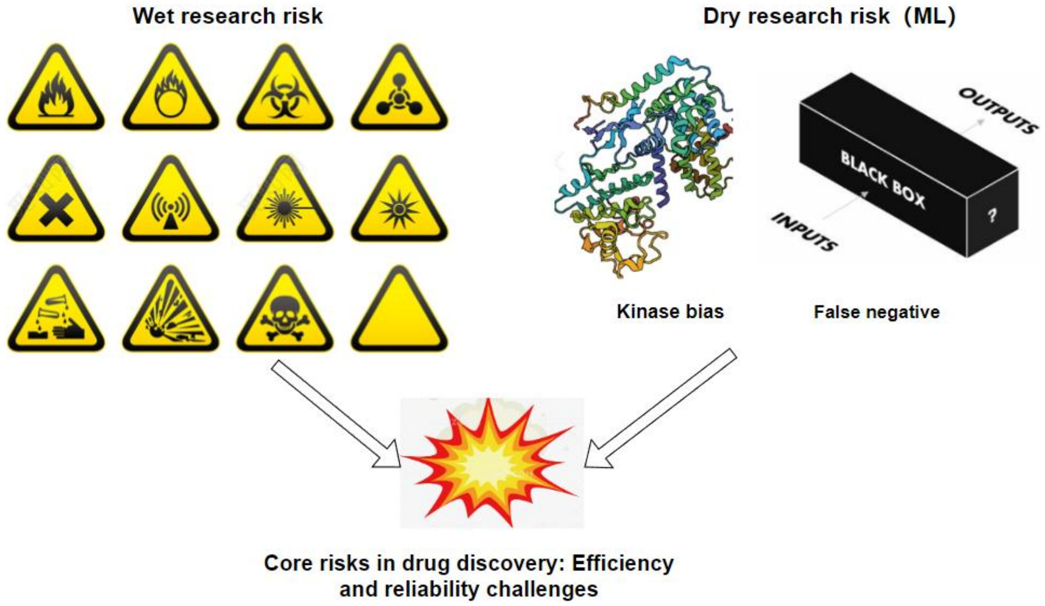


Figure 1: Potential risk in traditional research

Besides that, these processes typically require years or months and only validate a few drug-target pairs. Besides that, laboratory procedures have built-in safety hazards, including exposure to toxic chemicals, biohazard contamination, and device failure, that can incapacitate scientists and render experimental reproducibility inconvenient. For instance, mishandling of reagents can cause systematic errors, and ill-calibrated equipment can make screening data worthless[2]. Figure 1 depicts the danger of traditional methods.

Although machine learning is an effective computational method for drug-target interaction (DTI) prediction, its application is still within the realm of adverse risk factors. First, data bias risks prevail, where most public databases are over-represented towards highly investigated protein families, and prediction accuracy is highly biased toward rare targets[3]. Such a bias can result in potentially active drug-target pairs being wrongly excluded and, hence, new drug discovery disabled. Second, model uninterpretability raises decision-making risks. The "black-box" character of deep learning models (such as CNNs and GNNs) renders predictions not to have mechanistic justification at a biological level, which may result in excessive dependence on statistical correlation over causal relationships and ineffective or risky wet-lab validation designs. Moreover, technology access limitation depends on computational resources equally[4]. Training complex models requires high-performance GPUs and massive storage resources, creating technological barriers for resource-constrained institutions and exacerbating inequalities in research resource allocation. Finally, overfitting risks may foster spurious correlations: when models excessively depend on noise or coincidental patterns in training data, their generalizability in real-world scenarios will sharply decline, ultimately compromising the success rate of drug development.

1.1.2 Challenge

Traditional computational methods primarily relied on the similarity measures between the chemical structure of drugs and the amino acid sequences of targets, calculating the similarity between drugs and targets to infer whether an interaction might occur. In cheminformatics, drug molecules are often represented using chemical fingerprints such as the Extended-Connectivity Fingerprints (ECFP) and Molecular ACCess System (MACCS) keys [5]. These fingerprints encode molecular structures into binary vectors, facilitating computational analysis. In target proteins, sequence alignment techniques are typically used to determine areas of similarity, which aid in comprehending functional, structural, and evolutionary relationships [6].The techniques can shine with trivial examples but reveal deep deficiencies in more difficult tasks.

Firstly, traditional methods are unable to process high-dimensional data. Both drug chemical structure and target amino acid sequence are high-dimensional, and traditional similarity measures (e.g., cosine similarity, Tanimoto coefficient, etc.) can only conduct matching in low feature space. In DTI prediction, traditional methods cannot process high-dimensional data and thus are inappropriate for complex tasks. For instance, Huang et al. (2020) note that molecular representation's high dimensionality hindered existing deep-learning-based DTI prediction models, which may provide less precise and interpretable outcomes [7]. Similarly, Pei et al. (2022) posit the issues given by data paucity and high dimensionality for Drug-Target Affinity (DTA) prediction, which requires models must be able to process such challenges in order to improve prediction accuracy [8]. Traditional methods are anticipated to lag in terms of simulating the complex, non-linear drug-target interaction because they are mainly founded on straightforward similarity calculations that may not be able to capture intricate interactions. For instance, the interaction of some drug classes with some target protein structure domains is of the utmost importance to the activity and selectivity of a drug. However, it is possibly hard to model with conventional methods [9]. Finally, conventional methods usually rely on hand-crafted features, requiring expert-level experience to specify the feature space. Although these methods can perform well on specific datasets, they possess poor generalization ability to handle diverse drug and target pairs. They are additionally rigid and lack practical robustness[10].

Despite all these challenges, slowly, with the success of deep learning techniques in computer vision and natural language processing, these techniques have been used in drug-target interaction prediction. Deep learning techniques can automatically learn high-dimensional features from raw data, uncover subtle non-linear interactions between targets and drugs, and significantly improve the accuracy and reliability of predictions. Deep learning can overcome the limitations of conventional methods, especially in handling complex data, high-dimensional features, and non-linear interactions[11]. Because of this, deep learning models have received significant interest as a research target for next-generation drug-target interaction prediction and introduce new ideas and technical backup to next-generation drug discovery.

1.2 Aim

In Drug-Target Interaction (DTI) prediction, conventional methods usually employ individual models to handle drug and target features separately, and individual models are ineffective in dealing with complex information. In order to solve this problem and improve the prediction quality, the current project will design and develop a novel deep-learning model for accurate DTI prediction. This project proposes a hybrid model combining an FCS mining module, Convolutional Neural Networks (CNN), and a Transformer self-attention mechanism.

Firstly, the Frequent Continuous Substructure (FCS) Mining module separates drugs (SMILES strings) and proteins (amino acid sequences) into biologically significant substructures. Afterward, the model applies CNN to process drug chemical and protein structure information, automatically extracting the local features of drug molecules. After that, the Transformer module constructs context-aware vectors with rich chemical semantics for drug and protein substructures by combining content embeddings, positional encodings, and self-attention mechanisms, capturing complex biochemical correlations among substructures in an efficient manner. This fusion is planned to take the benefits of every one of the three technologies over solitary models, enhancing the reliability and precision of DTI prediction. The project plans to achieve an efficient and more precise drug discovery and target study solution through this model, promoting the further development of DTI prediction technology.

1.3 Objectives

This project aims to develop a drug-target interaction (DTI) prediction model with deep learning that would be capable of accurately predicting target and drug binding affinity, thereby accelerating target identification and drug discovery. The project will utilize multiple drug and target representation methods, combining FCS, CNN and Transformer deep learning models to extract complex features and interactions between drugs and targets.

The project will collect data from public databases such as DAVIS and BindingDB, the FCS Mining Module decomposes drugs (SMILES strings) and proteins (amino acid sequences) into biologically meaningful substructure sequences. Subsequently, the Transformer Module processes these substructure sequences by mapping each discrete symbol into continuous vector representations through an embedding layer. Positional encodings are incorporated to preserve sequential order information, and the self-attention mechanism dynamically captures contextual dependencies between substructures, ultimately generating chemically rich, context-aware embedding vectors. Finally, the Interaction Module performs structured interaction modeling using the drug and protein embedding vectors. A convolutional neural network (CNN) extracts local patterns from the interaction map and outputs the final drug-target interaction probability.

The core goal of this project is to build a hybrid deep learning regression model combining FCS, Transformer and CNN. The model will optimize several hyperparameters, such as batch size, learning rate, dropout rate, etc., with the aim of improving the prediction accuracy of drug-target interactions. Finally, the project will be deployed on a website, allowing users to upload drug and target data (such as SMILES and protein sequences) and receive predicted binding affinity results. This will help accelerate drug discovery and target identification processes.

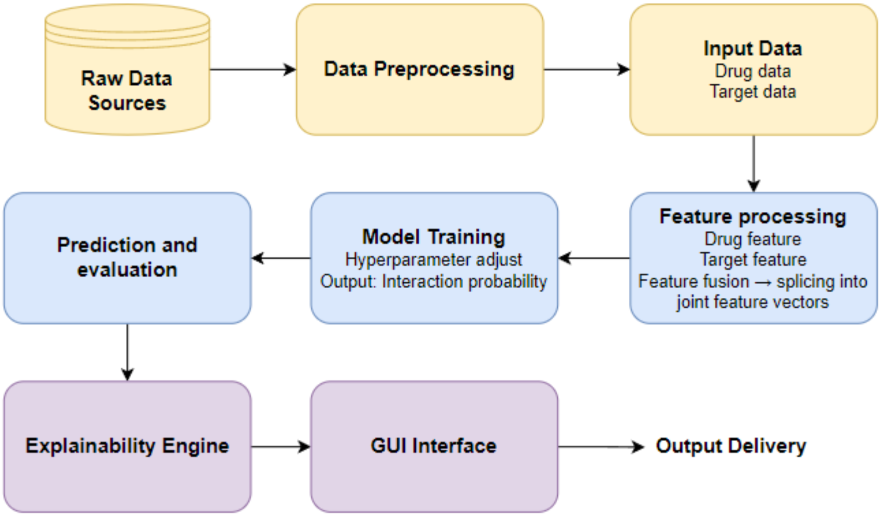


Figure 2: Project Overview

1.4 Project Overview

Deep learning techniques have revolutionized various fields of research, including drug discovery and drug-target interaction (DTI) prediction. Convolutional Neural Networks (CNNs) and Transformers are highly effective in extracting complex patterns from both chemical structures of drugs and biological sequences of targets. However, models that are not tailored to the unique characteristics of drug-target interaction prediction often face challenges, such as overfitting, poor generalization, and inefficiency in handling high-dimensional data.

1.4.1 Scope

By harnessing a high-accuracy drug–target interaction model, this project stands to dramatically accelerate the drug discovery process by enabling rapid, in silico screening and prioritization of candidates, while simultaneously uncovering novel therapeutic targets to broaden treatment options. Early prediction of efficacy and off-target interactions reduces costly late-stage failures and enhances patient safety by flagging potential adverse effects before clinical testing. Moreover, by incorporating patient-specific data, the approach paves the way for truly personalized medicine, tailoring therapies to individual molecular profiles. Short of that, these roles should be capable of enhancing overall healthcare outcomes with better drugs and fewer side effects and maximizing development resource utilization, moving the industry away from the historically trial-and-error approach towards a more cost-effective and practical approach.

1.4.2 Audience

The creation of an advanced drug-target interaction prediction system will significantly benefit the stakeholders involved in the pharmaceutical and healthcare sectors. First, drug companies can make more informed decisions about drug candidates and save time and money along the drug development process by improving the efficiency and accuracy of drug-target interaction prediction. Secondly, researchers will gain access to an advanced tool for analyzing drug-target interactions, enabling them to explore new therapeutic avenues, understand drug mechanisms, and discover novel drug-target pairings. Third, through more accurate prediction of possible drug-target interactions, physicians will make better treatment decisions when prescribing, resulting in better patient outcomes and fewer side effects. Furthermore, improved drug-target interaction prediction can result in the design of safer and more effective drugs, which ultimately will be advantageous to patients by having greater drug efficacy and fewer side effects. Healthcare Facilities like Hospitals and doctors will be positively impacted by a more efficient drug development process, minimizing the gap between drug discovery and clinical use.

In short, The new DTI prediction model is of enormous benefit to pharmaceutical companies, researchers, physicians, patients, and medical organizations. To achieve the highest accuracy and efficacy of drug-target prediction, the model can transform drug discovery, drug development, and patient treatment and enable personalized medicine.

Chapter 2 Background Review

2.1 Traditional methods to solve problems

Drug-target interaction prediction is the central part of drug repurposing and drug discovery. Good drug and target prediction would accelerate the development of new drugs, minimize experimental work and time in validation, optimize therapeutic activity, and limit side effects. The conventional approaches to drug-target interaction prediction tend to use the chemical properties of drugs and the biological attributes of targets and perform similarity comparisons to predict potential interactions. For instance, similar measures based on chemical fingerprints and amino acid sequence similarity were some of the first methods used for this prediction. Nonetheless, these approaches have serious drawbacks when handling high-dimensional, nonlinear relationships and sparse data. For example, most drug-target pairs in DTI databases do not have known interaction information, preventing classical methods from describing well-complicated interaction relationships between targets and drugs. In addition, classical methods are based on handcrafted features, leading to poor generalization performance when encountering complicated biomedical data.

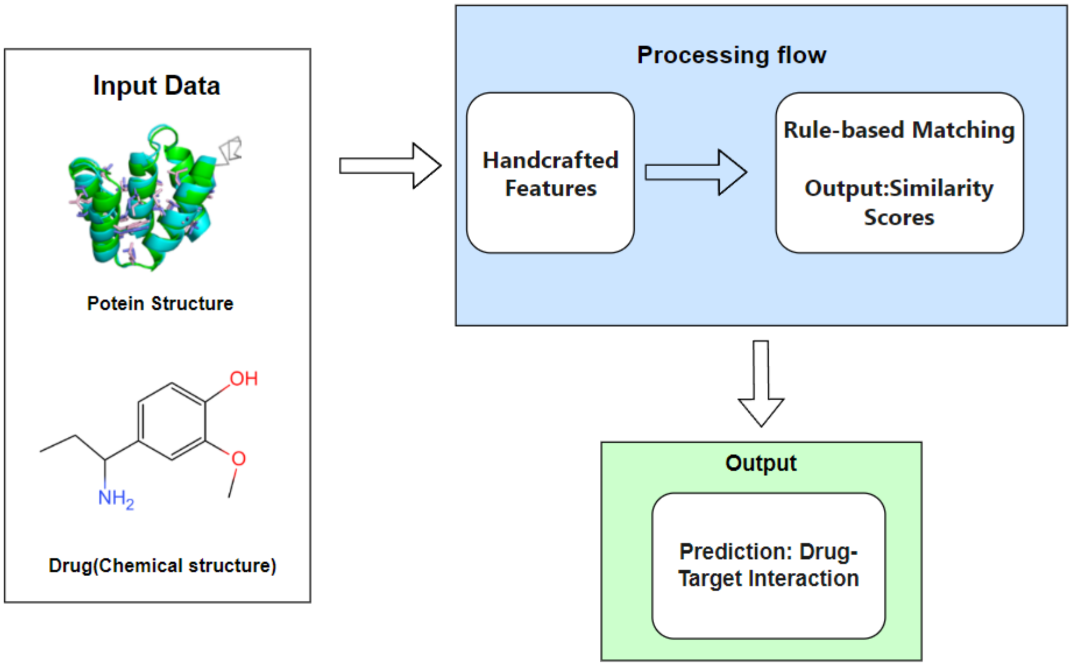


Figure 3: Traditional research method

2.2 Topic-based machine learning methods

In drug-target interaction (DTI) prediction, conventional machine learning (ML) methods have been fundamental in early DTI prediction studies. Support Vector Machines (SVMs), Random Forests (RFs), and Gradient Boosting Machines (GBMs) were used traditionally to predict DTIs from hand-crafted features derived from molecular and biological information[12]. The features are typically physicochemical drug properties, sequence-based target descriptors, and structural fingerprints. While successful, such approaches are limited by their dependency on feature engineering, which is domain-dependent and susceptible to an inability to capture the complex interactions between drugs and targets well.

