Multi-Task Learning Framework for Drug–Target Interactions and Adverse Effects Prediction

by

Md. Sabbir Hossain Mirza	23241086
Fahmid Hasan Chowdhury	21201286
Zakaria Ibne Rafiq	24341150
Radito Dhali	24141150
Md. Rakibul Hasan Talukder	23241100

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University

January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Sabbir Hossain Mirza

23241086

Fahmid Hasan Chowdhury

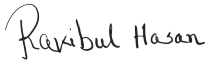
21201286

Zakaria Ibne Rafiq

24341150

Radito Dhali

24141150



Md. Rakibul Hasan Talukder

23241100

Approval

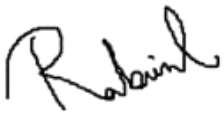
The thesis titled "Multi-Task Learning Framework for Drug–Target Interactions and Adverse Effects prediction" submitted by

1. Md. Sabbir Hossain Mirza (23241086)
2. Fahmid Hasan Chowdhury (21201286)
3. Zakaria Ibne Rafiq (24341150)
4. Radito Dhali (24141150)
5. Md. Rakibul Hasan Talukder (23241100)

of Fall 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in January 31, 2026.

Examining Committee:

Supervisor:
(Member)



Dr. Md. Golam Rabiul Alam

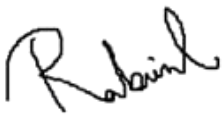
Professor
Department of Computer
Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Dr. Swakkhar Shatabda

Professor
Department of Computer
Science and Engineering
BRAC University

Thesis Coordinator:
(Member)



Dr. Md. Golam Rabiul Alam

Professor
Department of Computer
Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor
Department of Computer
Science and Engineering
BRAC University

Abstract

Drug-target interactions (DTIs) and adverse drug reactions (ADRs) are biological processes that interact closely but are difficult to model together, preventing the investigation of off-target effects and system perturbations that underlie drug safety. We introduce a single deep-learning system to predict both DTI and ADR using diverse molecular and protein representations of drug SMILES sequences and three-dimensional molecular graphs as well as protein sequences and structural features provided by AlphaFold. Curated DTI and drug-ADR data have a common RxNorm identifier that facilitates the cross-task correspondence between these two data.The proposed context-aware multi-task model uses variational auto-encoder bottleneck to both regularize shared latent space and multihead predictors for binary DTI classification and multi-label ADR with strong label imbalance. In contrast to the previous methods where interaction and safety modeling are decoupled or a single-modality evidence is used, in the model, drug-protein-ADR context is learned jointly. At the drug and protein concentrations, cold-start split analysis show decision-relevant predictions, which are calibrated, and good extrapolation to unfamiliar objects. In addition to predictive performance, the pipeline has a modular and reproducible prediction framework between featurization and inference enabling scalable experimentation. It continues the integration of DTI-ADR modeling as a conceptual method of early safety triage, risk-conscious virtual screening and translational drug discovery. Under stringent protein-level cold-start evaluation, the proposed framework achieves strong and stable performance, attaining a DTI AUROC of 91.60%, AUPRC of 85.46%, and F1-score of 78.39%, alongside ADR prediction with a weighted AUROC of 98.80% and weighted AUPRC of 94.54%, consistently reproduced across multiple random seeds.

Keywords: *Drug–Target Interaction (DTI), Adverse Drug Reaction (ADR), Multi-Task Deep Learning, Multimodal Representation Learning, ChemBERTa, ESM, Graph Neural Networks (GNNs), EGNN, GVP-GNN, Fusion Encoder,Variational AutoEncoder, Computational Drug Discovery*

Acknowledgement

First and foremost, we are thankful to the Almighty for giving us the chance and the path we needed to finish this project on time. Secondly, we would like to sincerely thank our thesis supervisor, Dr. Md. Golam Rabiul Alam and our co-supervisor, Dr. Swakkhar Shatabda for their constant encouragement, mentorship, and guidance as we explored a challenging topic. Their relentless assistance and continuous input enabled us to face and overcome the challenges. Finally, we intend to use this opportunity to express our gratitude to every faculty member for their assistance and support during our stay at Brac University.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the background of the study, the motivation and objectives of the research, and the reasons why a single framework to predict Drug-Target Interactions and Adverse Drug Reactions is necessary. Also, we will examine the theoretical context of the study, how it combined the various biological modalities of data, the overall scope of the study, and what issues it aimed to resolve.

1.1 Background

Drug-Target Interaction (DTI) prediction is an essential part of drug discovery in the modern era and the concept is to determine the interaction between the chemical compounds and the protein targets. The study of such interactions is of central significance since it defines the therapeutic potential of a drug in addition to the undesired biological effects of the drug. The classical methods of screening of the discovery of DTIs via the traditional laboratory techniques are accurate but time-intensive, costly and inefficient to conduct at the necessary scale to test millions of potential compounds. This has promoted the application of computational techniques that use machine learning and deep learning models on predicting the probability of interactions in a more effective way.

Defining whether a drug is bound to a protein is however not the only aspect of drug safety. A higher number of approved drugs, despite having good affinity to bind to their targets of interest, may result in Adverse Drug Reactions (ADRs) - unwanted or unintended physiological reactions including minor side effects or potentially fatal complications. ADR is a key challenge to pharmaceutical development and is a significant cause of drug post-marketing withdrawal. Computational prediction of such reactions has therefore become another significant field of study.

The recent advances in deep learning and specifically in multimodal structures able to incorporate heterogeneous biological information such as molecular graphs, SMILES representations, protein sequences and 3D structural data have provided new opportunities to improve both DTI and ADR prediction. However, current researchers tend to consider these two issues as two different tasks without the biological connection to each other. The mechanisms that regulate the drug-protein interactions are usually closely related to the mechanisms that lead to adverse effects. Therefore, as a combined predictor of DTI and ADR, it is possible that joint

modelling of the two factors in a single framework could provide a more thorough perspective of the drug behaviour and potentially yield more interpretable and safer computational drug discovery systems.

The study proposes one multi-task deep learning system to predict drug-target and adverse drug reaction in a single architecture. The model is based on the joint representation of multimodal drugs and proteins, which include chemical sequence encodings, molecular graph encodings, protein sequence encodings and structure-sensitive encodings, in a common interaction space. Variational latent bottleneck regulates the combined representation and facilitates compact and generalizational feature learning. The framework will discover biologically meaningful interactions between binding behavior and downstream clinical outcomes and can significantly improve extrapolation in drug-protein perturbations that have never been experimentally observed by uniting prediction of interactions and adverse reactions of drugs based on a shared latent representation.

1.2 Rational of the Study or Motivation

In the era of big data, drug discovery is becoming a data-rich activity that requires more than just the successful identification of therapeutic targets: it must increasingly account for the possibility of side effects that could be present at an earlier stage. Traditional computational pipelines address these issues separately. In Drug-Target Interaction (DTI) models, the goal is to predict molecular binding strength or affinity. In Adverse Drug Reaction (ADR) models, the aim is to find connections between drugs and side effects reported in clinical practice. While this division seems useful methodologically, it overlooks the biological continuity of a drug from its molecular action to its effects on human physiology.