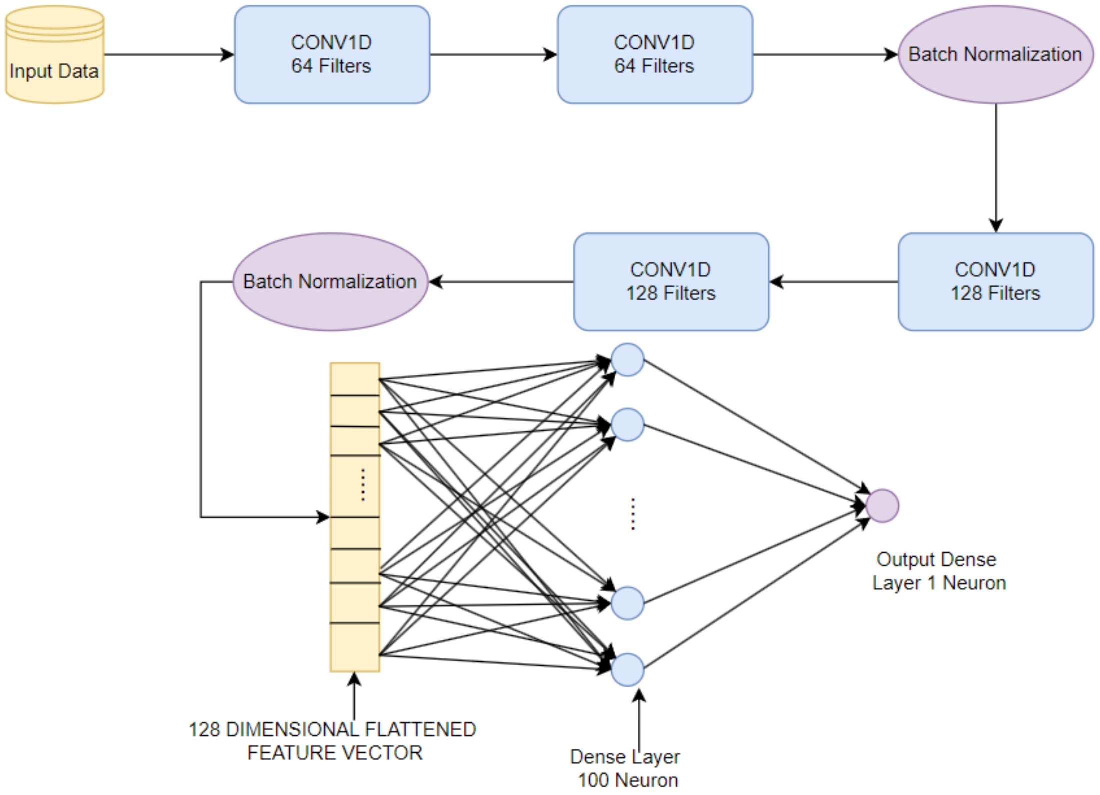


Figure 4: Example of machine learning application in DTI [10]

2.3 Topic-based deep learning approach

However, over the last few years, DL methods have been increasingly used for DTI prediction, and models like CNNs, GNNs, and Transformer models have been proposed, significantly expanding the field of research. Unlike traditional approaches, deep learning models can learn high-dimensional features from primary data and identify intricate nonlinear relationships between targets and drugs[13]. For instance, the DeepDTA model employs convolutional neural networks to learn target sequences' local amino acid sequence patterns and substantially enhance drug-target interaction prediction accuracy[14]. Likewise, Graph Neural Networks (GNNs) depict drug molecules as graph structures, which can effectively learn the interactions among various atoms of a molecule [15]. Apart from that, Transformer models have been broadly used to resolve long-range dependencies between targets and drugs because of their strong self-attention mechanisms.

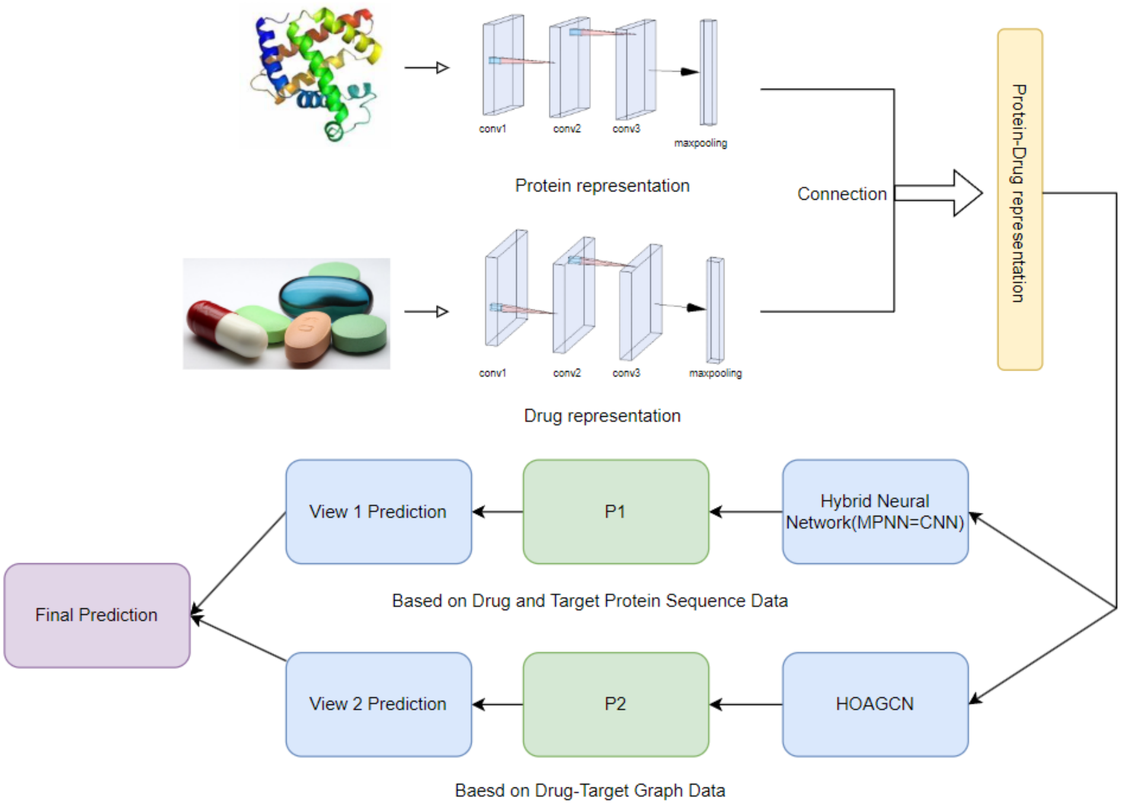


Figure 5: Structure of DeepDTI model [16]

Although drug-target interaction (DTI) prediction models using deep learning have seen improvement, model generalization and sparsity problems remain. Various researchers have addressed these problems through possible solutions like data augmentation and Generative Adversarial Networks (GANs) to create extra DTI data and alleviate the data shortage effect. For example, the paper "GANsDTA: Predicting Drug-Target Binding Affinity Using GANs" proposes a semi-supervised GAN-based method that can learn informative features from labeled and unlabeled data, allowing for improved prediction performance in situations where there is limited labeled data [17].

Overall, with the development of deep learning techniques, significant progress has been made in drug-target interaction prediction research. As more high-quality data accumulates and more advanced deep learning methods emerge, DTI prediction will play an increasingly important role in drug discovery and drug repurpose. A summary of the different researchers and their findings and research results can be found in Table 1.

Table 1: Summary of the Related works

Author	Datasets	Methods & Models	Results	Limitation
Yang et.al[16]	DAVIS, KIBA, BindigDB	Hybrid deep network, High-order graph attention convolutional network, multi-view adaptive integrated decision module	AUPRC: 0.871 ± 0.008 AUROC: 0.851 ± 0.004 Accuracy: 0.831 ± 0.007	Multimodal feature fusion may introduce irrelevant information, requiring more refined weight allocation strategies
Wang. Y and Yin.Z[18]	NR, GPCR, IC	Compare six advanced DTI prediction techniques, DTINet5, DRLSM28, NRMLF29, MSCMF30, NetMF31, and NEDTP32, with FBRWPC	DTINet: 0.902 NetMF: 0.917 NeDTP: 0.929 MSCMF: 0.922 FBRWPC: 0.953	The author emphasizes the shortcomings of the model when comparing with other methods (such as the high false positive rate of DTINet and the coarse-grained feature extraction problem of DDR).

Chen et.al[19]	DAVIS, BIOSNAP, BindigDB	CNN, Morgan and CNN, PubChem and CNN, Self-Attention	CNN:0.82 Morgan and CNN:0.813 PubChem and CNN:0.828	The authors acknowledged the model's suboptimal performance in predicting interactions involving Nuclear Receptor and GPCR families.
Wang et.al[20]	BindigDB, DAVIS	Large-scale Drug target Screening Convolutional Neural Network	AUC: 0.96 AUPRC: 0.95 accuracy: 90.13%	Molecular docking methods require 3D structural information, whereas LDS-CNN relies solely on sequence data. For interactions dependent on structural details (such as allosteric effects), the model may fail to capture critical features.
Li er.al[21]	NR, GPCR, IC, Enzyme	Based on CNN DeepConv-DTI, LSTM-CNN model combined with attention mechanism	NR (AUPR/AUC)：0.84/0.96 GPCR (AUPR/AUC)：0.72/0.93	AutoDTI++ relies on ECFP fingerprints during the preprocessing stage to fill in the missing values of the sparse DTI matrix. If certain drugs lack chemical structure information (e.g., novel compounds) or if the fingerprint representations are inaccurate, this may lead

				to biases in feature learning.
Majumdar er.al[10]	KIBA	An architecture based on 1D convolutional neural networks (CNNS) specifically designed to predict drug-target interaction (DTI) values.	MSE: 0.70 MAE: 0.63	The model was trained using only 19,000 entries from the KIBA dataset (the original dataset contains 118,254 entries). This limited data scope, constrained by computational resources, may affect model training.
Sun et.al[22]	DAVIS, KIBA, BIOSNAP	Nested graph neural networks (NGNN): Enhanced local structure representation through k-hop subgraph pooling. Cross-attention Module (Cross-AFT): Captures information about drug interactions with target substructures.	AUROC(Davis):0.931 AUROC(KIBA): 0.915 AUROC(BIOSNAP): 0.934	The high proportion of negative samples (>90%) in Davis and KIBA may introduce bias and require more complex sampling strategies or loss function optimization.

Chapter 3 Methodology

This chapter will introduce the basic techniques used to implement the proposed model, including dataset information, data preprocessing, the architecture of proposed model and the settings of training hyperparameters and other information about model implementation.

3.1 Approach

This project tackles binary drug–protein interaction prediction using the BIOSNAP dataset, which provides labeled pairs of drug SMILES strings and protein amino-acid sequences. We split the data into training, validation and test sets, ensuring a balanced mix of interacting and non-interacting examples in each. The prediction model builds on the BIN_Interaction_Flat framework but replaces the original DenseNet component with a lightweight 2D convolutional layer. First, each SMILES and protein sequence is decomposed into high-frequency substructure tokens via FCS + BPE and mapped to indices and masks. These token indices are then mapped to content and positional embeddings, summed and passed through LayerNorm and Dropout. Drug and protein embeddings are each fed into separate multi-layer Transformer encoders to capture intra-sequence context. Model forms an interaction tensor by elementwise multiplying the two Transformer outputs and summing across the embedding dimension, yielding a (batch, D, P) interaction matrix. A single 3×3 Conv2d layer followed by Dropout extracts local high-order interaction features, which are then flattened and decoded through fully connected layers to produce a logit. We apply a Sigmoid activation and train with binary cross-entropy loss, using the Adam optimizer (initial learning rate 1e-4) and an early-stopping schedule on the validation set to prevent overfitting.

3.2 Dataset

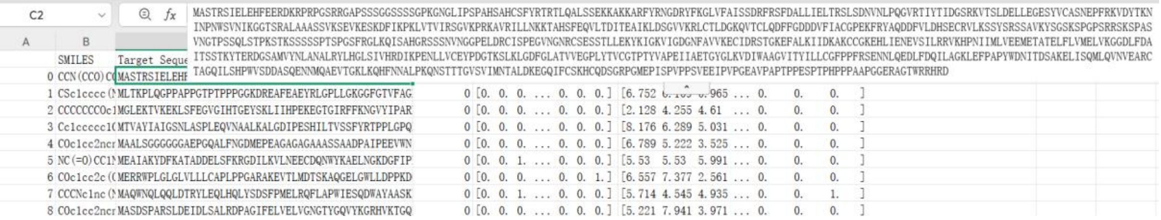
In this project, three datasets are used BindingDB, BIOSNAP, and DAVIS. Each dataset consists of drug-protein pairs labeled as either interacting or non-interacting. These datasets are used to train and evaluate the performance of the deep learning model.

3.2.1 Dataset 1 - BindingDB

A dataset of experimentally validated drug-target interactions, which contains 13288 drug-protein pairs. The drug molecules are represented by their SMILES strings, and the target proteins are represented by their amino acid sequences. Each pair is labeled as either interacting or non-interacting. The BingdDB dataset is available on BindingDB website.

3.2.2 Dataset 2 - BIOSNAP

A large-scale dataset containing 19237 drug-protein pairs with labels indicating whether there is an interaction between the drug and the protein. Like BindingDB, drugs are represented by their SMILES strings and proteins by their amino acid sequences. This dataset is available on the BIOSNAP website.



```
C2    fx  MASTRSIELEHFEERDKRPRPGSRRGAPSSSGGSSSSGPKGNGLIPSPAHSAHCSFYRTRTLQALSSEKKAKKARFYRNGDRYFKGLVFAISSDRFRSFDALLIELTRSLSDNVNLPQGVRTIYTIDGSRKVTSLDELLEGESYVCASNEPFRKVDYTKN
          INPNWSVNIKGGTSRALAAASSVKSEVKESKDFIKPKLVTVIRSGVKPRKAVRILLNKKTAHSFEQVLTDITEAIKLDSGVVKRLCTLDGKQVTCLQDFFGDDDVFIACGPEKFRYAQDDFVLDHSECRVLKSSYSRSSAVKYSGSKSPGPSRRSKSPAS
          VNGTPSSQLSTPKSTKSSSSSPTSPGSFRGLKQISAHGRSSSNVNGGPELDRCISPEGVNGNRCSESSTLLEKYKIGKVIGDGNFAVVKECIDRSTGKEFALKIIDKAKCCGKEHLIENEVSILRRVKHPNIIMLVEEMETATELFLVMELVKGGDLFDA
          ITSSTKYTERDGSAMVYNLANALRYLHGLSIVHRDIKPENLLVCEYPDGTKSLKLGDFGLATVVEGPLYTVCGTPTYVAPEIIAETGYGLKVDIWAAGVITYILLCGFPPFRSENNLQEDLFDQILAGKLEFPAPYWDNITDSAKELISQMLQVNVEARC
          TAGQILSHPWVSDDASQENNMQAEVTGKLKQHFNNALPKQNSTTTGVSVIMNTALDKEGQIFCSKHCQDSGRPGMEPISPVPPSVEEIPVPGEAVPAPTPPESPTPHPPPAAPGGERAGTWRRHRD
A   B
    SMILES     Target Seque
  0 CCN(CCO)CC MASTRSIELEHF
  1 CSc1cccc(N MLTKPLQGPPAPPGTPTPPPGGKDREAFEAEYRLGPLLGKGGFGTVFAG   0 [0.  0.  0.  ...  0.  0.  0.] [6.752 [illegible] 965 ... 0.    0.    0.    ]
  2 CCCCCCCOc1 MGLEKTVKEKLSFEGVGIHTGEYSKLIIHPEKEGTGIRFFKNGVYIPAR   0 [0.  0.  0.  ...  0.  0.  0.] [2.128 4.255 4.61  ... 0.    0.    0.    ]
  3 Cc1ccccc1( MTVAYIAIGSNLASPLEQVNAALKALGDIPESHILTVSSFYRTPPLGPQ   0 [0.  0.  0.  ...  0.  0.  0.] [8.176 6.289 5.031 ... 0.    0.    0.    ]
  4 COc1cc2ncr MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWN   0 [0.  0.  0.  ...  0.  0.  0.] [6.789 5.222 3.525 ... 0.    0.    0.    ]
  5 NC(=O)CC1N MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIP   0 [0.  0.  1.  ...  0.  0.  0.] [5.53  5.53  5.991 ... 0.    0.    0.    ]
  6 COc1cc2c(( MERRWPLGLGLVLLLCAPLPPGARAKEVTLMDTSKAQGELGWLLDPPKD   0 [0.  0.  0.  ...  0.  0.  1.] [6.557 7.377 2.561 ... 0.    0.    0.    ]
  7 CCCNc1nc(N MAQWNQLQQLDTRYLEQLHQLYSDSFPMELRQFLAPWIESQDWAYAASK   0 [0.  0.  1.  ...  0.  0.  0.] [5.714 4.545 4.935 ... 0.    0.    1.    ]
  8 COc1cc2ncr MASDSPARSLDEIDLSALRDPAGIFELVELVGNGTYGQVYKGRHVKTGQ   0 [0.  0.  0.  ...  0.  0.  0.] [5.221 7.941 3.971 ... 0.    0.    0.    ]
```

Figure 6: Data instance of BIOSNAP

3.2.3 Dataset 3 - DAVIS

A dataset containing 2085 drug-protein pairs that are labeled as interacting or non-interacting. The dataset provides the SMILES strings for the drugs and the amino acid sequences for the proteins. This dataset is available on the Kaggle website.

3.3 Data Preprocessing

In this project, the main task of data preprocessing is to convert the drug SMILES representations and protein amino acid sequences into numerical inputs that the model can process. The specific preprocessing steps are as follows.

3.3.1 Drug SMILES Encoding

The drug SMILES representations are processed using the Byte Pair Encoding (BPE) method. The SMILES string is split into substructure tokens (e.g., CCO might be split into ["C", "CO"]). These substructure tokens are mapped to corresponding index values based on a pre-trained BPE vocabulary. This process allows the SMILES representation to be input to the model as a numerical vector.

3.3.2 Protein Sequence Encoding

The protein amino acid sequences are converted into numerical representations. Each amino acid is mapped to a unique numerical index, using a predefined mapping table that converts amino acid characters (such as A, R, M, etc.) into numbers. For each protein sequence, if its length is less than the predefined maximum length max_p, which is set as 545, it is padded with zeros. The padding is masked to prevent it from interfering with the model training. The final output is a numerical vector of length 545, accompanied by a mask array that marks the actual data and the padded portion.

3.3.3 Padding and Masking

During the encoding process, if the length of the drug or protein sequence is less than the predefined maximum length, it is extended by zero-padding. The padded part does not affect the model's learning and is thus handled with masking. The padded sections are masked to prevent them from interfering with the model's learning during training. A mask is created using a matrix of ones, where the non-padded part corresponds to a mask value of 1, and the padded part corresponds to a mask value of 0. This ensures that the model does not consider the padded parts when computing the loss function.

3.4 Data Split

For each dataset, the data is divided into three parts. The Training Set (Train) part is used to train the model. It contains the majority of drug-protein pairs. Validation Set (Val) is used to evaluate the model during training and tune hyperparameters. Test Set (Test) is used to assess the final performance of the model after training. Regarding the dataset splitting, each dataset is randomly divided into training, validation, and test sets. While specific proportions may vary based on the dataset, a typical split is 80% for training, 10% for validation, and 10% for testing. By applying this approach, ensuring each set has enough samples for training, validation, and testing purposes.

3.5 Component Modules and Model Architecture

The proposed model is a three-stage deep learning framework for drug–protein interaction prediction. First, mine frequent substructures from SMILES and amino-acid sequences using FCS + BPE to obtain discrete tokens. These tokens are then converted into content and positional embeddings and fed into multi-layer Transformer encoders to capture rich contextual relations. Finally, we construct an interaction map by element-wise combining drug and protein embeddings, apply a small CNN to extract local high-order features, and decode through an MLP to produce a binary interaction probability.

3.5.1 FCS Mining Module

The FCS Mining Module integrates the substructure vocabulary obtained via Frequent Consecutive Sub-sequence (FCS) mining with the data pipeline, converting raw SMILES and amino-acid sequences into model-ready index sequences. First, it loads the drug and protein substructure vocabularies—previously mined from large unlabeled molecular databases and stored in text files—and uses them to initialize two BPE tokenizers (pbpe for proteins, dbpe for drugs) that split each input string into high-frequency substructure tokens. Next, it reads the corresponding mapping tables to assign each token a unique integer index. During the Index Mapping (words2idx) step, the module applies BPE tokenization, converts tokens to their indices by lookup, then pads or truncates each index sequence to fixed lengths (maximum 545 for proteins, 50 for drugs) and generates 0/1 mask vectors to mark valid token positions. Finally, these index sequences and masks are packaged by BIN_Data_Encoder into a PyTorch Dataset, ready for the Embedding and Transformer encoding stages.

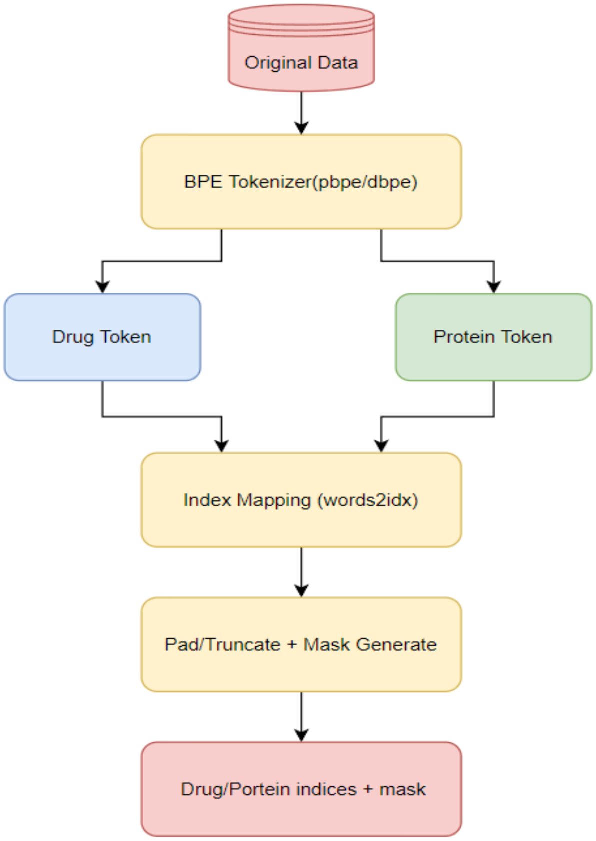

Original Data
BPE Tokenizer(pbpe/dbpe)
Drug Token
Protein Token
Index Mapping (words2idx)
Pad/Truncate + Mask Generate
Drug/Portein indices + mask

Figure 7: Model architecture of FCS Mining Module

3.5.2 Enhanced Transformer Embedding Module

The Enhanced Transformer Embedding Module, as shown in the diagram, first receives the substructure index sequences and their masks output by the FCS Mining Module. With a default batch size of 32 during training, these tensors have shapes of (32, 50) for drugs and (32, 545) for proteins. They are then passed into the Embedding layer, which maps each substructure index to a d_model-dimensional content embedding (word embedding) and a positional embedding; these two embeddings are summed and then normalized via LayerNorm and regularized with Dropout to produce initial representations of the same shapes. Besides, in the Embedding step, we compute for the i-th protein substructure by using Formula 1.

$$E_i^p = E_{\text{cont},i}^p + E_{\text{pos},i}^p \qquad (1)$$

Where $E_{\text{cont},i}^p$ is the content embedding and $E_{\text{pos},i}^p$ is the positional embedding; the drug side follows the same form. Then, the full embedding matrix is passed through the multi-layer Transformer encoder for contextual enhancement, the processing procedure is shown in formula 2.

$$\widetilde{E^p} = \text{TransformerProtein}(E^p),\ \widetilde{E^d} = \text{TransformerDrug}\left(E^d\right) \qquad (2)$$

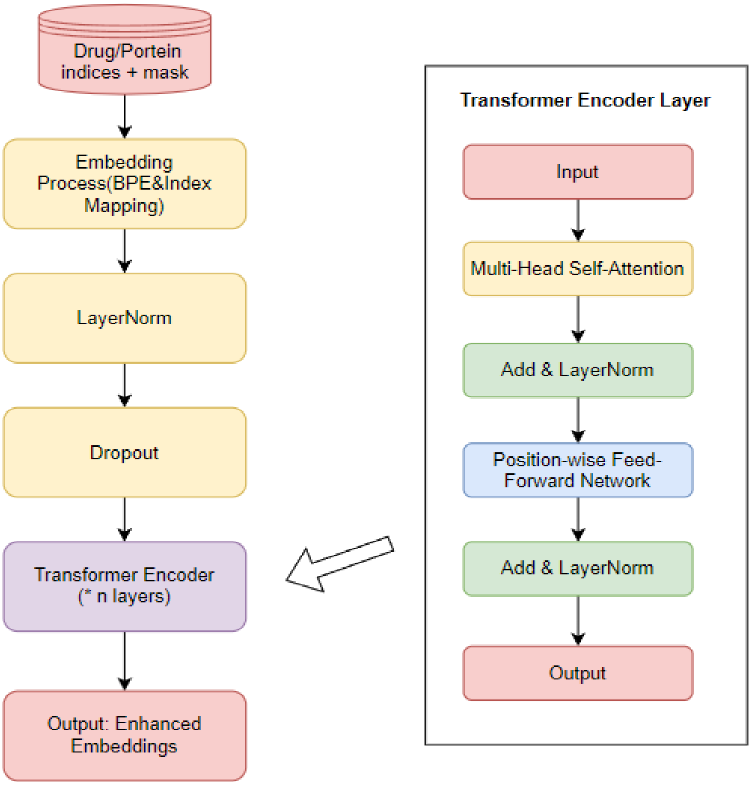

Drug/Portein indices + mask
Embedding Process(BPE&Index Mapping)
LayerNorm
Dropout
Transformer Encoder (* n layers)
Output: Enhanced Embeddings
Transformer Encoder Layer
Input
Multi-Head Self-Attention
Add & LayerNorm
Position-wise Feed-Forward Network
Add & LayerNorm
Output

Figure 8: Model architecture of Enhanced Transformer Embedding Module

Next, these representations are fed into a stack of n standard Transformer encoder layers—each layer consisting of multi-head self-attention, residual addition and LayerNorm, followed by a position-wise feed-forward network, another residual addition and LayerNorm. After all layers, the output tensors retain shapes of (32, 50, 384) and (32, 545, 384), but now each position's vector is enriched with global contextual information from every other substructure in the sequence, forming the "augmented substructure embeddings" used by the subsequent interaction prediction module.

3.5.3 Interaction Prediction Module

The Interaction Prediction Module first takes the drug and protein embeddings from the Transformer encoder, performs elementwise multiplication at each position, and sums over the embedding dimension to produce a single-channel interaction matrix. It then adds a channel dimension to this matrix and applies dropout to prevent overfitting. Next, this single-channel matrix is passed through a 3×3 convolutional layer (CNN), expanding the number of channels from 1 to 3 to scan and extracting local high-order interaction patterns. After the convolution yields multi-channel feature maps, the module flattens these maps into a one-dimensional vector.

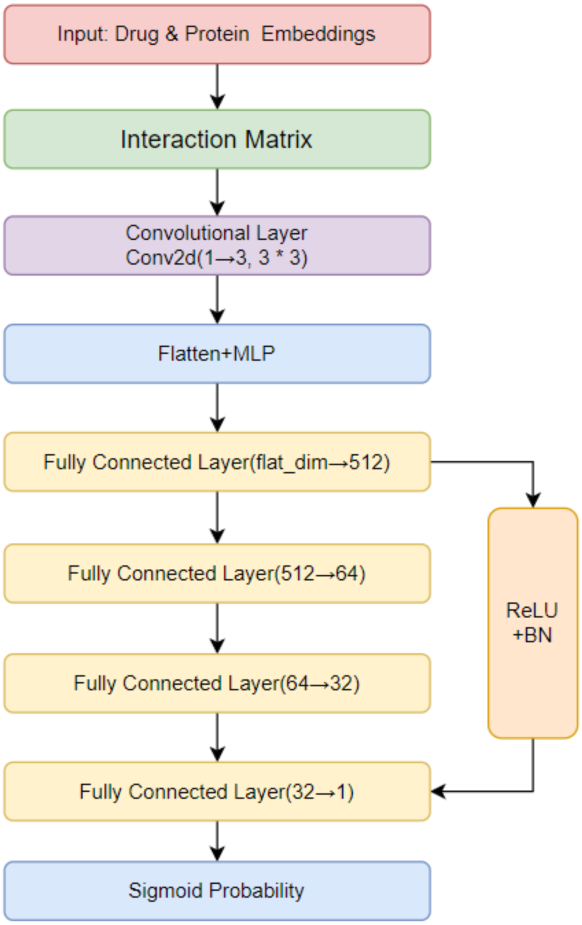

Input: Drug & Protein Embeddings
Interaction Matrix
Convolutional Layer
Conv2d(1→3, 3 * 3)
Flatten+MLP
Fully Connected Layer(flat_dim→512)
Fully Connected Layer(512→64)
Fully Connected Layer(64→32)
Fully Connected Layer(32→1)
Sigmoid Probability
ReLU
+BN

Figure 9: Model architecture of Interaction Prediction Module

This vector is then fed into a multilayer perceptron (MLP decoder): first through a 512-unit fully connected layer with Rectified Linear Unit (ReLU) activation and batch normalization (BN), then through a 64-unit fully connected layer (also with ReLU and batch normalization), followed by a 32-unit fully connected layer, and finally mapped to a single neuron that outputs a logit. The logit is passed through a Sigmoid activation function to map it into the [0,1] range, yielding the model's predicted probability of drug–protein interaction.

3.5.4 Model Workflow

In this project, the proposed model first read the drug SMILES strings and protein amino acid sequences and use the FCS algorithm combined with BPE to decompose them into high-frequency substructure tokens, mapping these tokens to index sequences with corresponding masks. Model then perform content and positional embedding for each token index, sum the two embeddings, and apply LayerNorm and Dropout to obtain the initial representations.

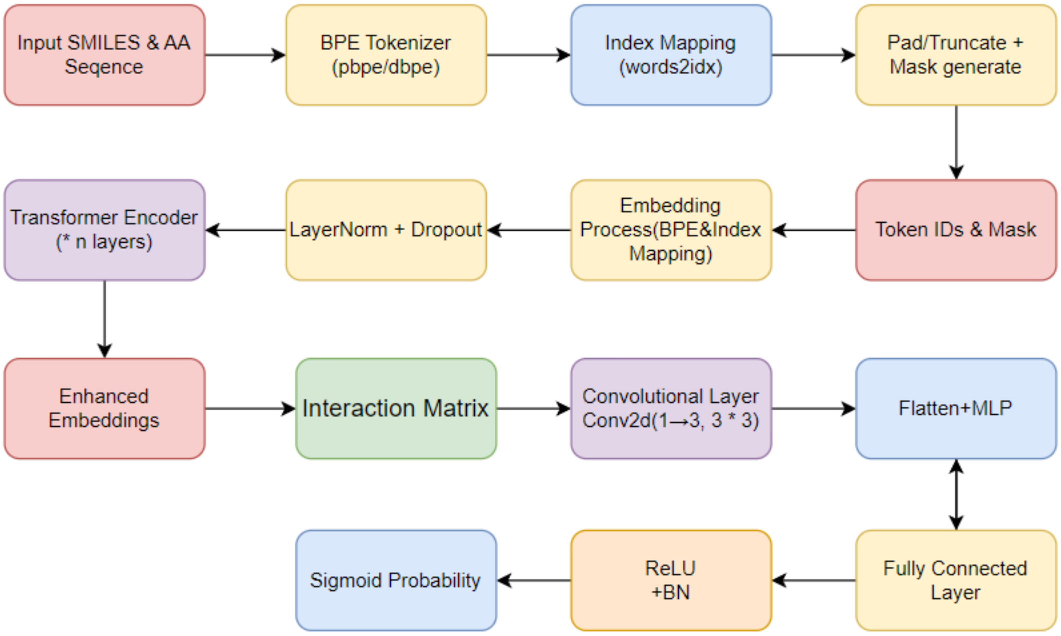

Input SMILES & AA Seqence
BPE Tokenizer (pbpe/dbpe)
Index Mapping (words2idx)
Pad/Truncate + Mask generate
Token IDs & Mask
Embedding Process(BPE&Index Mapping)
LayerNorm + Dropout
Transformer Encoder (* n layers)
Enhanced Embeddings
Interaction Matrix
Convolutional Layer Conv2d(1→3, 3 * 3)
Flatten+MLP
Fully Connected Layer
ReLU +BN
Sigmoid Probability

Figure 10: Overview of Model Workflow

Next, we feed the drug and protein substructure representations separately into multi-layer Transformer encoders to capture internal contextual relationships and output enhanced embeddings; we then construct an interaction matrix of shape (batch, D, P) by element-wise multiplying the two embeddings and summing along the embedding dimension; after that, we apply a 3×3 Conv2d convolution followed by Dropout on the interaction matrix to extract local high-order interaction features; finally, we flatten the convolution output and pass it through fully connected layers of sizes 512→64→32→1 to produce a single logit, which is activated by a Sigmoid to yield the binary probability of drug–target interaction.

3.5.5 Experimental Setup and Technology

The model is optimized using the Adam optimizer, which is well-suited for training deep learning models due to its efficiency in handling sparse gradients. The objective function used for training is binary cross-entropy loss, which is appropriate for the binary classification task of predicting drug-protein interactions. To avoid overfitting, dropout regularization is used by the model. It temporarily overwrites some percentage of the neurons during training to avoid the model relying on a specific neuron. Furthermore, early stopping is implemented during training to avoid training when the model may no longer improve performance on the validation set.

Table 2. The technologies of the project

Software	Framework	Tensorflow 2.10.0; CUDA 11.2 cuDNN 9.1
	Language	Python 3.8.20
	Libraries	Pandas 2.0.3; matplotlib 3.7.5; seaborn 0.13.2; h5py 3.11.0; ipykernel 6.29.5; ipython 8.12.2;
	Version Management	GitHub
	Operation System	Windows 11
Hardware	CPU	AMD Ryen 9 5900HX with Radeon Graphics
	GPU	NVIDIA_GeForce_RTX_3060

3.6 Evaluation Metrics

3.6.1 Area Under the Curve (AUC)

AUC is a popular performance metric of binary classification models. It approximates how well the model ranks positive and negative instances for any possible classification threshold[23]. In drug-protein interaction prediction, AUC is the probability that the model ranks a randomly selected positive interaction higher than a randomly selected non-interaction. AUC varies from 0 to 1, with 1 indicating perfect classification and 0.5 indicating random guessing. AUC is derived from the Receiver Operating Characteristic (ROC) curve, which graphs the actual positive rate (TPR) against the false positive rate (FPR) at different thresholds, as in equation (3).

$$AUC = \int_{-\infty}^{+\infty} P(TPR|FPR)\ dFPR \qquad (3)$$

Actual Positive Rate (TPR) denotes the ratio of correctly classified true positives, and False Positive Rate (FPR) denotes the ratio of misclassified false negatives as positives. The model will better differentiate between positive and negative classes as AUC rises.

3.6.2 Area Under the Precision-Recall Curve (AUPRC)

AUPRC is another important metric, particularly when dealing with imbalanced datasets. It measures the trade-off between precision and recall for different thresholds. AUPRC is the area under the Precision-Recall (PR) curve, which plots precision against recall. In the context of drug-protein interaction prediction, precision refers to the proportion of correctly predicted interactions out of all predicted interactions, while recall measures how many actual interactions were correctly predicted[24]. AUPRC provides a more informative metric than accuracy when the dataset contains a large class imbalance, which is common in biological data. AUPRC equation as seen in equation (4).

$$AUPRC = \int_{0}^{1} P(Recall|Precision)\ dPrecision \qquad (4)$$

Precision is the ratio of true positive interactions to all predicted interactions. Recall is the ratio of correctly predicted interactions to actual interactions. Larger values of AUPRC are better model performances, particularly under imbalanced settings with many negative cases.

3.6.3 F1 Score

The F1 score is the harmonic means of precision and recall, offering a balance between the two metrics. It is particularly useful when the class distribution is imbalanced, as it ensures that both false positives and false negatives are considered in the evaluation[25]. The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates poor performance in both categories[26]. AUPRC equation as seen in equation (5).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (5)$$

Where, Precision is the ratio of true positive predictions to all positive predictions. Recall is the ratio of true positive predictions to all actual positive cases. A high F1 score indicates that the model performs well in both identifying positive interactions (high recall) and correctly labeling them as positive (high precision).