The weakness of these approaches becomes obvious when considering safety issues after approval. Drugs with a good degree of target selectivity can still result in unexpected side effects as a consequence of interactions with other proteins or off-target binding. As discussed in [2] and [3] even more exact DTI predictors may leave downstream phenotypic consequences in the cold, if adverse outcomes are not accounted for at the prediction level. Likewise, chemistry-based, text co-occurrence based and ADR-based (ADR stands for Application Data Rate) predictors that predict only the occurrence of the side effect have limited interpretability and fail to explain why a particular ADR occurs.

The recent work by [4] and [24] suggests that structural and spatial context from proteins and compounds can highly assist the interaction prediction when exploited in conjunction with a graph-based or cross-attention design. But these studies have not yet been connected to molecular interactions that are related to the actual pharmacological effects in the real world, such as ADRs. The lessons learned from these experiments hint at the fact that multimodal embeddings can represent interaction, including the adverse effects.

This fact inspires the emergence of a DTI–ADR prediction model that can represent both dimensions of drug action, therapy, and side effects. A common model can use

shared representations of molecular and biological characteristics to uncover interdependent patterns, which are otherwise obscured, when tasks are learned separately. In addition, the inclusion of contrastive learning [7] principles and enabling parameter sharing across chemical, protein, and phenotypic domains enables the model to learn from consistent mappings across modalities, leading to better generalization toward new compounds and targets.

Through integrating these two predictive fields, the present proposed approach seeks to advance a more comprehensive and interpretable computational paradigm for drug discovery, one capable of prioritising safer compounds, enabling faster repurposing of drugs, and decreasing expensive late-stage development failures.

1.3 Problem Statement

Despite the dramatic advances of deep learning for Drug–Target Interaction (DTI) prediction, prior computational modeling has been hindered by limitations in capturing intricate biological and chemical relationships that lead to both efficacy and side effects. Most previous works such as [2], [3] and [22] only predict the binding affinities or binary interaction labels, with no consideration of bridging these latent representations to the pharmacological responses. Their predictions, when accurate in vitro, typically do not generalize to reliable in vivo performances. Likewise, Adverse Drug Reaction (ADR) predictions based on text-mining or similarity-based approaches have not captured the molecular and target-level reasoning for how those reactions originated. This disassociation of DTI and ADR prediction results in a lack of knowledge regarding how the drug's binding profile relates to side effects, which is an essential gap when designing safer and more explainable drug discovery pipelines.

In order to bridge this gap, we propose to develop a unified framework that can simultaneously learn from DTI and ADR data, based on a shared representation that explores the relationships between drug binding patterns and the resultant adverse observations. This method faces three primary challenges: heterogeneity in biological data types, imbalance between positively and negatively labeled interactions, and lack of direct mapping between DTI and ADR signals provided by existing datasets. In this direction, we plan to build a multi-task deep learning model, which incorporates heterogeneous drug and protein representations into a shared latent space, which is jointly optimized to predict interaction and adverse reactions profiling using end-to-end multi-task supervision. This further allows interpretable and generalizable prediction of drug-target interactions and ADR potential.

1.4 Objective

The main goal of this research is to create a multi-task deep learning framework that can predict Drug-Target Interactions (DTIs) and Adverse Drug Reactions (ADRs) together using a single structure. The framework aims to bridge the current divide between therapeutic interaction modeling and adverse effect prediction by learning shared representations that reflect the underlying biochemical and structural dependencies among drugs, proteins, and clinical outcomes.

To achieve this goal, the study is guided by the following specific objectives:

- To construct an integrated dataset that aligns DTI and ADR data through standardized drug identifiers (e.g., RxNorm), ensuring consistent mapping between drug-protein pairs and associated adverse effects.
- To generate and compare multiple types of embeddings for both drugs and proteins using diverse encoders- ChemBERTa, EGNN and SMILES2Vec for chemical representations; ProtVec, ESM, GVP for protein representations, capturing both sequence-level and structural information.
- To create a unified multi-task deep learning model, which has a shared latent representation, where the fused drug-protein features are shared and sent through task-specific prediction heads to undertake binary Drug-Target Interaction classification and multi-label Adverse Drug Reaction prediction.
- To regularize the shared drug protein representation with a variational latent bottleneck, that allows frugal and sensible feature learning to enhance generalization to unseen drug protein pairs and concurrently learn interaction prediction and adverse reaction profiling.

Through these objectives, the study seeks to demonstrate that a multi-head, multimodal, and geometry-aware deep learning framework can effectively capture both therapeutic and toxicological dimensions of drug behavior—offering a more reliable and mechanistically grounded approach to computational drug discovery.

1.5 Methodology in Brief

This study is based on the systematic pipeline that incorporates the areas of dataset integration, generating multimodal embeddings, and training multi-task models. These steps involve curation and mapping of the DTI and ADR data sets, and the identifiers of the drugs are standardized via RxNorm mapping to offer some degree of uniformity across the data sets. Every drug protein pair in the DTI data is associated with the known adverse reactions and this is used to construct a single foundation in which they can be predicted together.

Feature representations are created using various biological modalities. SMILES2Vec and ChemBERTa are used to obtain drug embeddings, whereas ProtVec, ESM, GVP, and EGNN are used to obtain protein embeddings, which represent both sequence-level and 3D structure-based information. The encoding of ADRs is based on multi-label representations obtained with the help of text descriptions of adverse events.

The embedded fused drug and protein representations are trained using a common context encoder and variational latent bottleneck which regularises the joint representation and encourages compact and generalizable features learning. Following this shared latent space two task-conditioned prediction heads are trained in parallel: one predictor is trained to make a binary prediction to make predictions of Drug-Target Interaction; the other is a multi-label prediction to make predictions

of the probability of adverse drug reactions. Training The model is trained in an end-to-end manner through multi-task supervision.

This model has been trained and tested on different drug and protein embedding combinations to determine the most successful one. The standard metrics of performance are used to measure performance, with special consideration on extrapolation of unobserved drugs and targets.

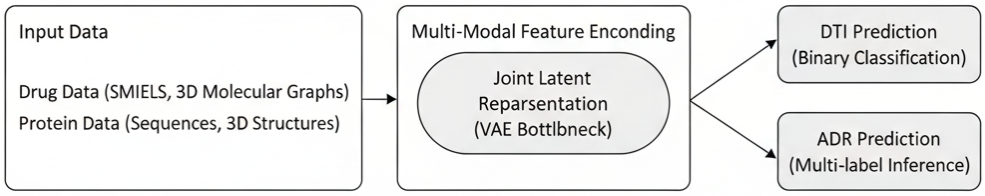


Figure 1.1: Multi-Task Learning

1.6 Scopes and Challenges

The suggested architecture has the potential of helping discover drugs through computational methods because it allows predicting Drug-Target Interactions (DTIs) and Adverse Drug Reactions (ADRs) at the same time, in one model. Its uses cover an extensive number of areas in pharmaceutical research such as drug repurposing, side-effect profiling and safety estimation of targets. With a mixture of multimodal sources of data, including chemical sequences, molecular graphs, and structure of a protein and even the textual description of ADR, the model will build a more comprehensive view of drug behavior that is expected to be generalizable to new targets and new molecules. Moreover, it is easy to extend it with other forms of embedding or even to new biological modalities since it is implemented in a modular fashion, and reconfigured to new biomedical data.