3.6.4 Loss (Binary Cross-Entropy Loss)

Loss functions are essential for training the model, and in binary classification tasks, Binary Cross-Entropy Loss is commonly used. It measures the difference between the predicted probabilities (from the model's output) and the actual binary labels. The loss function penalizes the model more when it makes predictions further from the true label. During model training, the goal is to minimize this loss to improve the model's performance[27]. Loss equation as seen in equation (6).

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \qquad (6)$$

y_i is the true label (0 or 1). p_i is the predicted probability for the positive class (the probability that the drug interacts with the protein). N is the total number of samples. A lower loss value indicates that the model's predictions are closer to the true labels, leading to better model performance.

3.7 Data Testing

For evaluation purposes, for the test set of drug-protein pairs, ensure that the test set includes all two examples of interaction and non-interaction, spanning multiple protein families and drug categories, to ensure broad biological representation. Each pair's annotation has been strictly checked, so each SMILES string, protein sequence and interaction label is complete and accurate. To ensure fairness, we apply exactly the same preprocessing pipeline used during training to the test data, thereby eliminating any differences or biases in performance evaluation.

Chapter 4 Implementation and Result Analyses

In this section, this part will focus on introducing the specific implementation of the proposed model, including training with different datasets, and at the same time introduce a series of parameters such as the learning rate, epochs and batch size during the model training. Subsequently, the obtained prediction results will be analyzed, and finally, the design related to the GUI will be introduced.

4.1 Results of Model Training

4.1.1 Initial result

This section presents the final performance of the proposed model, which evaluates results using AUC, AUPRC, F1 scores and Loss. The initial learning rate was set to 1e-4 , and a batch size of 32 was adopted during training to balance computational efficiency with the ability to capture data diversity. The displayed results are all the results of the model after 50 epochs of training.

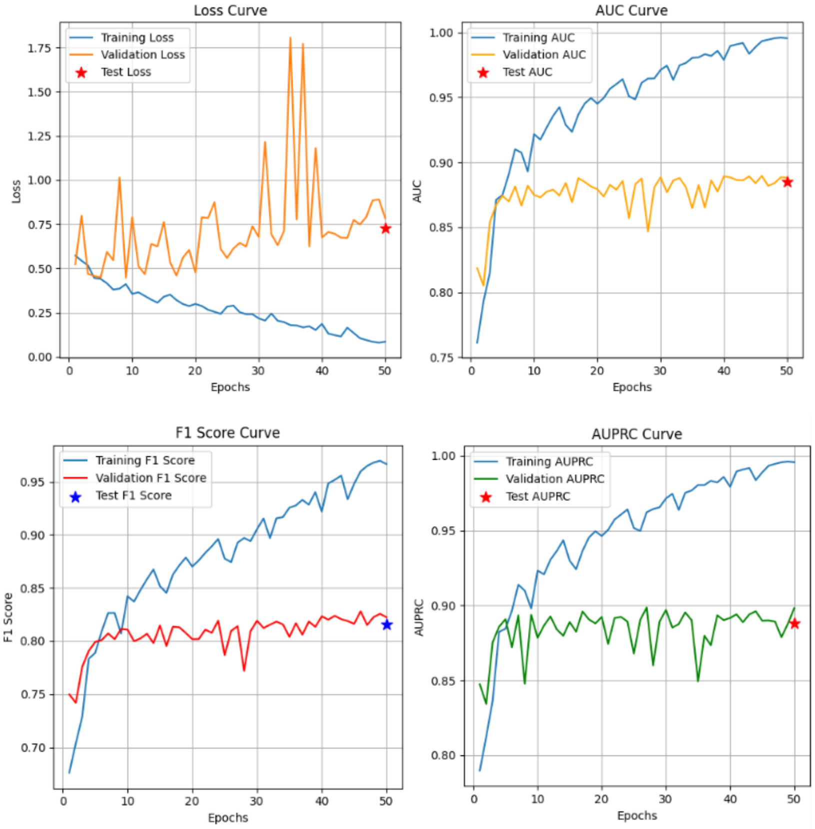


Figure 11: BIOSNAP Performance plots

The model was trained on the BIOSNAP, BindingDB, and DAVIS datasets for drug-protein interaction prediction. Preliminary experiments revealed significant performance variations across datasets due to differences in interaction complexity and diversity. Consequently, the BIOSNAP dataset was selected for final evaluation, as it contains a larger number of drug-protein pairs and exhibits a more balanced ratio of positive to negative samples.

After running multiple training experiments, the final model achieved the following performance metrics on the test set:

Table 3: Main parameters (BIOSNAP)

	AUC	AUPRC	F1	Loss
Training	0.9955	0.9957	0.9664	0.0855
Test	0.8853	0.8881	0.8158	0.4622
Validation	0.8882	0.8982	0.8223	0.4356

In terms of AUC, the model achieved near-perfect performance on the training set of 0.9955, while its AUC on the validation and test sets are 0.8882 and 0.8853, respectively. The small difference between validation and test AUCs indicates that the model's discriminative ability on unseen data (data from test and validation sets) remains robust, despite a drop from its training performance, consistently holding around 0.89 on novel samples.

For AUPRC, the model scored 0.9957 on the training set, 0.8982 on the validation set, and 0.8881 on the test set—closely mirroring the AUC trend. Due to training AUPRC is very good, the validation and test AUPRCs both remain around 0.89, demonstrating that the model effectively balances precision and recall even under class imbalance.

Regarding F1 score, the model attained 0.9664 on the training set but dropped to 0.8223 on the validation set and 0.8158 on the test set. This pattern highlights very balanced discrimination between positive and negative examples during training, although with some reduction in validation set, the F1 score on unseen data remains above 0.82.

Looking at loss, the average training loss was just 0.0855, compared to 0.4356 on validation and 0.4622 on test. The large gap between training and unseen-data losses indicates an excellent fit to the training data but some errors on novel samples. However, the validation and test losses are similar, which also indicates that the overall generalization performance is relatively stable.

4.1.2 Final result

Proposed model delivers strong performance on all metrics, but its validation and test losses begin to fluctuate noticeably during the final epochs. To stabilize training, the optimization was enhanced by adding L$_2$ weight decay to the Adam optimizer and employing a ReduceLROnPlateau scheduler, which automatically lowers the learning rate when improvement stalls. An early-stopping rule was also implemented to halt training if the validation loss fails to improve for a predefined number of epochs. At the end of each epoch, the model is evaluated on both the validation and test sets, ensuring that these adaptive strategies are driven by the most recent performance feedback.

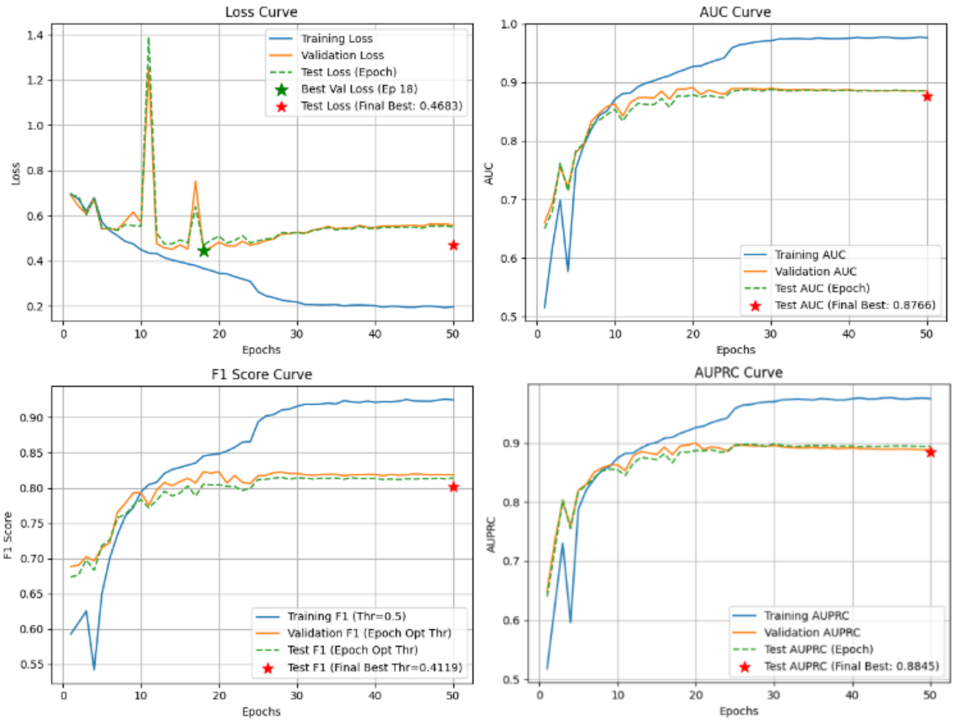


Figure 12: BIOSNAP Performance plots (After improvement)

In addition, to more accurately reflect the model’s balance between precision and recall across epochs, a strategy was introduced that dynamically searches for the optimal decision threshold based on validation-set F1 scores. At the end of each training epoch, the script evaluates a range of cut-off points on the validation set, selects the threshold that maximizes F1, and saves it as that epoch’s optimal threshold (best_f1_threshold). All subsequent F1 evaluations in the training, validation, and testing phases use this optimal threshold for binarization, and the results are plotted as an F1 curve. This ensures that the reported F1 scores always correspond to the best possible decision boundary, providing a more faithful depiction of the model’s precision–recall trade-off throughout training.

Table 4: Main parameters (After improvement)

	AUC	AUPRC	F1	Loss
Training	0.9755	0.9757	0.9258	0.1864
Test	0.8853	0.8881	0.8287	0.4683
Validation	0.8958	0.8998	0.8223	0.4374

The model achieves a training AUC of 0.9755 and AUPRC of 0.9757, demonstrating excellent discrimination between positive and negative samples. On the validation and test sets, it attains AUCs of 0.8958 and 0.8853 and AUPRCs of 0.8998 and 0.8881, respectively—only about 0.08 lower than on the training set—indicating stable and reliable performance on unseen data. The test set F1 score is 0.8287 and the validation set F1 is 0.8223, both exceeding 0.82, further confirming a balanced precision–recall trade-off. In terms of loss, the training, validation, and test losses are 0.1864, 0.4374, and 0.4683; the validation and test loss curves exhibit no significant oscillations, suggesting that the introduced L2 regularization and learning-rate scheduling effectively improved model stability and addressed previous loss-curve fluctuations. Furthermore, the incorporation of a learning-rate scheduler and early-stopping mechanism has made the training process more controllable, enabling automatic learning-rate adjustments and timely termination of training to maintain consistent model performance.

4.1.3 Result analyses with Confusion Matrix

The confusion matrix shows that the model correctly identified 2,142 true negatives. Produced 574 false positives, which means misclassifying 574 non-interacting pairs as interacting; produced 462 false negatives and correctly captured 2,294 true positives. The relatively high false-positive count indicates occasional misclassification of negative samples, while the relatively low false-negative count demonstrates that the model misses a few real interactions, reflecting a high recall rate.

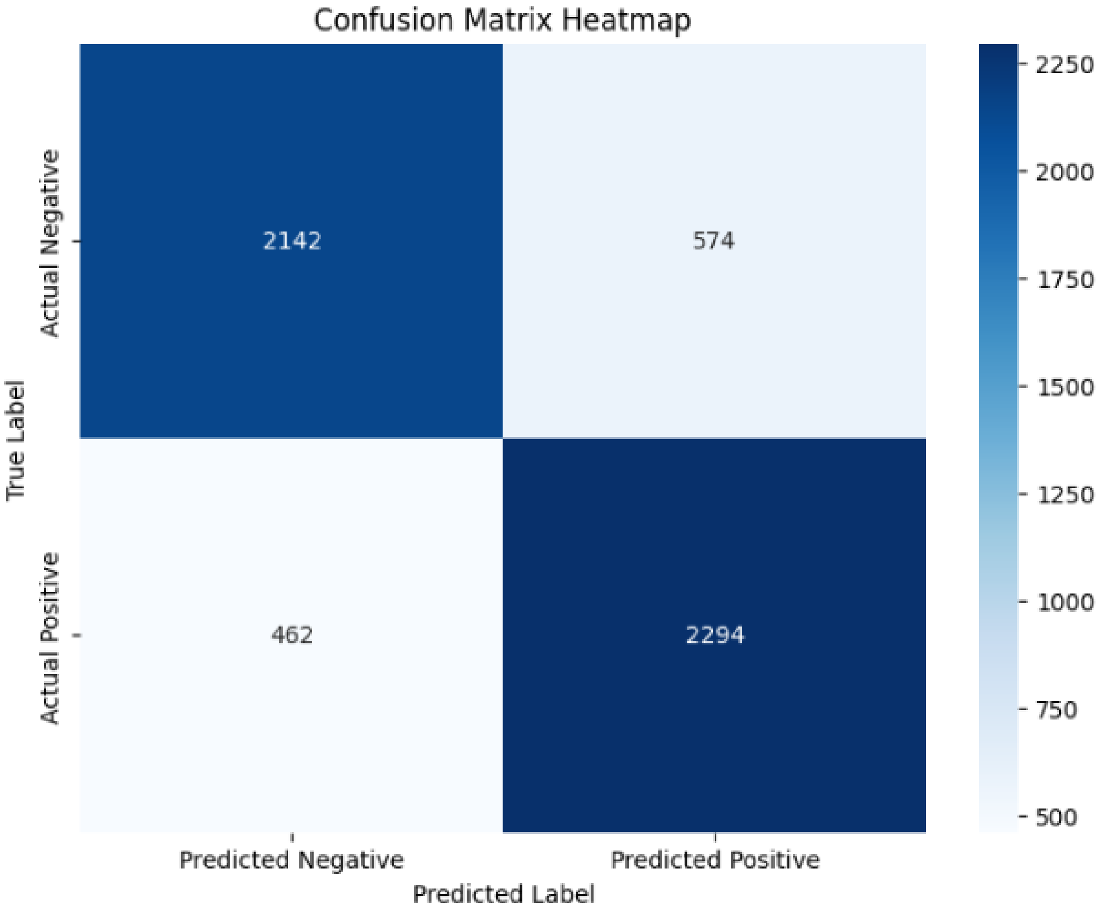


Figure 13: Confusion Matrix Heatmap

4.2 Model Explainability

To achieve Model Explainability, the results of the model will be interpreted and analyzed using five types of SHAP plots, namely Force Plot, Summary Plot, Dependency Plot, Waterfall Plot, and Bar Plot.

4.2.1 Force Plot

This force plot illustrates the prediction process for sample 0 (true label 0, means prediction result is non-binding). The gray vertical line marks the base value (expected value) of approximately 0.49, which is the model's predicted probability before accounting for any feature contributions. Red bars pointing right indicates features that increase the prediction probability, while blue bars pointing left indicate features that decrease red denotes positive contributions and blue denotes negative ones. Here, the total negative contributions slightly outweigh the positive ones, yielding a final model output of about 0.49, below the 0.5 decision threshold, and thus correctly predicting “non-binding.”

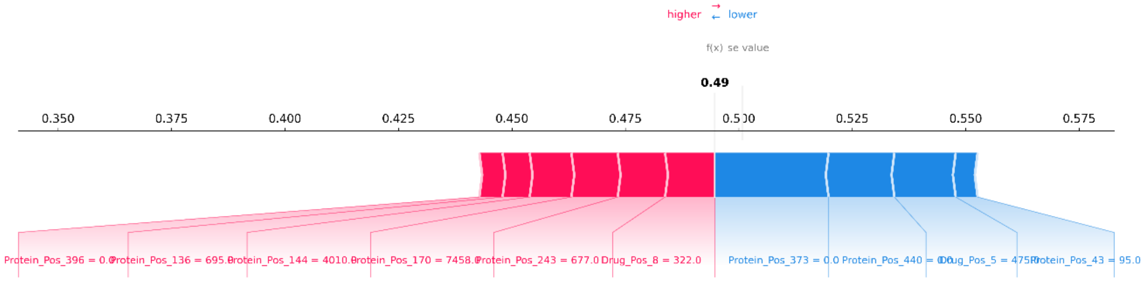


Figure 14: Force Plot

4.2.2 Summary Plot

This summary shows the global impact distribution of each feature—on the model’s output. The vertical axis lists the feature positions. The horizontal axis represents SHAP values, indicating both the magnitude and direction of each feature’s contribution to the predicted binding probability. Positive SHAP values (points to the right) push the prediction toward “binding,” while negative SHAP values (points to the left) suppress it. Point color encodes the actual feature value in each sample: red indicates a high feature value, and blue indicates a low feature value.

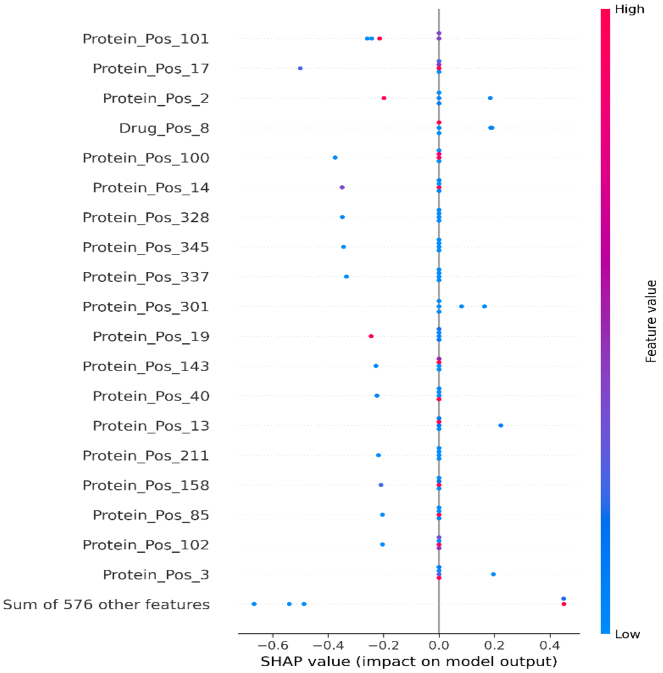


Figure 15: Summary Plot

4.2.3 Dependency Plot

This dependence plot illustrates how Drug_Pos_7 values affect its SHAP contribution to the model output: the x-axis shows Drug_Pos_7's raw values, the y-axis its SHAP values, and each point's color encodes the corresponding Drug_Pos_19 value (blue = low, red = high). Almost all points lie near zero, indicating that Drug_Pos_7 alone has virtually no direct impact on the prediction. Only when Drug_Pos_19 is very high (the purple point) and Drug_Pos_7 is around 1 000 does the SHAP value rise slightly (to about +0.025), suggesting a minor positive interaction between these positions. Even at extreme Drug_Pos_7 values (~8 000) with high Drug_Pos_19, the SHAP values do not shift negatively. Overall, the model does not depend on Drug_Pos_7 by itself, and its influence emerges only through subtle joint effects with Drug_Pos_19.