However, a lot of challenges are linked to this ambition. One of the common problems is data imbalance (the number of non-interaction drug-protein pairs is much larger than the positive interaction ones), which may result in biased learning for models. Another issue is the heterogeneity of data modality (chemical, structural and textual) that makes it hard to match embeddings in a common latent space. Moreover, since there is no one-to-one mapping between some DTIs and their ADRs, there is an indirect relational learning and careful loss design. Lastly, some of the high-dimensional 3D encoders like GVP, EGNN unavoidably result in weighty models which incur higher training costs or even need some sort of hardware acceleration.

Despite these limitations, the conceptual model proposed in this study offers a viable route to a truly multimodal and general-purpose predictive pharmacology system, contributing to safer and more expeditious drug discovery.

1.7 Thesis Organization

The thesis is designed to achieve the gradual structure of motivation, methodology, implementation, and assessment of a single multi-task deep learning model to predict both Drug-Target Interaction (DTI) and Adverse Drug Reaction (ADR).

Chapter 1 presents the problem domain and sets the motive of joint modeling of DTIs and ADRs. It provides an overview of the computational drug discovery background, states the problem statement, outlines the objectives of the research, and briefs the proposed methodology, and comment on scope and major issues considered in this research.

Chapter 2 is a review of related literature, which includes a number of introductory concepts, classical machine learning techniques, deep learning techniques to predict DTI and ADRs, multi-modal representation learning, graph neural networks, and up-to-date multitask learning models. This chapter critically discusses research that is available and shows gaps that drive the proposed integrated DTI-ADR framework.

Chapter 3 details the system requirements, constraints and the greatest impacts of the proposed work. It outlines the functional and technical standards of the framework and the ethical, societal, environmental and economic concerns, and sets out the plan of project management and risk mitigation.

The proposed methodology is described in **Chapter 4**. It describes the general system structure, construction of multimodal features with drugs and proteins, the encoding strategy of ADR fusion and joint context, variational autoencoder bottleneck, task-specific prediction head, loss form, training setup, and the entire data preprocessing and integration process.

Chapter 5 gives the experimental assessment, and evaluation results. It also reports performance in stringent cold-start conditions at the protein level, architectural and design choices, statistical and stability analysis over a number of different random seeds, and compares the proposed framework to the corresponding baseline models on both DTI and ADR prediction.

Lastly, **Chapter 6** brings the thesis to an end by re-recapping the most interesting findings and contributions of the study, the importance of the proposed unified framework to computational drug discovery and provides the future research directions and possible extensions.

Chapter 2

Literature Review

This chapter initially covers the concepts, definition, and mechanisms of multiple machine learning (ML) and deep learning (DL) models to understand the baseline related to our paper. Secondly, several papers in relation to DTI and the ADR prediction were searched, including their methodology and pre-processing stages. Finally, the main results of those research papers are extracted: multiple frameworks on DTI and ADR technique preprocessing and prediction.

2.1 Preliminaries

2.1.1 Theories and Concepts

Drug–Target Interaction (DTI)

DTI is the biochemical action of a molecule (drug) binding to a protein target that results in a physiological or therapeutic effect. The ability to predict such interactions using computational approaches would be advantageous for the early identification of potential candidate drugs, without relying only on expensive lab results.

Binding Affinity and Specificity

How strongly a drug binds to its target, also known as binding affinity, is what determines the drug's potency; specificity represents how selectively it can bind to the intended target versus unintended targets. Both are key to understanding whether a drug works and what its side effects may be.

Ligand-Based and Structure-Based Paradigms

Two paradigms drove classical computational drug discovery. Ligand-oriented methods, like QSAR models, are based on the assumption that compounds sharing similar structural features possess similar biological properties. Structure-based techniques, such as molecular docking, model how a drug binds in three dimensions to the binding site of a target protein in order to predict interaction strength. Every such strategy relies a lot on handcrafted features and experimentally determined structures.

Data-Driven Learning in Drug Discovery

With the emergence of large bioactivity data sets, DTI prediction was recast as a supervised learning problem based on machine-learning techniques. Deep neural networks now directly extract novel, latent molecular representations from raw sequences or graphs, so that we no longer depend on expert-designed features and will have better generalization ability for unseen drug–protein pairs.

Adverse Drug Reactions (ADRs)

ADRs are unwanted and mostly adverse effects resulting from drug use. Most ADRs are due to drug off-target interactions that induce the drugs of interest on unintended proteins. This concept overlap between DTI and ADR prediction suggests a common biochemical basis of the therapeutic effect as well as of side effects.

Multi-Task Pharmacological Modeling

Especially since DTIs and ADRs are causally associated, joint modeling of DTIs and ADRs in an integrated framework enables a system to capture both binding associations as well as their clinical implications. In this way multi-task learning forms a theory-inspired bridge between molecular-level prediction and organism-level phenotypes.

2.1.2 Models and Methodologies

Conventional Computational Models

Earlier computational approaches in drug discovery were based on machine learning techniques such as Support Vector Machines (SVMs), Random Forests (RF), and k-Nearest Neighbors (k-NN). These models relied on predefined molecular descriptors like fingerprints or physicochemical features to classify drug–target pairs. While interpretable and computationally efficient, their performance was limited by the expressiveness of handcrafted features and their inability to generalize to unseen drugs or proteins.

Deep Learning-Based DTI Models

Deep learning introduced a major paradigm shift by enabling end-to-end feature learning from raw molecular and sequence data. Models such as [2] used convolutional neural networks (CNNs) to extract features from SMILES strings and amino acid sequences, while [5] and [22] expanded on this by integrating word-level and graph-based encodings. These architectures learn complex, nonlinear mappings between drug and protein representations, yielding superior binding-affinity prediction performance.

Graph Neural Networks (GNNs) for Structural Learning

To capture spatial and relational information inherent in molecular structures, Graph Neural Networks have become increasingly popular. Models like GVP-GNN and EGNN incorporate 3D geometric data by maintaining equivariance to spatial transformations, allowing them to model atomic-level interactions more accurately. These

methods move beyond 1D sequence models, directly learning from the topology and geometry of molecular graphs derived from RDKit or AlphaFold structures.

Transformer Architectures and Pretrained Models

Transformer-based encoders such as ChemBERTa for molecules and ESM for proteins leverage self-attention to learn contextualized representations from large unlabeled corpora. By fine-tuning these pretrained embeddings for DTI prediction, researchers achieve improved generalization across diverse chemical and biological domains. Such models also serve as modular components that can be reused across multiple pharmacological tasks.

Multi-Task Deep Learning Frameworks

Modern research recognizes the interdependence between DTI and ADR prediction. Multi-task frameworks share a common backbone to learn generalized representations and employ separate output heads for distinct objectives—such as DTI interaction and adverse effect classification. This structure enables knowledge transfer between related tasks, improving performance on both through shared inductive signals

Variational Autoencoder

A Variational Autoencoder (VAE) is a generative neural network that is trained to encoder input data into a compact and continuous latent type of data. A VAE does not map inputs to one fixed encoding, but instead the latent space is a probability distribution, which enables the network to express uncertainty and variability in the data. In training, the model is trained to restructure the input whilst regularizing the latent space to a known distribution, which is usually a Gaussian. Such regularization promotes regular and smooth latent representations, capable of enhancing generalization to unobserved samples and more robust downstream predictions.

Proposed Multi-Head Architecture

The suggested framework uses an integrated multi-task deep learning framework that takes into account multimodal drug and protein presentations into a common latent space. The features of drugs are obtained with the help of ChemBERTa and EGNN and protein ones are acquired with the assistance of ESM and GVP-GNN encoders. These fused representations are put in a joint context encoder and variational autoencoder bottleneck to encourage learning compact and generalizable features. Two task-specific heads are trained simultaneously out of the common latent representation: one is used in binary Drug-Target Interaction prediction and the other one is used in multi-label Adverse Drug Reaction prediction.

2.1.3 Embedding and Data Representation Techniques

Molecular Embeddings for Drugs

Drug molecules are typically encoded by their Simplified Molecular Input Line Entry System (SMILES) strings, which capture the atomic composition and interac-

tion in a linear, text-based abstract. These sequences are then passed into deep learning models such as SMILES2Vec and ChemBERTa [6] etc., to map them to high-dimensional representations, which are able to capture structural and chemical semantics. SMILES2Vec gains distributed representation by introducing infrastructure based on the RNN or CNN models, and ChemBERTa exploits context information between atoms and substructures by leveraging the Transformers. These embeddings allow models to learn useful representations of molecular similarity, polarity, and reactivity based on pattern matching instead of handcrafted descriptors.

3D Structural Representations of Drugs

In addition to 1D sequence encoding, 3D coordinates of molecular structures (from conformers generated by RDKit) contain more complete spatial information. Graph-based encoders such as EGNN (Equivariant Graph Neural Network) directly act on such 3D graphs, learning features invariant under rotation and translation. Including such geometric embeddings enables the model to encode molecular shape, charge distribution, and interaction potential, which is essential for accurate binding affinity prediction.

Protein Sequence Embeddings

Protein targets are primarily represented through their amino acid sequences. ProtVec and ESM (Evolutionary Scale Modeling) are widely adopted sequence encoders. ProtVec uses skip-gram models inspired by natural language processing to learn representations of overlapping k-mers, capturing evolutionary motifs and functional similarities. ESM, a large-scale Transformer-based model trained on millions of protein sequences, produces contextualized embeddings that encode secondary structure and residue-level interactions, providing a robust basis for downstream DTI prediction. [1]

Protein Structural Embeddings

With the advent of AlphaFold, 3D structural data for proteins has become widely accessible. Graph-based models like GVP-GNN (Geometric Vector Perceptron GNN) effectively encode these structures by processing both scalar (amino acid features) and vector (spatial orientation) information. This dual representation allows the model to understand geometric dependencies within protein structures, enabling it to capture the spatial complementarity between drug molecules and their binding sites.

Adverse Drug Reaction (ADR) Representation

Adverse Drug Reactions (ADRs) are represented using a multi-label binary encoding scheme based on standardized MedDRA identifiers. A fixed ADR vocabulary is constructed from the curated dataset, where each ADR corresponds to a unique index in the label space. For each drug, the presence or absence of an ADR is encoded as a multi-hot vector, allowing multiple adverse reactions to be associated with a single compound simultaneously. Unlike drug and protein features, ADRs are not embedded as input representations; instead, they are incorporated as supervision

targets during training. This formulation preserves the discrete clinical nature of ADRs while enabling joint optimization with drug–target interaction prediction.

Joint Context Encoding

The fused drug and protein representations are combined to form a joint interaction context, which is then processed by a context encoder. This component consists of a lightweight feed-forward neural network that transforms the concatenated drug–protein features into a compact interaction-aware representation. The context encoder enables the model to capture higher-level dependencies between molecular and biological characteristics that are not evident when the modalities are considered independently. By projecting heterogeneous features into a common interaction space, it provides a coherent representation that serves as the foundation for downstream prediction tasks.

2.2 Review of Existing Research

Ye et al. (2021)[14] addressed that DTI prediction is one of the key problems for drug discovery. A major challenge in DTI prediction is that both drugs and targets have complex graph based structures and, hence, are not amenable to traditional deep learning methods such as deep neural networks (DNNs) and convolutional neural networks (CNNs). The authors make a proposal of a multiple output graph convolutional network (MOGCN) which has the ability to enhance learning capabilities and increase the accuracy of DTI prediction. The method introduces auxiliary classifier layers with effective gradient propagation and effectively optimizes feature use at multiple levels. Moreover, the study uses a new graph calculation method which takes advantage of label information and manifold learning to form a meaningful adjacency matrix. The approach involves using stacked autoencoders (SAE) to obtain low level features, concatenating features and using graph convolutions. The proposed MOGCN was evaluated on five benchmark datasets, NR, GPCR, IC, E, and DB, and compared with baseline DNN baseline and original GCN. Experimental results show that MOGCN performs better than existing methods, and especially so when SAE based low level feature extraction is used. Based on the obtained results, MOGCN can tackle the gradient vanishing problem and at the same time improve the graph structures for DTI prediction, resulting in improved classification performance. The authors also discuss future work in increasing the depth and width of GCNs using residual network techniques and improving the graph calculation framework to capture more accurately relationships between interacting and non-interacting drug-target pairs.

Thafar et al. (2020)[8] addressed the challenge of silico drug target interaction (DTI) prediction as it is crucial for drug discovery and repurposing. Currently, the known computational methods suffer from an inordinate amount of false positive rates, and in short supply are experimentally validated DTI pairs. In order to solve this, the authors proposed DTiGEMS+, a framework that combines Graph Embedding, Graph Mining, and Similarity-based techniques to predict DTIs as a link prediction problem in a heterogeneous network. In this regard, the proposed approach builds a larger network with available DTIs, the drug–drug similarity network, and the

target–target similarity network, which reflects a wider view of the drug target relationships. Multiple computational strategies have been employed in DTiGEMS+, namely similarity fusion, graph embedding using node2vec, path based feature extraction and machine learning classifiers like neural networks, random forests and AdaBoost for final classification. DTiGEMS+ was evaluated on four benchmark datasets (NR, GPCR, IC, and E) and proved the highest average AUPR (0.92), 33.3% relative error rate reduction over the state of the arts. Biomedical literature and databases such as DrugBank and KEGG were used by the authors to validate novel predicted DTIs. As a future work, they propose improving the model by adding more similarity measures in the model, fine tuning the negative sampling strategies and enabling the framework to predict interactions for brands and new drugs and targets.

Mahmud et al. (2019)[4] addressed the challenge of drug discovery as a costly and time consuming process that is grounded on wet-lab experiments. A major limitation of DTI datasets and thus classification accuracy is the class imbalance issue identified in the study. In order to solve this issue, they came up with a high throughput computational model iDTi-CSsmoteB, which uses drug chemical structures and protein sequences to enhance DTI identification. Protein sequence features are extracted from PSSM-Bigram, AM-PseAAC, and dipeptide PseAAC descriptors and drug chemical structures are represented using molecular substructure fingerprint (MSF). The synthetic minority over-sampling technique (SMOTE) was utilized to handle the class imbalance problem and the XGBoost was used as the main classifier to predict DTIs. It was evaluated on four benchmark datasets (enzyme, ion channel, GPCR, and nuclear receptor) by means of five fold cross validation. The experimental results showed that iDTi-CSsmoteB achieves better area under the ROC curve (auROC) than the other existing methods, which verified the effectiveness of the proposed feature extraction techniques, data balancing methods, and classification model. Further, the study proposes that the model can be used in drug repurposing and novel drug discovery. As future works, semi supervised learning techniques and introducing precision recall metrics can be considered to improve the performance further.