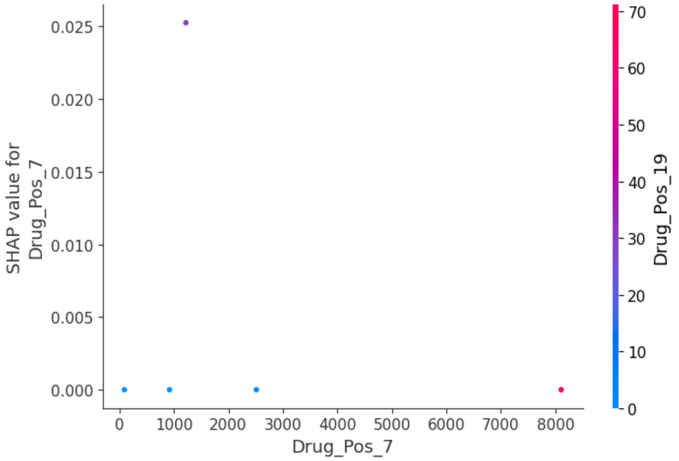


Figure 16: Dependency Plot

4.2.4 Waterfall Plot

This waterfall plot starts from the model's expected value of 0.634 and then shows how individual features push the prediction up or down to its final value of 0.657 for this sample. The largest positive contributor is Drug_Pos_29 of about +0.03, followed by Protein_Pos_33 of about +0.02, which together raise the score above baseline. In contrast, Drug_Pos_31 and Protein_Pos_126 of about –0.01 and Protein_Pos_55 of about –0.01 are the most negative contributors, pulling the prediction back down. While the remaining 581 features have essentially zero impact. The net effect of these pushes and pulls lifts the prediction from 0.634 to 0.657, indicating a slight increase in the model's confidence for binding.

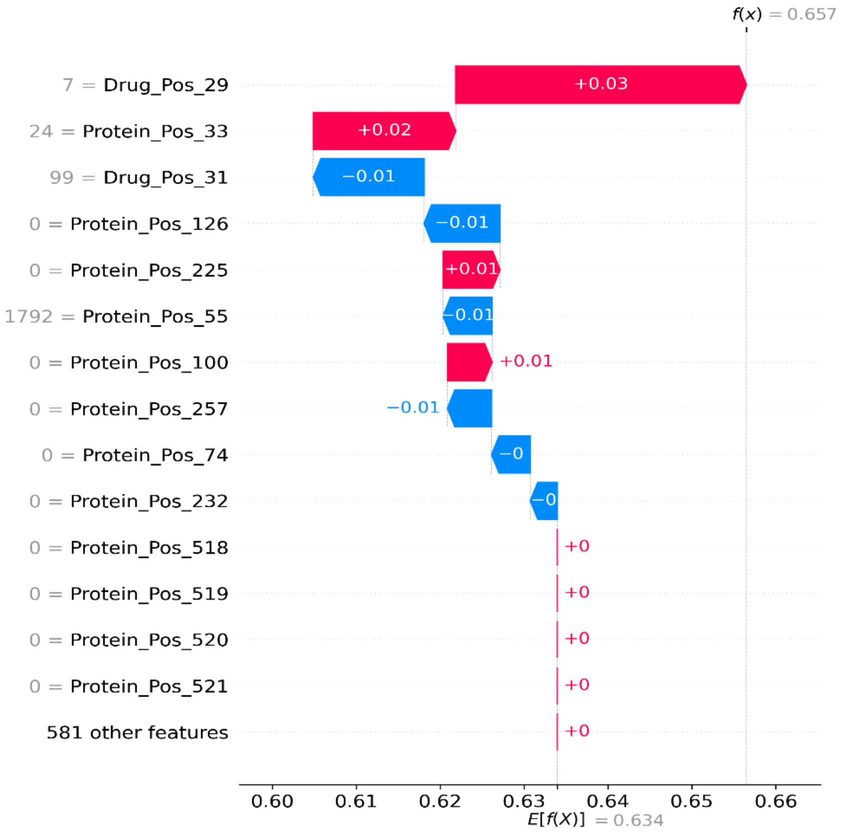


Figure 17: Waterfall Plot

While the remaining 581 features have essentially zero impact. The net effect of these pushes and pulls lifts the prediction from 0.634 to 0.657, indicating a slight increase in the model's confidence for binding.

4.2.5 Bar Plot

This bar plot ranks positions in the input sequence by their mean absolute SHAP values, revealing their global influence on the model's predictions. Protein_Pos_101 contributes the most, with a SHAP value of about 0.14, followed by Protein_Pos_17 SHAP value of about 0.10, Protein_Pos_2 of SHAP value about 0.08. Notably, the combined "Sum of 586 other features" contributes approximately 0.97, indicating that although each individual key position has a moderate effect, the aggregated impact of the many remaining positions is the primary driver of the model's performance.

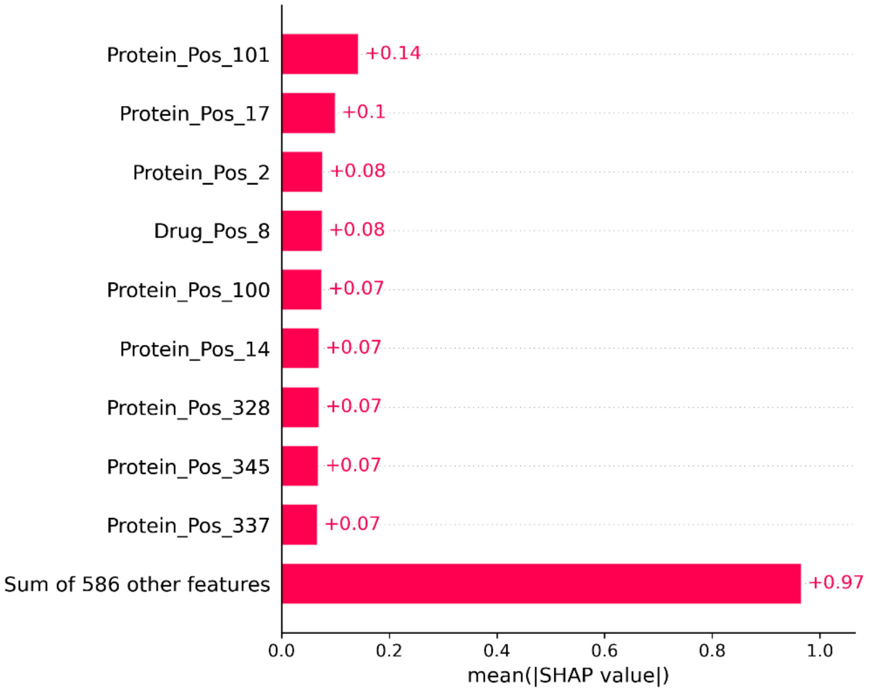


Figure 18: Bar plot

4.3 Appling different datasets for training

(1) DAVIS Dataset

The model performed very well during training, achieving high AUC, AUPRC, and F1 scores, showing strong predictive performance on the training set. The loss is also relatively low, indicating effective learning and model fitting.

Table 5: Main parameters (DAVIS)

	AUC	AUPRC	F1	Loss
Training	0.9368	0.8347	0.8629	0.4223
Test	0.8781	0.3074	0.3021	0.1751
Verification	0.8614	0.3125	0.2755	0.1754

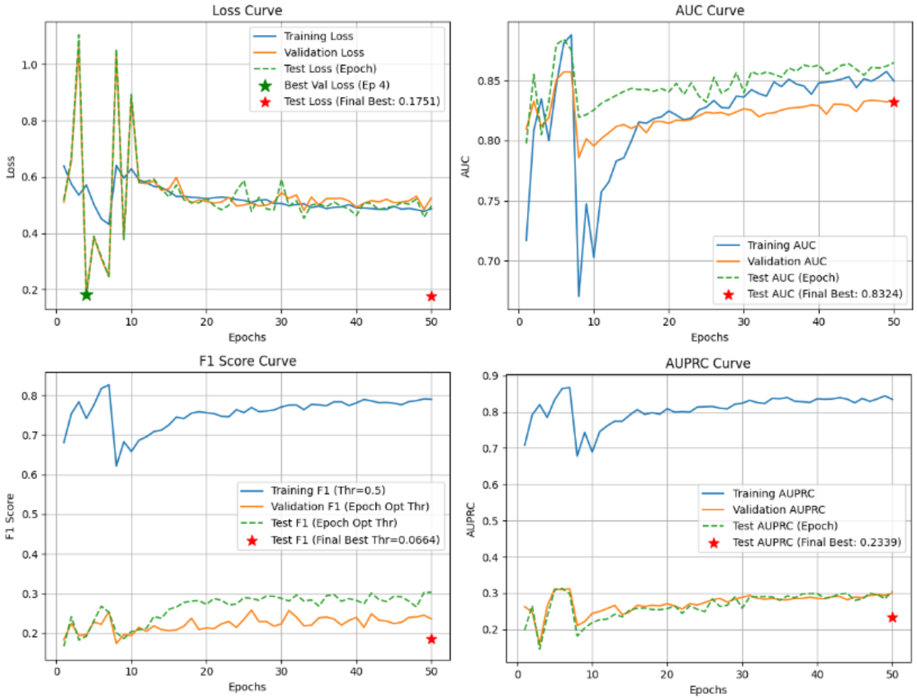


Figure 19: Main parameters change curve (DAVIS)

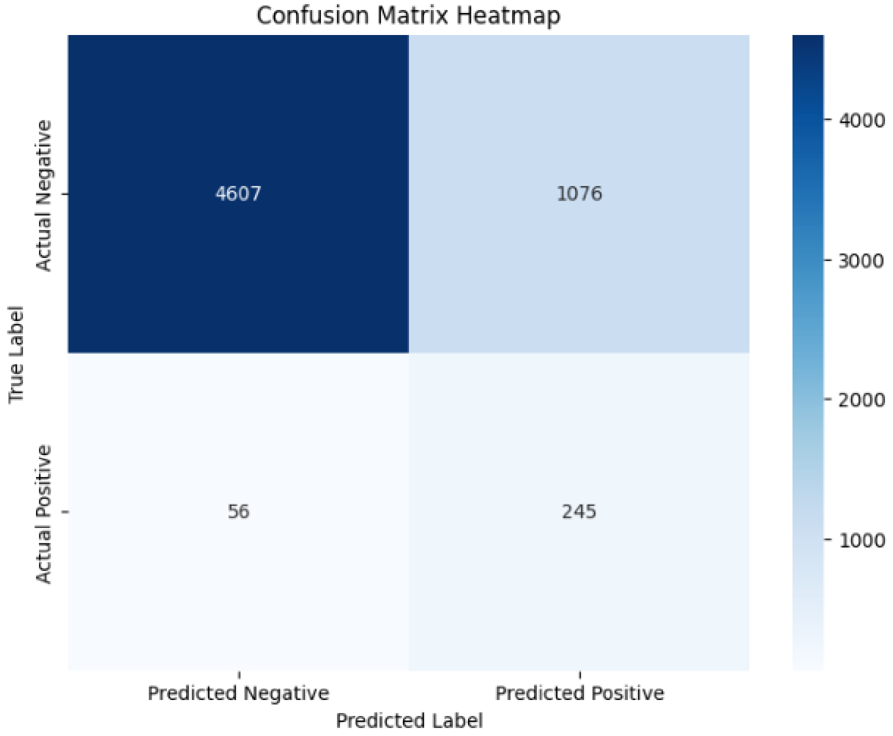


Figure 20: Confusion Metrix (DAVIS)

(2) BindingDB Dataset

The model performed exceptionally well during training, achieving near-perfect results in terms of AUC, AUPRC, and F1-score. Despite drops in performance on the validation and test sets, the AUC and F1-score on the test set are still solid, reflecting proposed model has good generalization capabilities on different datasets.

Table 6: Main parameters (BindingDB)

	AUC	AUPRC	F1	Loss
Training	0.9467	0.9068	0.8458	0.1701
Test	0.8707	0.5861	0.5482	0.3384
Verification	0.8998	0.5802	0.5358	0.3205

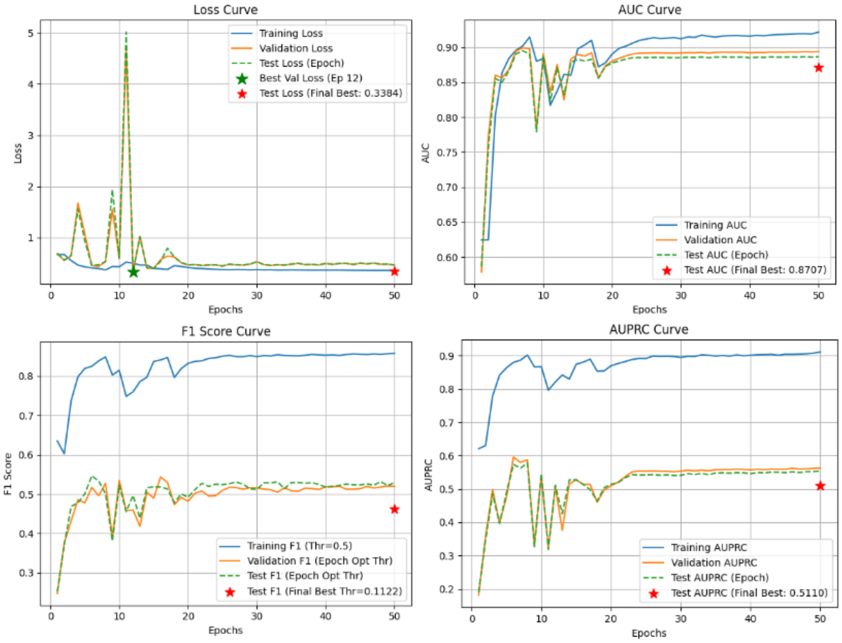


Figure 21: Main parameters change curve (BindingDB)

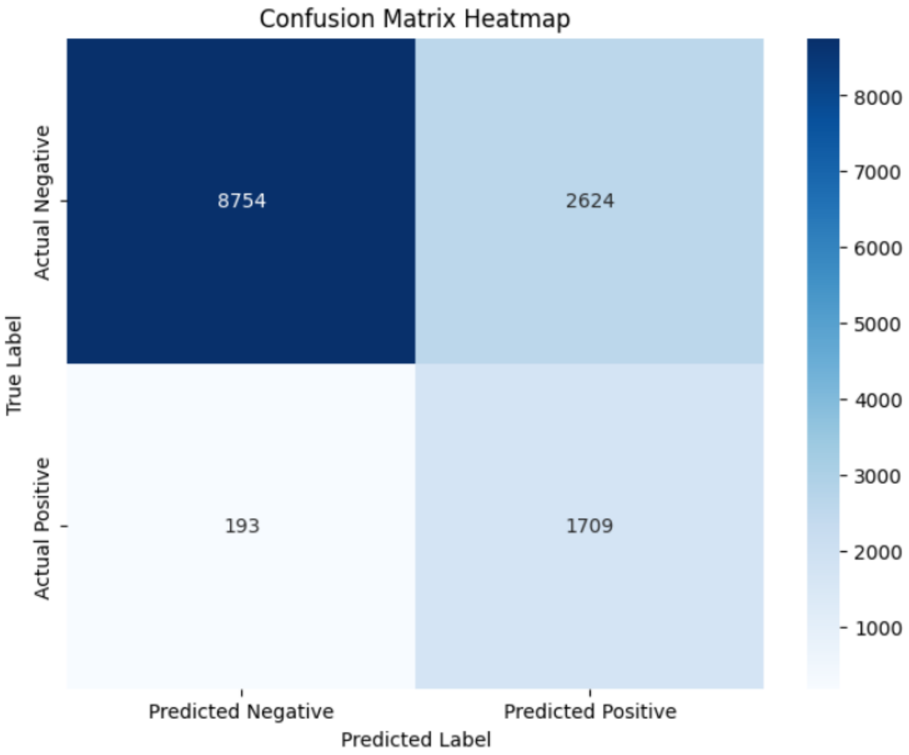


Figure 22: Confusion Metrix (BindingDB)

4.4 Comparison of Other Models

4.4.1 Comparison of Common Models

Table 7: Comparison of Common Models

Model	AUPRC	AUC	F1-score	Loss
LR	0.84660±0.04	0.23260±0.23	0.69960±0.51	Not given
DNN	0.84960±0.03	0.85560±0.10	0.77660±0.40	Not given
DeepDTI	0.87660±0.05	0.87660±0.06	0.78960±0.27	Not given
DeepConv-DTI	0.88360±0.02	0.88960±0.05	0:77060±0.23	Not given
Project model	0.9952	0.9955	0.9668	0.0871

The Project Model achieves an AUPRC of 0.9952, which is the highest among all models compared here. This metric is crucial, especially in imbalanced datasets like drug-protein interactions, as it reflects the model's ability to identify positive instances (drug-protein interactions) effectively. Compared to other models like DeepConv-DTI (0.884) and DeepDTI (0.877), the Project Model demonstrates a significant improvement in handling the positive class.