Zhan et al. (2020)[9] study aimed to overcome the challenge of predicting drug–target interactions (DTIs) due to their importance in drug discovery and understanding biological processes. Identifying DTIs is costly, time consuming and labor intensive, and reliable computational methods are required to fill the gap. An ensemble learning approach using Local Optimal Oriented Pattern (LOOP), Position Specific Scoring Matrix (PSSM), and Rotation Forest (RF) is proposed by the authors to predict DTIs effectively. In particular, protein sequences are converted by PSSM while preserving evolutionary information, feature vectors are then extracted by LOOP from PSSM to reflect local patterns, and drug molecules are represented by fingerprint features. Finally, RF is applied as the classification model to infer DTIs. Four benchmark datasets, enzyme, ion channel, G protein-coupled receptors (GPCRs), and nuclear receptor, are used to evaluate the model with high average prediction accuracies of 89.09%, 87.53%, 82.05%, and 73.33% respectively. A comparison is performed with the state of the art methods like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and it is shown that the proposed approach is

superior. Additionally, the study shows that the method enhances predictive performance but it relies on manual feature extraction that brings noise as well as ignores global structural information. To further boost the prediction accuracy of DTI, the authors propose that more robust feature extraction and augmentation or boosting of classification models would help.

Thafar et al.(2022)[17]addressed that drug-target binding affinity (DTBA) prediction is crucial for drug repositioning and virtual drug screening.. Most existing methods are based on three dimensional (3D) structural data of targets, which are frequently not available, or difficult to obtain. The authors propose Affinity2Vec, a regression based model that formulates the DTBA problem as a graph based task without the need for 3D structural information to overcome the above limitation. The method builds a weighted heterogeneous graph using drug-drug similarity, target-target similarity, and drug-target binding affinities. It is based on a mix of feature representation learning, graph mining, machine learning techniques. More specifically, drug representations are obtained from Simplified Molecular-Input Line-Entry System (SMILES) embeddings and protein sequences are encoded as ProtVec embeddings. Graph based meta path scoring and embedding based feature concatenation are applied in feature extraction, and the final prediction is made by Extreme Gradient Boosting (XGBoost). The model is evaluated against benchmark datasets Davis, KIBA and the PDBBind Refined, and performs better than existing state-of-the-art methods. It is evaluated using mean squared error (MSE), concordance index (CI) and area under the precision recall curve (AUPR) and found to be effective. Although Affinity2Vec has good predictive power, the authors concede that it does not have the ability to predict interactions of new drugs or targets and that it could be improved off future work with more large scale datasets and other molecular representations.

Yang et al. (2024)[27] proposed MINDG (Multi view Integrated Learning Network) which is a Multi view Integrated Learning Network with deep learning and graph learning to improve the drug target interaction (DTI) prediction. Traditional approaches suffer from the use of first order neighboring nodes in graph based models and lack of combination of both sequence and structural features of drugs and targets. To solve these problems, we propose a three steps framework, first, we use a hybrid deep network to extract the sequence based feature of the drugs and the targets; second, we use a high order graph attention convolutional network (HOAGCN) to discover deeper structural relationships. And, lastly, we use a multi view adaptive integrated decision module (MAIDM) to refine the prediction by combining the outputs from the first two steps. The evaluation of the model was done on the BindingDB dataset and DAVIS dataset for which it showed better performance than the state of art methods DeepCDA, TripletMultiDTI, graph attention networks (GAT) in terms AUPRC, AUROC, F1 Score. Experimental results demonstrate that MINDG is effective in combining deep learning and graph learning to well capture both intrinsic properties of molecules and relational interaction from outside. However, the authors concede that MINDG does not exploit all available intrinsic structural information of drugs and proteins and propose to further improve it in future work using more advanced graph learning techniques. Furthermore, the absence of wet-lab validation is also a limitation and future studies will involve the

incorporation of experimental verification of the model to increase its reliability.

Ren et al.(2023) [21] developed DeepMPF which is a deep learning framework for DTIs prediction by using a multi-modal representation with meta-path semantic analysis. However, traditional DTI prediction methods are difficult to work with as it requires tremendous biological experiments and single modal computational models are limited to capture complex association behaviors in heterogeneous biological networks. In order to address these issues, DeepMPF integrates protein–drug–disease heterogeneous networks, and extracts feature information from three modalities: sequence, heterogeneous structure, and similarity. Six high order semantic representative meta-path schemas are designed to preserve hidden structural information by describing high order relationships. For the drug sequences and 3-mer sparse matrices used for the proteins, sequence embeddings are gained via natural language processing (NLP) tools; for the structural and the similarity features, Smith-Waterman scores and SIMCOMP similarity are used for this purpose. In order to generate comprehensive feature descriptors, the framework applies joint learning to improve predictive accuracy. DeepMPF is evaluated against state-of-the-art methods on four benchmark datasets, and it achieves better results than those methods in drug repositioning experiments on both COVID-19 and HIV. Moreover, the predictions of it are validated by molecular docking experiments, whose reliability is hereby reinforced. The authors admit that some biological features are not captured and recommend generalization through the addition of self attention mechanisms and larger datasets. Future work will focus on scaling up DeepMPF's effectiveness particularly in drug discovery applications.

Wang et al. (2022) [18] proposed the RoFDT model to overcome the challenge in predicting drug–target interactions (DTIs) that are critical in drug discovery and development. Because finding DTIs with traditional wet-lab methods is laborious and costly, computational models are used to predict large amounts of potential DTIs. The authors created RoFDT, an improvement of DTI prediction accuracy with the integration of feature weighted Rotation Forest (FwRF) with protein sequence and drug molecular structure information. The protein sequences were numerically encoded by Position Specific Score Matrix (PSSM) and the features were extracted by using Pseudo Position Specific Score Matrix (PsePSSM). Drug molecular structures were represented by molecular fingerprint descriptors in the meantime. The FwRF classifier was fed these features for prediction. The model was tested on four benchmark datasets, namely Enzyme, GPCR, Ion Channel and Nuclear Receptor, and the accuracy scores were 91.68%, 88.11%, 84.72%, 78.33% respectively. The result of RoFDT was found superior in its performance compared to support vector machines (SVM) and other existing models with 7 out of the top 10 predicted DTIs validated by the SuperTarget database. However, the study did stipulate some limitations of its effectiveness, like the dependence on PSSM for protein sequence representation and the manual feature selection. Eventually, improved automation will be provided; feature extraction techniques will be refined; and the model will be made more sensitive to newly found drug targets.

Mesrabadi et al. (2023)[19] where a computational model for DTI prediction was developed to overcome the high cost, time and complexity of experimental methods.

The three main phases of their approach are feature extraction, feature selection and classification. The authors extracted various protein sequence features like EAAC, PSSM, etc., and also fingerprint features from drugs to form a complete feature representation in the feature extraction phase. However, the extracted data is high dimensional, a wrapper feature selection method called IWSSR was utilized in the next phase to eliminate redundant and less informative features. The Rotation Forest model was used to enhance predictive performance by classifying the selected features. This work innovates by combining different feature extraction techniques with IWSSR based feature selection and finally optimized classification. The model was evaluated on the gold standard datasets (enzyme, ion channel, G protein coupled receptor, and nuclear receptor) with accuracy of 98.12%, 98.07%, 96.82%, and 95.64% respectively. The high predictive performance of these results shows that the model is compatible with other state of the art methods. A key point of the study is based on the effectiveness of computational methods for DTI prediction, without explicitly mentioning limitations or future research directions.

Wu et al. (2021)[12] developed BridgeDPI, a novel graph neural network (GNN) based methodology to improve the prediction of drug–protein interaction (DPI) and to address the issues of existing computational methods such as poor quality of protein-protein associations (PPAs) and drug-drug associations (DDAs). The proposed model utilizes hyper nodes as bridges to connect proteins and drugs to enhance representation learning of PPAs and DDAs in an end-to-end supervised manner. Drug representations are created using molecular fingerprints and sequence-based features, while protein sequences are processed using k-mer features and convolutional neural networks (CNNs). Subsequently, these embeddings are fed into a GNN to combine hyper nodes for learning the interactions of drugs and proteins more accurately. Evaluation of BridgeDPI on various datasets such as customized BindingDB, C.elegans, and human datasets showed better performance than state-of-the-art models with AUC values of 0.989 for seen proteins and 0.952 for unseen proteins in BindingDB, 0.995 in C.elegans dataset, and 0.990 in human dataset. Moreover, BridgeDPI was also validated in a real world case study, by predicting interactions between viral proteins of COVID-19 and antiviral drugs, and the results matched those of the World Health Organization and the FDA, which further validated the practical usefulness of BridgeDPI. It is acknowledged by the authors that benchmark datasets suffer from data bias, resulting in high artificial performance metrics, and they conduct cross validation experiments on independent datasets to avoid this disadvantage. Results indicate that BridgeDPI can improve DPI prediction by using hyper-nodes, and thus may aid in drug discovery.

Öztürk et al. (2018)[2] developed DeepDTA, a deep learning based model for the drug-target binding affinity prediction as the traditional binary classification methods have the limitations of not predicting the strength of the drug-target interactions. Different from existing methods that depend on 3D protein-ligand complex structure or 2D molecular representation, DeepDTA uses only drug and protein sequence information and models them as CNN to predict binding affinity values. The model uses two separate CNN blocks to represent the drug SMILES and protein sequences and extracts high level feature representations before feeding them into the fully connected layers for affinity prediction. As a comparison test, the per-

formance of DeepDTA was evaluated on the Davis and KIBA benchmark datasets and it already surpassed the traditional methods including KronRLS and SimBoost. DeepDTA performed better on the KIBA dataset, with a Concordance Index (CI) of 0.863, higher than that of KronRLS (CI: 0.782) and SimBoost (CI: 0.836) with statistical significance (P-value ¡ 0.0001). Results indicate that the use of CNNs to model sequence information helps in capturing interaction patterns which helps in improving the prediction of binding affinity. The authors recognize that CNN architectures may not be the best way to model the sequential dependencies in protein sequences, and therefore suggest that Long Short-Term Memory (LSTM) networks would improve predictive accuracy. The future work will also include integration of ligand based protein representations and investigating the alternative deep learning architectures to improve more drug target interaction prediction.

An et al., (2021)[10] developed a computational method named WELM SURF that increases drug target interaction (DTI) prediction accuracy by considering protein evolutionary information and molecular structure of the drugs. The computational approaches described in the study aim to address the limitations of experimental DTI validation, which can be very costly and time-consuming. To extract the evolutionary features from protein sequences, WELM-SURF uses Position Specific Scoring Matrix (PSSM), and to further refine the key sequence attributes from PSSM, SURF is used. Molecular substructure fingerprints are utilized as a means of creating feature vectors for drug representation. The Weighted Extreme Learning Machine (WELM) is used to classify DTIs because of its capability of efficiently optimizing the loss function while keeping the fast training speed and strong generalization performance. Evaluation of the model on four benchmark datasets, enzymes, ion channels, G-protein-coupled receptors (GPCRs), nuclear receptors was performed using fivefold cross validation. The accuracy scores for WELM-SURF are 93.54%, 90.58%, 85.43% and 77.45%, which is better than the well known traditional classifiers such as extreme learning machine (ELM) and support vector machine (SVM). It is demonstrated via comparative analysis with existing methods that WELM-SURF has much better feature extraction and prediction ability for DTIs. The authors stress that WELM-SURF is robust and efficient for large scale bioinformatics applications. Additionally, they point towards future work in certain areas, including improvements in feature extraction methodologies and the use of advanced machine learning algorithms which may improve predictive accuracy.

Liu and Paliwal (2024) [24] present DualBind: a deep learning framework for protein–ligand binding affinity prediction where they tackle the problems of prior models that only utilize labeled data or assume a Boltzmann-distributed dataset that might not be viable in reality. However, traditional supervised models utilize mean squared error (MSE) loss and tend to overfit in the case of lack of the labeled data. On the contrary, unsupervised methods such as denoising score matching (DSM) are based on generative modeling and unable to predict the absolute affinity values. DualBind first does a dual-loss integration combining MSE loss for the prediction of accurate absolute affinity and DSM loss to enhance the generalizability by learning from both labeled and unlabeled data, and then integrates two approaches in a dual-loss framework. The model was trained on the PDBbind v2020 refined dataset and evaluated on the CASF-2016 benchmark, achieving a Pearson correlation (Rp)

of 0.757, an RMSE of 1.461, a Spearman correlation of 0.742 and performs better than AutoDock Vina, SimBoost and DSM-only models. Experiments also showed that DualBind is capable of leveraging the unlabeled data, sustaining high predictive performance with only half of the labeled dataset and surpassing purely supervised models. The paper by the authors emphasizes the ease of working with both labeled and unlabeled data on DualBind, and suggests future directions for research to be on hybrid training strategies for high performance models as well as their research on pretraining to improve model performance.

Zhao et al. (2024)[29] proposed a framework called MSI DTI, to integrate multi source information, i.e., sequence based biological features and network based features from bioinformatics knowledge graphs. The model has a drug representation module that uses 5 different types of representation (MACCS keys, Morgan fingerprints, Avalon fingerprints, Mol2Vec, and Graph2Vec) which are then combined to form a unified vector. This contains information like chemical substructure, connectivity patterns in molecular topology and molecular representation. After that the model builds a Drug-Target Knowledge Graph (DTKG) and uses multi head self attention with residual connection to capture both low and high order interaction information. Moreover, they used a pre-trained language model BioBERT to extract entities and relationships. The model also has a protein representation layer which employs two graph based embedding approaches for feature extraction from the heterogeneous biological networks. One of them is deep learning with random walks where they use the AttentionWalk algorithm to obtain node information. The other one is deep learning without random walk where they used combination-based multi-relation graph convolutional networks(CompGCN) which gave further insight into the representation and relationship of the edges in the entity representation when analyzing the node representation. After that the combined features are sent to the feature interaction and output module where they use PCA to decrease the dimension of the feature vector while retaining as much information as possible. By using a multi-head self-attention mechanism the model detects intricate relations between features while separate attention heads redirect their attention to different feature subspaces. The model combines all head outputs together with a mechanism that maintains primary features by using residual connections.