The Project Model also leads in AUROC, with a value of 0.9952. AUROC is a general metric for classification performance, measuring the model’s ability to distinguish between positive and negative classes. The Project Model outperforms other models like DeepConv-DTI (0.890) and DeepDTI (0.877), highlighting its overall superior discriminative ability.

The F1-score of the Project Model is 0.9668, which is comparable to the top models in this comparison. Specifically, the Project Model outperforms models like DeepConv-DTI (0.771) and DNN (0.777). The F1-score balances both precision and recall, indicating that the Project Model has effectively learned to minimize both false positives and false negatives in predicting drug-protein interactions.

4.4.2 Comparison of other related models

Table 8: Comparison of other Models of related studies

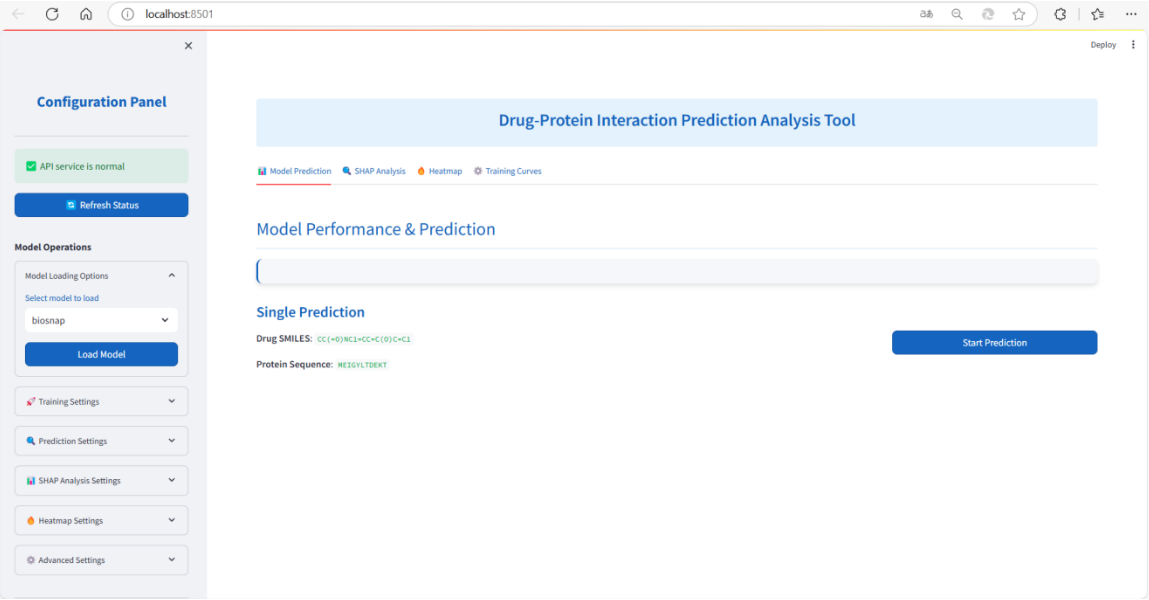
Model	AUPRC	AUC	F1-score	Loss
FBRWPC [1]	0.953	0.630	Not given	Not given
MINDG [16]	0.971	0.951	0.857	Not given
Project model	0.9952	0.9955	0.9668	0.0871

In terms of AUC, the Project model achieves 0.9955, far surpassing MINDG’s 0.951 and FBRWPC’s 0.630, indicating an almost perfect ability to discriminate between positive and negative samples; for AUPRC, the Project model also leads at 0.9952 versus MINDG’s 0.971 and FBRWPC’s 0.953, demonstrating that it maintains an excellent balance of precision and recall even under class imbalance; and for F1-score, the Project model reaches 0.9668, clearly outperforming MINDG’s 0.857 (FBRWPC not reported).

In summary, the project model performs exceptionally well in handling accuracy and recall rate, especially in imbalanced datasets, where identifying minority class-drug-protein interactions is crucial. By achieving an outstanding AUPRC, this model has demonstrated its ability to effectively address this challenge, ensuring accurate prediction of positive interactions. Furthermore, the project model has the highest AUROC and strong discriminative ability. It can effectively distinguish between interacting and non-interacting drug-protein pairs and is superior to other models. This model also achieved a high f1 score, reflecting its ability to strike a good balance between accuracy and recall. This balance is particularly important in applications where accurate prediction and comprehensive coverage of positive samples are necessary. Finally, the low loss value indicates that the project model has been well optimized and can be effectively generalized to unknown data. Compared with other models, this reduces overfitting and further enhances the robustness and efficiency of the model training process.

4.5 GUI Design Demonstration

This project aims to study the interaction between proteins and targeted drugs. An interactive website has been designed and implemented, with the goal of enhancing the user experience and demonstrating the practical application of drug-target interactions using protein structures. This website offers two main functions: predicting the possibility of interaction for a specific pair of drug and protein information and predicting the overall prediction results of the complete dataset.



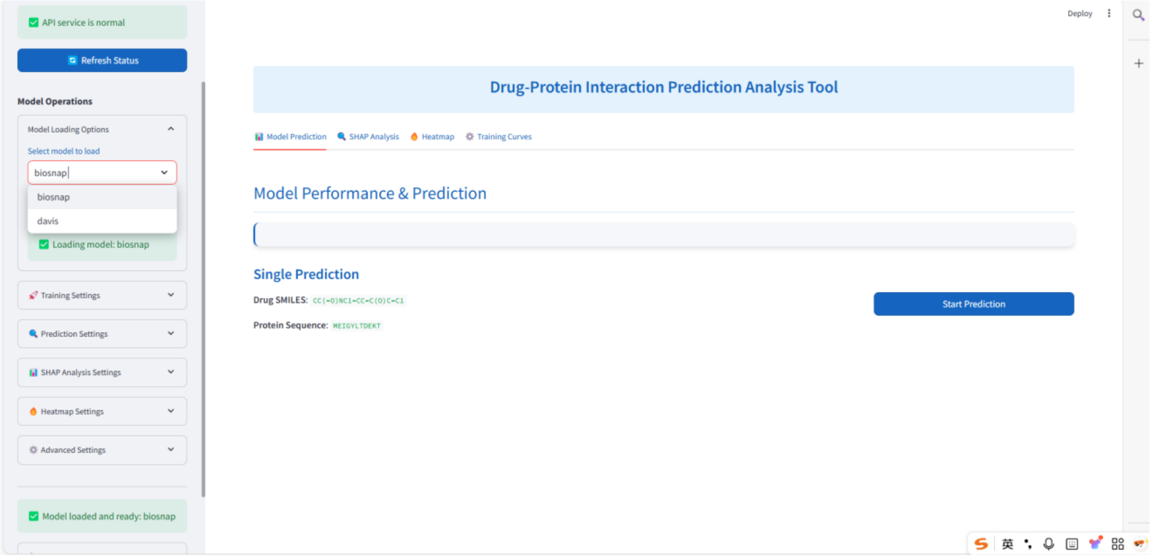
Select model to load: biosnap

Drug SMILES: `CC(=O)NC1=CC=C(O)C=C1`

Protein Sequence: `MEIGYLTDEKT`

Figure 23: Web GUI (home page)

Figure 23 shows the home page of the Website, which offers the single prediction function. By entering the protein and drug information, the prediction output will be shown in this page.



Select model to load: biosnap
biosnap
davis

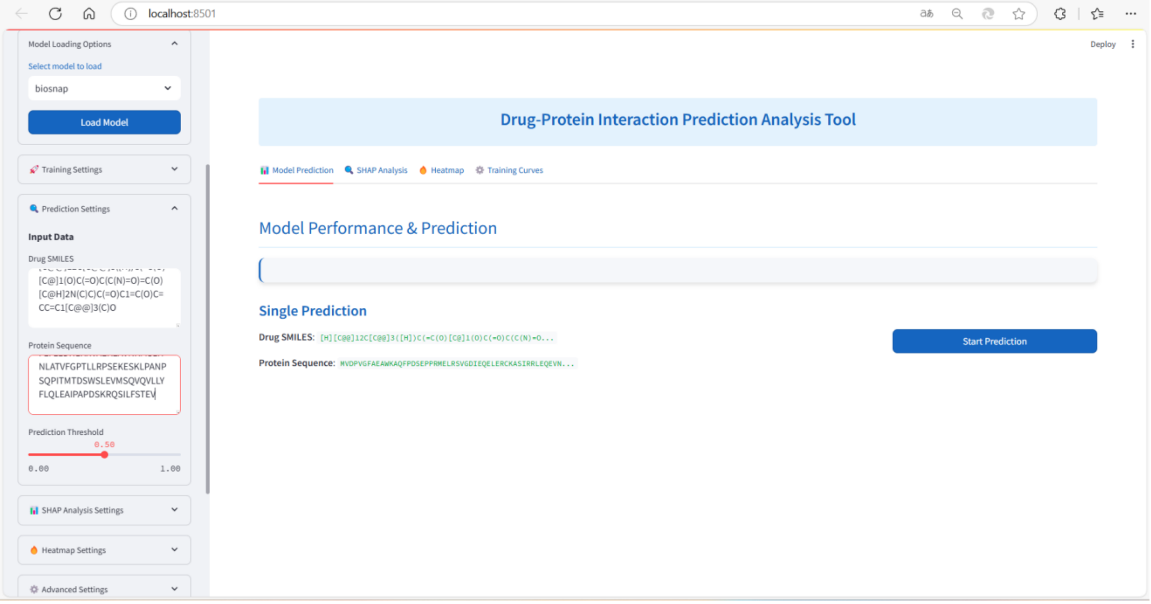
Loading model: biosnap

Model loaded and ready: biosnap

Drug SMILES: `CC(=O)NC1=CC=C(O)C=C1`

Protein Sequence: `MEIGYLTDEKT`

Figure 24: Web GUI (selecting saved model)



Select model to load: biosnap

Drug SMILES:

```
[C@]1(O)C(=O)C(C(N)=O)=C(O)
[C@H]2N(C)C)C(=O)C1=C(O)C=
CC=C1[C@@]3(C)O
```

Protein Sequence:

```
NLATVFGPTLLRPSEKESKLPANP
SQPITMTDSWSLEVMSQVQVLLY
FLQLEAIPAPDSKRQSILFSTEV
```

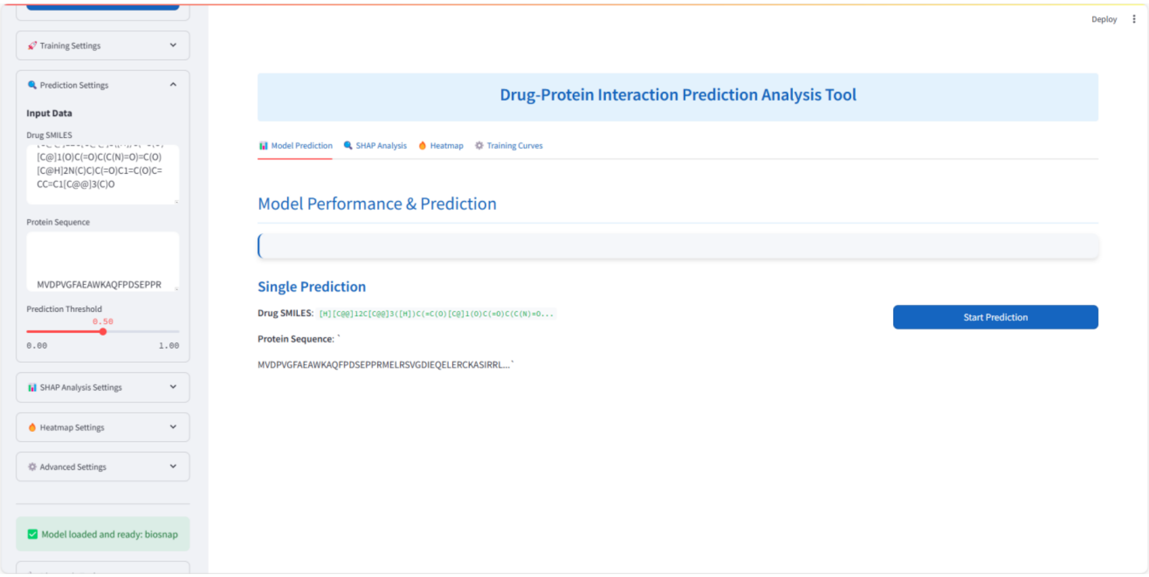
Prediction Threshold: 0.50 (0.00 – 1.00)

Drug SMILES: `[H][C@@]12C[C@@]3([H])C(=C(O)[C@]1(O)C(=O)C(C(N)=O...`

Protein Sequence: `MVDPVGFAEAWKAQFPDSEPPRMELRSVGDIEQELERCKASIRRLEQEVN...`

Figure 25: Web GUI (input correct date)

As figure 24 and 25 show, On the left sidebar of the web page, the user can select the saved model weights. After confirming the option, the single-item prediction will be made using the selected model. Meanwhile, before the prediction starts, the user needs to input the drug and protein information respectively.



Input Data

Drug SMILES
[C@]1(O)C(=O)C(C(N)=O)=C(O)
[C@H]2N(C)C)C(=O)C1=C(O)C=
CC=C1[C@@]3(C)O

Protein Sequence
MVDPVGFAEAWKAQFPDSEPPR

Prediction Threshold 0.50 (0.00 – 1.00)

Model loaded and ready: biosnap

Model Performance & Prediction

Single Prediction

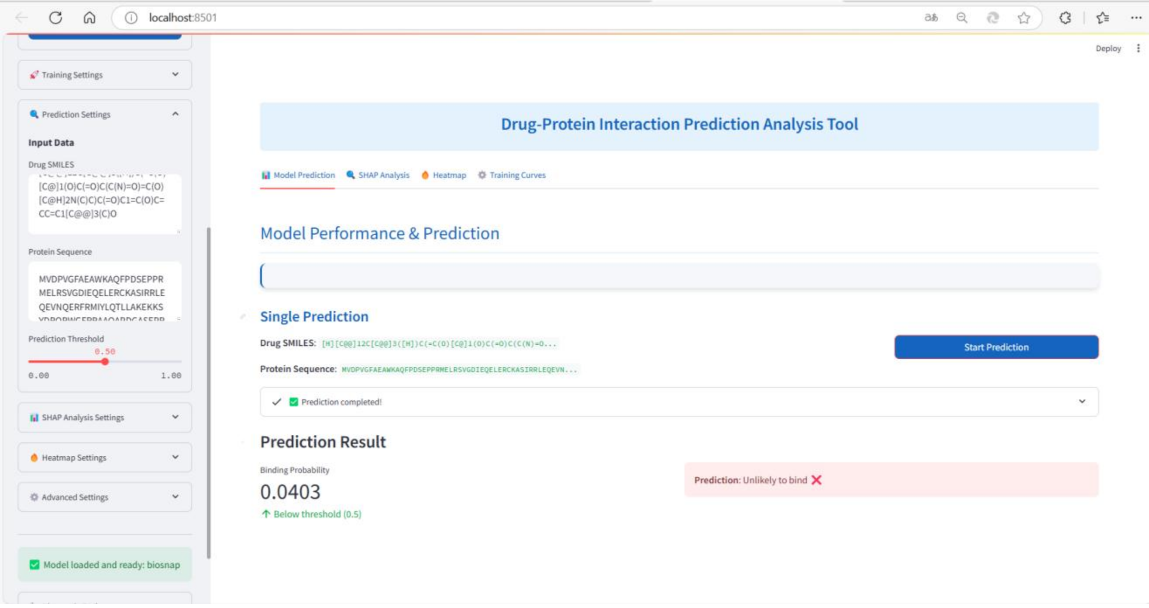
Drug SMILES: `[H][C@@]12C[C@@]3([H])C(=C(O)[C@]1(O)C(=O)C(C(N)=O...`

Protein Sequence: `

MVDPVGFAEAWKAQFPDSEPPRMELRSVGDIEQELERCKASIRRL...`

Figure 26: Web GUI (input wrong date)

Figure 26 shows if the input information is not the expected data format, the input item will not be green color, this function will remind the user to check the input information.