Yan et al. (2021)[13] introduced CNNEMS, a deep learning model designed to predict drug-target interactions (DTIs) by integrating protein evolutionary data and molecular structural information. Identifying DTIs is important in the drug discovery process but biological experiments for this purpose are time consuming, costly and often with high false negative rates. In order to overcome these challenges, CNNEMS employs features extracted by a Convolutional Neural Network (CNN) and a classification using an Extreme Learning Machine (ELM). First, conversion of target protein sequences into Position-Specific Scoring Matrices (PSSM) and visualization using molecular fingerprint for drug compound representation are performed to capture biological evolutionary features. The CNN layers are then applied on the extracted feature vectors to extract abstract hidden features, which are finally classified using ELM algorithm. To validate the model, prediction accuracies of 94.19%, 90.%, 87.95%, and 86.11% was tested on enzymes, ion channels, GPCR, and nuclear receptor datasets, respectively. The higher accuracy and AUC scores

in comparison to those of existing DTI prediction methods confirmed CNNEMS's superior performance. The authors note that although CNNEMS performs well, increased performance of predictive performance can be obtained with improved deep learning algorithms. Future work will refine the model by optimizing the methods of feature extraction and including more types of biological data in order to increase the robustness of the DTI predictions.

Li et al. (2019)[3] presented deepatom as their data-driven system that applies a 3D-CNN structure to predict protein-ligand binding affinity. The DeepAtom framework extracts patterns of interaction by processing voxelized data from protein-ligand structures as opposed to physics-based and empirical and knowledge-based scoring functions which need substantial biological feature engineering. The system decreases bias linked to manually made descriptors through effective prediction models that produce generalized outcomes. The DeepAtom procedure includes three sequential stages starting from voxelization that applies feature channel encoding to a 3D grid box positioned around the ligand atom. The framework employs a light-weight 3D-CNN design that obtains interaction characteristics by utilizing depthwise separable convolutions with an efficient structure to reduce overfitting risks. A global affinity regression block receives extracted features before producing a binding affinity score through its final prediction. The design of DeepAtom achieves both model complexity and generalization because it enables effective performance when trained with limited labeled data. Leaky ReLU activation functions together with batch normalization enhance the computational speed while using training processes to achieve more stable convergence results. DeepAtom was tested and evaluated thoroughly using PDBbind v.2016 along with the Astex Diverse Set. DeepAtom was also trained using PDBbind's refined and general subsets with data augmentation by applying random translations and rotations to develop robustness. According to the evaluation, DeepAtom outperformed the well-known RF-Score, Pafnucy, and X-Score models. DeepAtom demonstrated outstanding performance on PDBbind v.2016 core set with R of 0.83 and RMSE of 1.23 pK units better than other prevailing methods. The model received enhancement through creation of a benchmark dataset that served both as a performance metric and useful resource for future studies in computational drug discovery.

Suruliandi et al. (2024)[25] explore DTI prediction through ML techniques to strengthen drug discovery operations. The research analyzes multiple machine learning (ML) algorithms for DTI expectation while emphasizing the significance of detecting potent drug-protein interactions to develop less expensive new medications. The researchers demonstrate the effectiveness of ML technique predictions for drug and protein target interactions through their analysis of diverse classifiers and datasets and qualitative and quantitative data assessment. The authors established that multiple ML approaches lead to promising DTI prediction results, with chemogenomics-based methods showing the most potential. The research team investigated three major ML strategies that address different functions during DTI prediction, including binding site identification and analysis of drug and target chemical-genomic spaces through docking-based, ligand-based, and chemogenomics-based approaches. Future research in this field should develop new classifiers to improve prediction accuracy according to the study, which reports minimal accu-

racy performance because of limited true negative drug-target pair availability. The paper stresses the necessity to advance ML techniques handling extensive data quantities as well as the integration of continuously updating data sources to optimize prediction procedures. Scientists plan to develop both negative datasets for interaction predictions in addition to applying feature engineering techniques for enhancing dataset quality in order to achieve reliable DTI prediction through advanced ML models.

Zhou et al. (2024)[30] introduce a solution to improve accurate drug-protein interaction (DPI) predictions that serve important functions in drug discovery and personalized medicine. Deep learning models currently face challenges when representing nodes accurately since this issue leads to reduced predictive outcomes. The authors developed a new framework that combines global and local node characteristics from drug-protein bipartite graph systems. The method employs pre-trained extraction of initial features for drugs along with proteins. Local estimations occur through MinHash with HyperLogLog that measure subgraph similarity and set cardinality together with a transformer-based global diffusion mechanism that monitors node interdependencies. The final part merges local components with global elements before applying a multilayer perceptron (MLP) to determine DPI probability. The model shows high accuracy levels through various experimental validations which prove it predicts drug-protein interactions better than conventional techniques. Molecular docking experiments validate that the model identifies fresh DPIs which were absent from existing databases therefore showing potential for drug rejuvenation purposes. Additionally the authors note challenges in managing complex data sources as well as difficulties in making their predictions easier to understand. The future development plan includes integrating new data types alongside enhanced interpretability mechanisms alongside framework optimization for enhanced prediction accuracy.

Bal et al. (2024)[22] used their research to solve the prediction accuracy problem for drug-target interactions (DTIs), which are essential for drug discovery, by developing better computational models for drug-protein binding affinity measurements. Binary classification methods used traditionally do not represent the continuous characteristics of binding strength. The authors present PGraphDTA as their solution to this limitation through their combination of Protein Language Models (PLMs) with Contact Map information. The implementation of ProtBERT, DistilProtBERT, ESM2, and SeqVec serves as Protein Language Models to produce strong representations of amino acid sequences instead of using previous Convolutional Neural Networks (CNNs). The predictive framework includes contact maps as inductive biases which provide protein residue-based and molecular distance-based components to enhance binding affinity predictions. The investigation shows how PLMs enhance prediction accuracy by using DAVIS and KIBA datasets where SeqVec delivers the minimum mean squared error (MSE). The performance benefit from protein contact maps in systems outweighs the negative effects of molecular docking-based contact maps that introduce errors into results. The research shows how PLMs boost protein data representation, especially for short datasets, and contact maps generate valuable structural information. The main challenges of this work include the high processing cost of embedding generation using PLMs along with uncertainties coming from molecular docking predictions. The ongoing work

aims to merge physical element-based structural biases together with attention systems to enhance interaction design while conducting tests through larger available data collections. The study has developed an essential framework to improve DTI prediction by creating better computation methods, which will boost drug discovery speed.

Kawichai et al. (2021)[11] provide an ensemble learning approach as a means of predicting drug-disease relationships. Drug repositioning issues may be resolved using Meta-path-based Gene Ontology Profiles for Drug-Disease Associations (MGP-DDA). Drug development through traditional techniques takes an extended period with high costs, which leads research to explore computational approaches that find additional uses for existing medications. The authors developed a meta-path methodology that incorporates Gene Ontology (GO) terms as connecting nodes between drug and disease entities to generate new information features named meta-path-based GO profiles. The profiles apply drug-disease functional similarities to provide accurate predictions about drug-disease associations. The research design builds a three-part network that relies on information from DrugBank together with DisGeNET and the Gene Ontology Annotation (GOA) database. The meta-paths drug-GO-disease and drug-GO-drug-disease and drug-disease-GO-disease lead to the creation of GO profiles that undergo singular value decomposition (SVD) for dimensionality lowering. The combination of PU bagging with XGBoost as its base learner trains the features to produce predictions. The experimental study demonstrates that MGP-DDA produces superior performance to current methods by reaching 88.6% precision. The practical benefits of MGP-DDA become evident through the fact that 37.7% of predicted drug-disease associations receive validation in ClinicalTrials.gov. The research identifies processing expenses as well as the requirement for incorporating more medical attributes into the system as current limitations. The upcoming research efforts will work on making the results more understandable and include pathway data to enhance predictive capabilities.