Input Data

Drug SMILES
[C@]1(O)C(=O)C(C(N)=O)=C(O)
[C@H]2N(C)C)C(=O)C1=C(O)C=
CC=C1[C@@]3(C)O

Protein Sequence
MVDPVGFAEAWKAQFPDSEPPR
MELRSVGDIEQELERCKASIRRLE
QEVNQERFRMIYLQTLLAKEKKS

Prediction Threshold 0.50 (0.00 – 1.00)

Model loaded and ready: biosnap

Model Performance & Prediction

Single Prediction

Drug SMILES: `[H][C@@]12C[C@@]3([H])C(=C(O)[C@]1(O)C(=O)C(C(N)=O...`

Protein Sequence: `MVDPVGFAEAWKAQFPDSEPPRMELRSVGDIEQELERCKASIRRLEQEVN...`

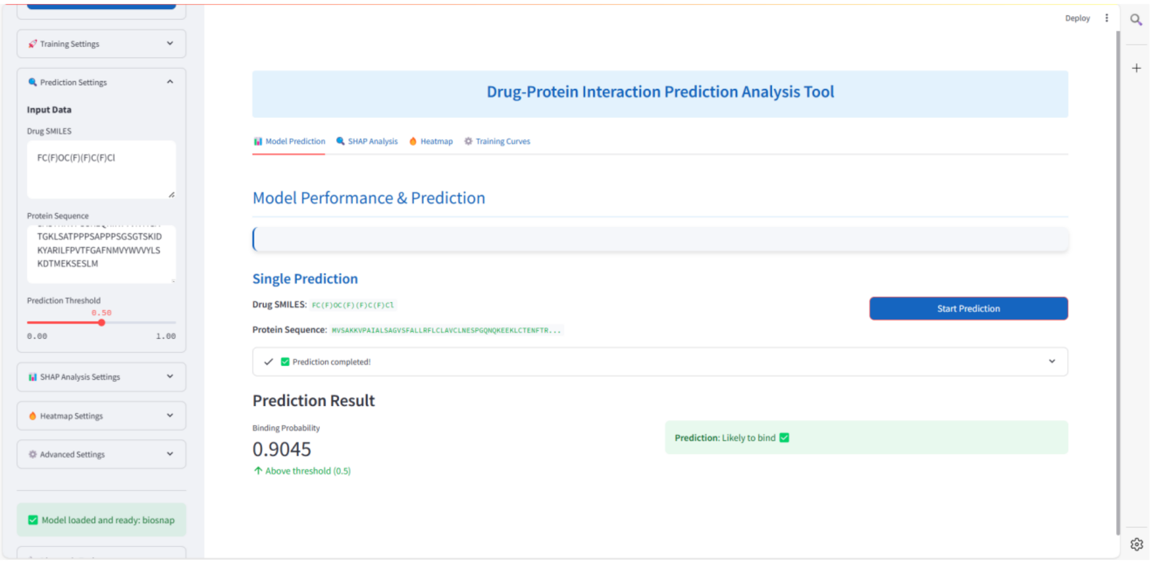
✓ Prediction completed!

Prediction Result

Binding Probability
0.0403
↑ Below threshold (0.5)

Prediction: Unlikely to bind ✗

Figure 27: Web GUI (output will not react)



Input Data

Drug SMILES: FC(F)OC(F)(F)C(F)Cl

Protein Sequence: TGKLSATPPPSAPPPSGSGTSKID KYARILFPVTFGAFNMVYWVVYLS KDTMEKSESLM

Prediction Threshold: 0.50 (0.00 – 1.00)

Model loaded and ready: biosnap

Drug-Protein Interaction Prediction Analysis Tool

Model Performance & Prediction

Single Prediction

Drug SMILES: FC(F)OC(F)(F)C(F)Cl

Protein Sequence: MVSAKKVPAIALSAGVSFALLRFLCLAVCLNESPGQNQKEEKLCTENFTR...

✓ Prediction completed!

Prediction Result

Binding Probability

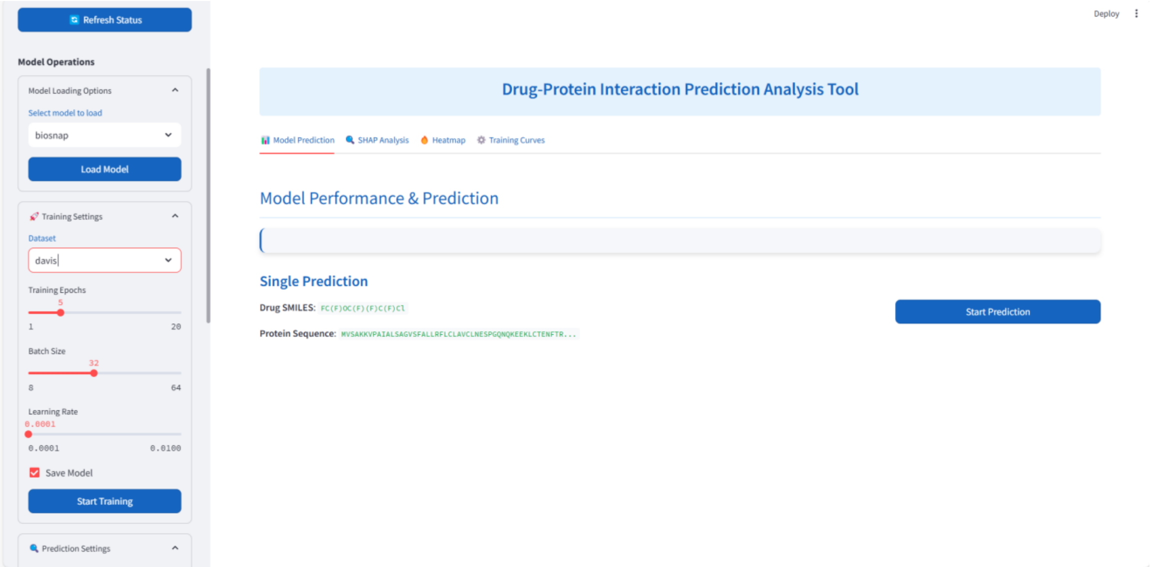
0.9045

↑ Above threshold (0.5)

Prediction: Likely to bind ✓

Figure 28: Web GUI (output will react)

Figure 27 and Figure 28 are the examples of two different results, if the prediction out is that protein and drug will react there will be a pop-up prompt (Likely to bind), otherwise the pop-up prompt will be unlikely to bind.



Model Operations

Model Loading Options

Select model to load: biosnap

Training Settings

Dataset: davis

Training Epochs: 5 (1 – 20)

Batch Size: 32 (8 – 64)

Learning Rate: 0.0001 (0.0001 – 0.0100)

☑ Save Model

Drug-Protein Interaction Prediction Analysis Tool

Model Performance & Prediction

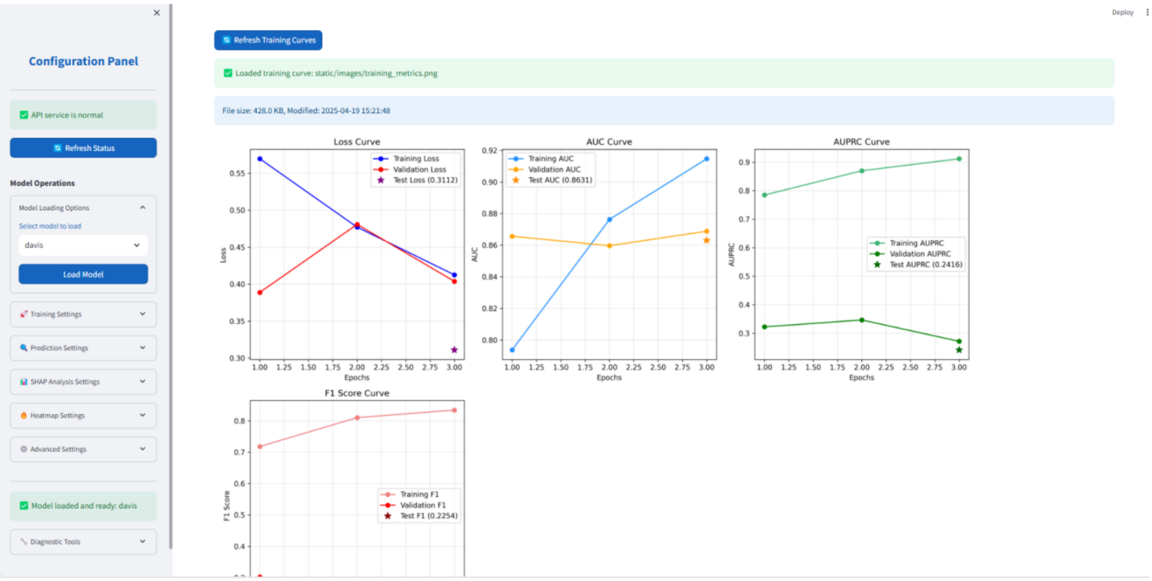
Single Prediction

Drug SMILES: FC(F)OC(F)(F)C(F)Cl

Protein Sequence: MVSAKKVPAIALSAGVSFALLRFLCLAVCLNESPGQNQKEEKLCTENFTR...

Figure 29: Web GUI (Set the training parameters)

Figure 29 shows how the user can set the training parameters for predicting the overall prediction results of the complete dataset. On the sidebar on the left, users can select the epochs to be trained, batch size, learning rate, and whether to save the model weights after training is completed.



☑ API service is normal

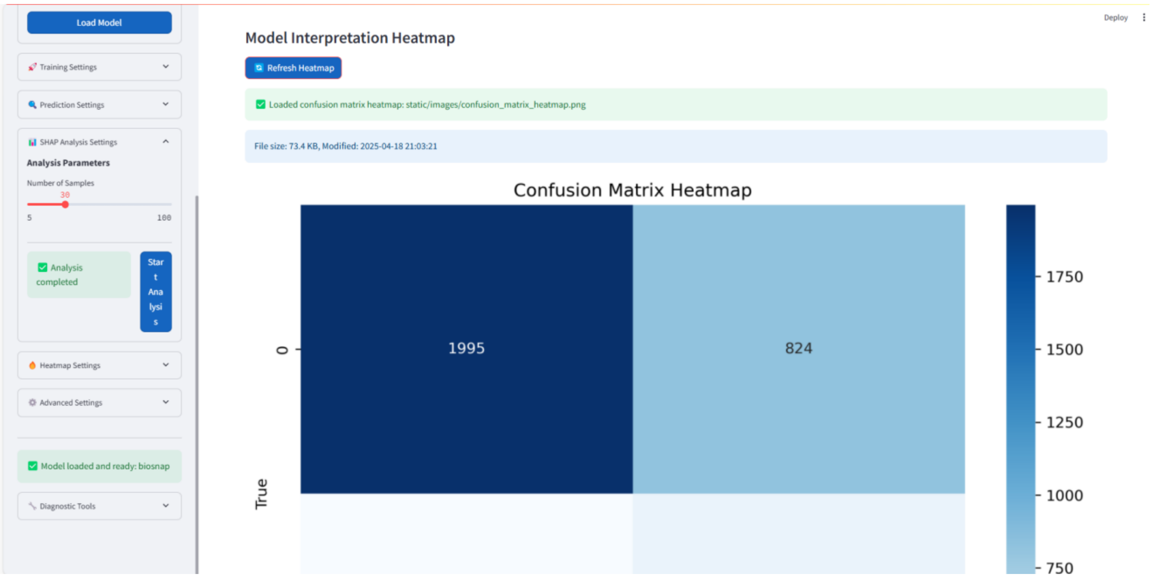
Select model to load: davis

☑ Model loaded and ready: davis

☑ Loaded training curve: static/images/training_metrics.png

File size: 428.0 KB, Modified: 2025-04-19 15:21:48

Figure 30: Web GUI (output Training curve)



Model Interpretation Heatmap

Number of Samples: 30 (5 – 100)

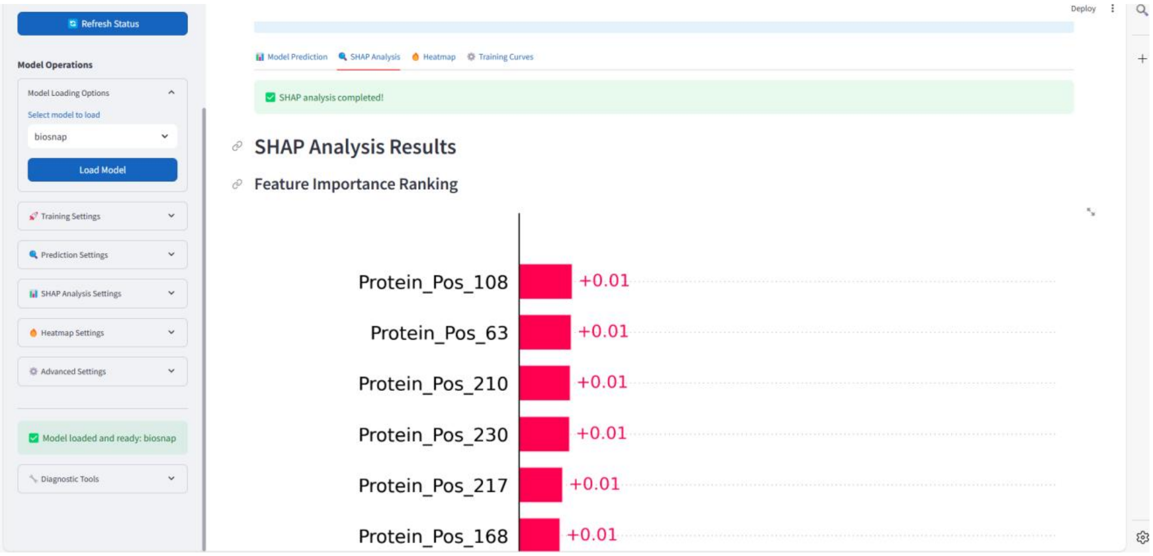
☑ Analysis completed

☑ Model loaded and ready: biosnap

☑ Loaded confusion matrix heatmap: static/images/confusion_matrix_heatmap.png

File size: 73.4 KB, Modified: 2025-04-18 21:03:21

Figure 31: Web GUI (output Confusion matrix heatmap)



Select model to load
biosnap

☑ Model loaded and ready: biosnap

☑ SHAP analysis completed!

SHAP Analysis Results

Feature Importance Ranking

Figure 32: Web GUI (output SHAP plots)

Figures 30,31 and 32 show the implications of explainable technologies of the prediction output, which contain training curves, confusion matrix heatmap and five kinds of SHAP plots.

Chapter 5 Professional Issues

5.1 Project Management

In this chapter, the content will focus on introducing the Activities of the Project, Time planning of the project progress, timetable for every deadline, project data management and deliverables, Risk analysis and the professional issues for this research project.

5.1.1 Activities

Table 9: Activities of the Project

Phase	Objectives
Preparation (Completed)	Review deep learning. Identify and narrow issues Seek possible solutions. Collect related articles.
Deep learning knowledge absorbing (Completed)	Research drug target interaction data Study CNN models and relevant programming libraries. Grasp loss functions, optimizers, model building and optimization. Investigate extra mechanism for suitability.
Data collection (Completed)	Collect 2 - 3 datasets from related websites. Split them into three classes: affinity, SMILES and target sequence. Decide the training, validation and test ratio
Development and Implementation	Build project model. Add Transformer mechanism. Train, analyze, and compare models. Optimize the chosen model and adjust hyperparameters if needed.

Improve models	Change other similar datasets to check the generalization capability of the model. Improve the model, including adding more mechanisms like attention, explainable.
Train and record the result	Change different parameters record the difference of model output. Analyze the results and summarize the work. Use different data sets for model training and check the generality of the model to different data sets
Design a GUI	Conceive the GUI design and draw the intended interface using drawing tools. Gather and learn techniques related to GUI design, for example FastAPI, Streamlit, PyQT. Complete GUI implementation

5.1.2 Schedule

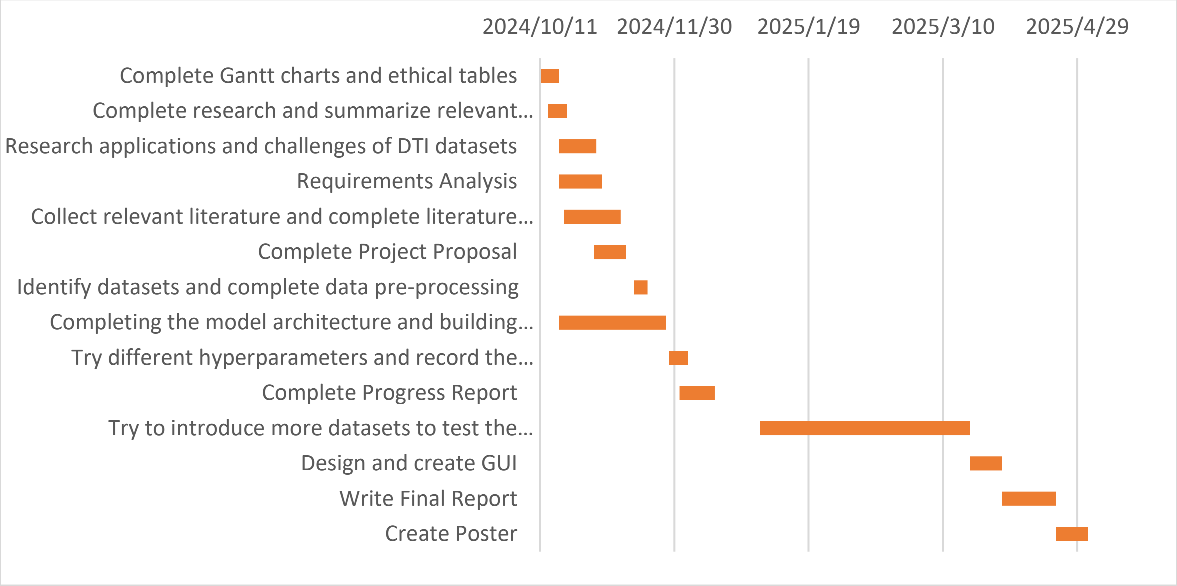


Figure 33: Time planning Gantt chart

Table 10: project timetable

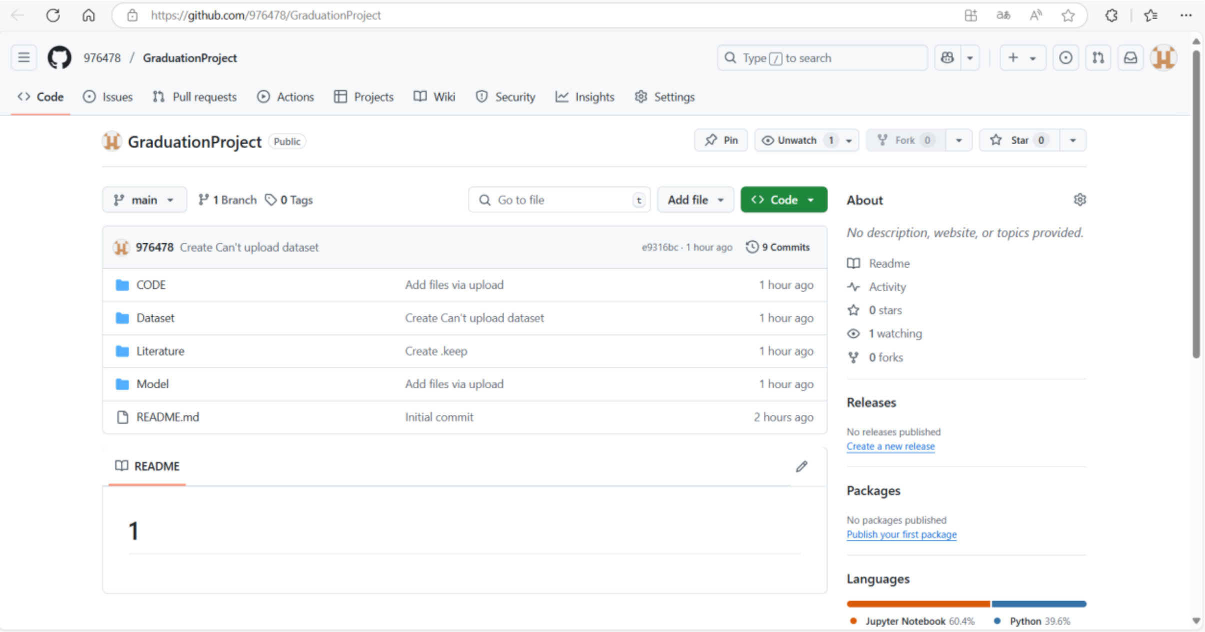
Task	Start Date	End Date	Duration
Complete Gantt charts and ethical tables	2024/10/11	2024/10/18	7
Complete research and summarize relevant literature	2024/10/14	2024/10/21	7
Research applications and challenges of DTI datasets	2024/10/18	2024/11/1	14
Requirements Analysis	2024/10/18	2024/11/3	16
Collect relevant literature and complete literature review	2024/10/20	2024/11/10	21
Complete Project Proposal	2024/10/31	2024/11/12	12
Identify datasets and complete data pre-processing	2024/11/15	2024/11/20	5

Completing the model architecture and building the base model	2024/10/18	2024/11/27	40
Try different hyperparameters and record the effects on the model results	2024/11/28	2024/12/5	7
Complete Progress Report	2024/12/2	2024/12/15	13
Try to introduce more datasets to test the generality of the model	2025/1/1	2025/3/20	78
Design and create GUI	2025/3/20	2025/4/1	12
Write Final Report	2025/4/1	2025/4/21	20
Create Poster	2025/4/21	2025/5/3	12

5.1.3 Project Data Management

All files including datasets, model codes, references, weekly reports and all sorts will be replicated into three copies for file safe, one on local computer, one on Baidu cloud disk, and one on GitHub. load the project to Baidu cloud disk for every modification, synchronize the project on three platforms, once there is a big progress then send the project on GitHub.

https://github.com/976478/GraduationProject



976478 Create Can't upload dataset — e9316bc · 1 hour ago — 9 Commits

CODE	Add files via upload	1 hour ago
Dataset	Create Can't upload dataset	1 hour ago
Literature	Create .keep	1 hour ago
Model	Add files via upload	1 hour ago
README.md	Initial commit	2 hours ago

README

1

No description, website, or topics provided.

No releases published

No packages published

Figure 34: GitHub repositories

5.1.4 Project Deliverables

a. Project Proposal

b. Ethics Form

c. Weekly Report

d. Progress Report

e. Final Report

f. Project Codes

g. Project PPT

h. Project poster

5.2 Risk Analysis

Table 11: potential risk analysis

Risk ID	Potential Risk	Cause ID	Potential Causes	Severity	Likelihood	Risk	Mitigation ID	Mitigation
R1.1	Missed deadline	C1.1.1	Illness	1	3	3	M1.1.1	Register exceptional circumstances if ill.
		C1.1.2	Hesitate in choosing a topic	1	1	1	M1.1.2	Communicate with supervisor as early as possible to complete the topic selection and background investigation
		C1.1.3	Poor time management	4	3	12	M1.1.3	Make a Gantt plan early
R1.2	Feature creep	C.1.2.1	Over-ambitious project spec.	3	2	6	M1.2.1	Discuss plan with supervisor early. Create basic (must-have) goals and enhancements (nice-to-have).
R1.3	Software bugs	C1.3.1	Non-modular design	1	3	3	M1.3.1	Create highly modular desigh before implementation

		C1.3.2	Poor test plan	4	3	12	M1.3.2	Create test plan at start
R1.4	Loss of data or code	C1.4.1	Poor version control	4	4	16	M1.4.1	Implement version control strategy at start.
R1.5	Insufficient processing capability for a large amount of data	C1.5.1	Limited computing resources	4	4	16	M1.5.1	Rent remote GPU resources
R1.6	progress slowly	C1.6.1	No idea of progress	3	4	12	M1.6.1	Ask supervisor or friends for help, also try to find solutions from articles collected

5.3 Professional Issues

Legal Issue:

In the legal context, using deep learning models for drug-target interaction (DTI) prediction involves critical issues such as data privacy, intellectual property, and compliance with regulations. Handling chemical and biomedical data often requires adherence to data protection laws such as General Data Protection Regulation (GDPR). It is crucial to anonymize sensitive data and prevent misuse[28]. Additionally, intellectual property rights surrounding drug-related data and AI-generated predictions must be carefully managed to ensure compliance with legal frameworks and avoid potential conflicts.

Social Issue:

Socially, deep learning access in DTI prediction needs to be balanced across geographies and sectors. The technology promises to speed up drug discovery for patients worldwide[29]. However, if it is only available to rich pharma firms or developed countries, it can exacerbate healthcare inequalities. Building trust based on transparency around model performance and constraints is necessary to get society's buy-in.

Ethical Issue:

Ethically, the implementation of deep learning models in DTI requires addressing biases that may arise from imbalanced datasets. Failure to do so could lead to inaccurate predictions for underrepresented drug classes or target proteins, affecting fairness and reliability. It is essential to maintain the interpretability and transparency of the models, ensuring stakeholders understand the basis of predictions[30]. Furthermore, there must be explicit consent for using proprietary or sensitive datasets in model training.

Environmental Issue:

From an environmental aspect, training deep learning models for DTI prediction consumes massive computational resources, translating into large carbon footprints. This is particularly the case for large datasets like BingDB and DAVIS. Both energy efficiency in training and the application of greener computation practices must be optimized to reduce the environmental cost. Researchers should strive to balance computational demands with sustainability practices[31].

In conclusion, deep learning application in DTI prediction is a complicated mix of legal, social, ethical, and environmental factors. By conforming to professional codes of practice and responsible workflows, it is possible to successfully get past these, enabling the technology to make valuable contributions to drug discovery and medicine.

Chapter 6 Conclusion

In this project, a very efficient deep-learning model for drug-protein interaction prediction is developed to counteract the issues of precisely determining intricate biological relationships in data sets with imbalances. The model uses advanced techniques, i.e., FCS Mining Module, Improved Transformer Embedding Module, and Interaction Prediction Module. Through training and testing on different data sets, i.e., BindingDB, BIOSNAP, and DAVIS, the model in this work achieved better AUC, AUPRC, and F1 scores, indicating higher prediction capability in drug-protein interaction prediction. Furthermore, there were plenty of comparative experiments with other traditional models, once again proving the excellence of the proposed model in accuracy, generalization, and efficiency.

A few limitations of the existing model remain. Due to restricted training materials, techniques such as fixed-length sequence padding were still employed, which could result in the loss of important contextual information and a mild compromise in the model's ability to capture high-level sequence relationships. Class imbalance within the dataset was also a concern, especially concerning infrequent interaction prediction. In order to counter these issues, techniques such as weight balancing of classes in the loss function were adopted, but they were not enough to fully address the issue. Moreover, even though the model used 2D convolutional layers in an attempt to make feature reuse feasible, it still suffers from fine-grained features, and further research would be able to push this field further with more sophisticated techniques. Despite such constraints, the model operated efficiently on validation and testing sets with extremely good AUPRC and AUC scores and could compress the number of parameters efficiently to facilitate efficient inference. Hence, under constraint resources, accurate and efficient prediction of complex drug-protein interactions can be made by applying deep learning approaches carefully and adjusting them.

For future work, this project will focus on enhancing the model’s ability to generalize across different datasets and improving its computational efficiency. Further optimization of the training process, such as through data augmentation or exploring more lightweight model architectures, will be explored to improve performance. In particular, dynamic scene processing and adaptive learning techniques such as transfer learning could be applied to extend the model’s capability to handle diverse biological scenarios. Moreover, beyond linear substructure tokens, integrating graph-based encodings (e.g., GNNs on molecular graphs) and 3D structural information such as using geometric deep learning on protein binding pockets could help capture fine-grained interaction determinants that lie outside the sequential embeddings. Overall, this work offers a modest contribution to drug–protein interaction prediction by highlighting key challenges and potential improvements and suggests several directions for future research to support the application of deep learning in precision medicine.

References

[1] Y. Zhang *et al.*, ‘Drug–target interaction prediction by integrating heterogeneous information with mutual attention network’, *BMC Bioinformatics*, vol. 25, no. 1, p. 361, Nov. 2024, doi: 10.1186/s12859-024-05976-3.

[2] Y. Li, C. Sun, J.-M. Wei, and J. Liu, ‘Drug–Protein interaction prediction by correcting the effect of incomplete information in heterogeneous information’, *Bioinformatics*, vol. 38, no. 22, pp. 5073–5080, Nov. 2022, doi: 10.1093/bioinformatics/btac629.

[3] S. M. H. Mahmud *et al.*, ‘PreDTIs: prediction of drug–target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques’, *Brief. Bioinform.*, vol. 22, no. 5, p. bbab046, Sep. 2021, doi: 10.1093/bib/bbab046.

[4] H. Gao, Y. Tian, R. Yao, F. Xu, X. Fu, and S. Zhong, ‘Exploiting Adversarial Examples to Drain Computational Resources on Mobile Deep Learning Systems’, in *2020 IEEE/ACM Symposium on Edge Computing (SEC)*, San Jose, CA, USA: IEEE, Nov. 2020, pp. 334–339. doi: 10.1109/SEC50012.2020.00048.

[5] A. Capecchi, D. Probst, and J.-L. Reymond, ‘One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome’, *J. Cheminformatics*, vol. 12, no. 1, p. 43, Dec. 2020, doi: 10.1186/s13321-020-00445-4.

[6] Y. Wang, H. Wu, and Y. Cai, ‘A benchmark study of sequence alignment methods for protein clustering’, *BMC Bioinformatics*, vol. 19, no. S19, p. 529, Dec. 2018, doi: 10.1186/s12859-018-2524-4.

[7] K. Huang, C. Xiao, L. Glass, and J. Sun, ‘MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction’, *Bioinformatics*, vol. 37, no. 6, pp. 830–836, May 2021, doi: 10.1093/bioinformatics/btaa880.

[8] K. Huang, C. Xiao, L. Glass, and J. Sun, ‘MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction’, *Bioinformatics*, vol. 37, no. 6, pp. 830–836, May 2021, doi: 10.1093/bioinformatics/btaa880.

[9] K. Huang, C. Xiao, L. Glass, and J. Sun, ‘MolTrans: Molecular Interaction Transformer for Drug Target Interaction Prediction’, *Bioinformatics*, vol. 37, no. 6, pp. 830–836, May 2021, doi: 10.1093/bioinformatics/btaa880.

[10] S. Majumdar *et al.*, ‘Deep Learning-Based Potential Ligand Prediction Framework for COVID-19 with Drug–Target Interaction Model’, *Cogn. Comput.*, vol. 16, no. 4, pp. 1682–1694, Jul. 2024, doi: 10.1007/s12559-021-09840-x.

[11] K. Y. Gao, A. Fokoue, H. Luo, A. Iyengar, S. Dey, and P. Zhang, ‘Interpretable Drug Target Prediction Using Deep Neural Representation’, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 3371–3377. doi: 10.24963/ijcai.2018/468.

[12] B. Liu, D. Papadopoulos, F. D. Malliaros, G. Tsoumakas, and A. N. Papadopoulos, ‘Multiple similarity drug–target interaction prediction with random walks and matrix factorization’, *Brief. Bioinform.*, vol. 23, no. 5, p. bbac353, Sep. 2022, doi: 10.1093/bib/bbac353.

[13] P.-Y. Kao, S.-M. Kao, N.-L. Huang, and Y.-C. Lin, ‘Toward Drug-Target Interaction Prediction via Ensemble Modeling and Transfer Learning’, in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2021, pp. 2384–2391. doi: 10.1109/BIBM52615.2021.9669729.

[14] H. Öztürk, E. Ozkirimli, and A. Özgür, ‘DeepDTA: Deep Drug-Target Binding Affinity Prediction’, *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, doi: 10.1093/bioinformatics/bty593.

[15] H. Wang, G. Zhou, S. Liu, J.-Y. Jiang, and W. Wang, ‘Drug-Target Interaction Prediction with Graph Attention networks’, Jul. 10, 2021, *arXiv*: arXiv:2107.06099. doi: 10.48550/arXiv.2107.06099.

[16] H. Yang *et al.*, ‘MINDG: a drug–target interaction prediction method based on an integrated learning algorithm’, *Bioinformatics*, vol. 40, no. 4, p. btae147, Mar. 2024, doi: 10.1093/bioinformatics/btae147.

[17] L. Zhao, J. Wang, L. Pang, Y. Liu, and J. Zhang, ‘GANsDTA: Predicting Drug-Target Binding Affinity Using GANs’, *Front. Genet.*, vol. 10, p. 1243, Jan. 2020, doi: 10.3389/fgene.2019.01243.

[18] Y. Wang and Z. Yin, ‘Drug–target interaction prediction through fine-grained selection and bidirectional random walk methodology’, *Sci. Rep.*, vol. 14, no. 1, p. 18104, Aug. 2024, doi: 10.1038/s41598-024-69186-w.

[19] Y. Chen, X. Liang, W. Du, Y. Liang, G. Wong, and L. Chen, ‘Drug–Target Interaction Prediction Based on an Interactive Inference Network’, *Int. J. Mol. Sci.*, vol. 25, no. 14, p. 7753, Jul. 2024, doi: 10.3390/ijms25147753.

[20] Y. Wang *et al.*, ‘LDS-CNN: a deep learning framework for drug-target interactions prediction based on large-scale drug screening’, *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 42, Sep. 2023, doi: 10.1007/s13755-023-00243-w.

[21] S. Z. Sajadi, M. A. Zare Chahooki, S. Gharaghani, and K. Abbasi, ‘AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders’, *BMC Bioinformatics*, vol. 22, no. 1, p. 204, Apr. 2021, doi: 10.1186/s12859-021-04127-2.

[22] Y. Sun, Y. Y. Li, C. K. Leung, and P. Hu, ‘iNGNN-DTI: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models’, *Bioinformatics*, vol. 40, no. 3, p. btae135, Mar. 2024, doi: 10.1093/bioinformatics/btae135.

[23] L. Zhou and K. K. Lai, ‘Weighted LS-SVM Credit Scoring Models with AUC Maximization by Direct Search’, in *2009 International Joint Conference on Computational Sciences and Optimization*, Sanya, Hainan, China: IEEE, Apr. 2009, pp. 7–11. doi: 10.1109/CSO.2009.333.

[24] S. A. Khan and Z. Ali Rana, ‘Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)’, in *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, Lahore, Pakistan: IEEE, Feb. 2019, pp. 1–6. doi: 10.23919/ICACS.2019.8689135.

[25] B.-C. Yan, H.-W. Wang, S.-W. F. Jiang, F.-A. Chao, and B. Chen, ‘Maximum F1-Score Training for End-to-End Mispronunciation Detection and Diagnosis of L2 English Speech’, in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan: IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/ICME52920.2022.9858931.

[26] J. Sepúlveda and S. A. Velastin, ‘F1 Score Assesment of Gaussian Mixture Background Subtraction Algorithms Using the MuHAVi Dataset’, in *6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15)*, London, UK: Institution of Engineering and Technology, 2015, p. 8 (6 .)-8 (6 .). doi: 10.1049/ic.2015.0106.

[27] L. Ran and C. Cai, ‘Topo-Loss: A Novel Loss Function for Learning Multiple Manifold Structures in Image Classification with Deep Neural Networks’, in *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Zhuhai, China: IEEE, Apr. 2024, pp. 703–707. doi: 10.1109/CVIDL62147.2024.10604103.

[28] A. Malloy, ‘Legal and ethical issues in the regulation and development of engineering achievements in medical technology: A 2006 perspective’, in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY: IEEE, 2006, pp. 6660–6662. doi: 10.1109/IEMBS.2006.260914.

[29] N. Quan, S. Ma, K. Zhao, X. Bi, and L. Zhang, ‘MFCADTI: improving drug-target interaction prediction by integrating multiple feature through cross attention mechanism’, *BMC Bioinformatics*, vol. 26, no. 1, p. 57, Feb. 2025, doi: 10.1186/s12859-025-06075-7.

[30] S. Du, T. Li, Y. Yang, and S.-J. Horng, ‘Deep Air Quality Forecasting Using Hybrid Deep Learning Framework’, *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2412–2424, Jun. 2021, doi: 10.1109/TKDE.2019.2954510.

[31] R. Xu, D. Wang, J. Li, H. Wan, S. Shen, and X. Guo, ‘A Hybrid Deep Learning Model for Air Quality Prediction Based on the Time–Frequency Domain Relationship’, *Atmosphere*, vol. 14, no. 2, p. 405, Feb. 2023, doi: 10.3390/atmos14020405.