Zhang et al. (2024)[28] tackle drug-protein interaction (DPI) prediction challenges because they are essential for drug discovery despite the prevalent bias in graph neural networks (GNNs) and negative transfer across bipartite networks. The authors present Meta-Graph Association-Aware Contrastive Learning (MGACL) to combine meta-knowledge transfer methods with contrastive learning for improving DPI prediction accuracy. A heterogeneous graph neural network within MGACL evaluates complex drug–target relationships through the combination with meta-local connected knowledge transfer networks for improved drug- and target-specific representation enhancement. The model uses an association-aware contrastive learning method that connects high-frequency drug vectors with learned auxiliary graph vectors to stop the negative transfer from occurring. The predictions of MGACL outperform existing approaches by approximately 3% according to AUC and AUPRC metrics through evaluations conducted on five different datasets. The model's effectiveness proves itself even more through confirmations from molecular docking testing of predicted drug–target connections. The research notes two weaknesses, which include difficulty processing uncommon drugs along with expensive training costs for large heterogeneous networks and the existence of improved yet observable negative transfer. Researchers will conduct future work to improve powerful

representation interpretability methods while also optimizing contrastive learning techniques and integration of features to enhance DPI prediction accuracy results.

Kyro et al. (2023)[20] explore the prediction of protein-ligand binding affinity because this ability drives drug discovery and protein engineering research. The authors propose HAC-Net as a Hybrid Attention-Based Convolutional Neural Network which combines 3D convolutional neural network functionality with two graph convolutional networks using attention techniques to boost prediction accuracy. Heatmap-based 3D-CNN deals with atomic-scale spatial data and incorporates GCNs to treat molecular connections as graph structures and applies node selection methods for feature aggregation. HAC-Net attains the top performance rating for PDBbind v.2016 core set benchmarks because of its integrated approach. The model's generalization ability is tested using three sets of training-test separations that optimize differences between protein structures and sequences together with ligand fingerprint variations. HAC-Net's practicality is confirmed when tested on lower-quality data in addition to having its robustness validated through ten-fold cross-validation procedures for ligand variation. HAC-Net achieves superior outcomes compared to prior models by delivering better results in root-mean-square error (RMSE) Spearman rank correlation and Pearson correlation measurement. The use of attention-based components leads to better prediction results because they help identify new interactions between proteins and ligands. The study authors identify two main disadvantages in their approach, such as deep learning network training expenses and potential performance biases in benchmarked data. The research field intends to maintain representation interpretability while optimizing attention elements for use across additional biomolecular property prediction responsibilities. HAC-Net exists on both GitHub and PyPI repositories for scientific community members to develop and validate its use through standardized frameworks.

Gao et al. (2024) [32] focused on the difficulties related to the prediction of Adverse Drug Reactions (ADRs), which is among the most important problems, with a goal in mind for an improvement of both patient's safety and healthcare resources consumption. Classical ADR prediction models tend to not easily handle complex properties involving individual patients due to parameter-dependent model foundations particularly characterised by demographic differences (e.g., age and sex), which also significantly affect the existence of ADRs in a given setting. To close this gap, the authors presented PreciseADR, a system with the goal of personalized ADR prediction at individual patient level. The model includes information from three sources which are categorised into demographic and clinical information for a new patient representation. By running GNNs, particularly the HGNN model, they model correlation relationships among patients, drugs, diseases and ADRs as nodes in a heterogeneous graph and link them with edges. Experiments reveal significant improvement in the prediction performance achieved by PreciseADR over the state of art machine learning methods, which achieves 3.2% AUC and 4.9% Hit@10 improvements respectively. These results corroborate the model's capability of learning intricate dependencies in patient data. In addition, we integrate demographic characteristics such as age and gender, showing they have a notable impact on ADR prediction and the potential of our method in advancing precision medication. Nonetheless, despite being able to incorporate multiple clinical infor-

mation, the model did not include omics data that could limit its ability to capture biological heterogeneity. Therefore, the authors suggest including omics-level data in future studies as well as advanced calculation data augmentation methods to increase predictive stability of ADR models.

Li et al. (2025) [23] aimed at mitigating the prediction of Adverse Drug Reactions (ADRs), one of the most persistent issues in global healthcare, owing to the tremendous effects they have on both patient morbidity and mortality. Conventional ADR prediction models usually encounter a problem of sparse and high-dimensional feature data, which not only deteriorates the prediction accuracy but also leads to training an individual model for each ADR. In order to tackle these shortcomings, the authors proposed a novel model that can address both KG embeddings and deep learning for accurate ADR prediction. In such a way, drug-related objects (i.e., enzymes, drugs, and ADRs) can be transformed into low-dimensional continuous vector spaces through employing the KGs for addressing the data sparsity problem while representing complicated biological essence. Moreover, their deep learning method and integrative features effectively learn the massive amount of data to improve the prediction significantly. Mixing the technology of KG embeddings and CNNs, our model provides unified prediction over multiple ADRs without constructing separate classifiers, which can be expensive in computational costs and challenging to scale. Experimental results suggest that our model is able to learn from the graph structure and outperform traditional machine learning algorithms, which are not powerful enough to catch the complicated relationship among drugs, enzymes, and side effects. The predictive robustness in the present study is also strengthened by the integration of a variety of features, including drug transporters and therapeutic indications. Yet, the paper indicates that there is a lack of standard benchmarking datasets, and the design of optimized embedding strategies are still an open question. In spite of these limitations, Li et al. 's method is a good step forward in developing more generalizable and efficient ADR prediction models.

Zhang et al. (2021) [15] addressed a major bottleneck, ADR prediction, which still represents one of the most crucial issues in drug research at present since today's clinical trials and animal testing fail to detect any potential ADRs before drugs can be put on the market. Given that these traditional approaches cannot fully process the complicated drug-drug interaction and handle a great diversity of compounds, they proposed a novel computational framework by integrating data mining and machine learning methods for more accurate ADR prediction. They were the first to suggest an approach based on KG embedding, with their method representing complex relationships between drugs, side effects, targets and indications as distributional embeddings. It allows the model to learn not only direct ADR hypotheses but also indirect connections between biomedical entities, facilitating a comprehensive ADR identification. To predict the ADRs better, we also integrated KG embeddings with logistic regression to form a general prediction model without requiring different models built for different ADRs. The approaches were tested on benchmark ADR datasets and showed a remarkable performance with high AUC up to 0.87, compared with other convolutional neural networks (CNNs) or deep heterogeneous joint models that tend to be fragile to model's interpretation or computationally expensive. The outcomes demonstrate that the KG embeddings can

help to provide more accurate and interpretable ADR predictions. However, Zhang et al. (2021) also noted that the sparsity of data when applied to achieve improved predictive performances, and scalability for large-scale biomedical research, remains a challenge. The support of new research to develop better semantic drug-disease representation and the implementation of an ensemble of DNN-based framework 32 are needed to improve both the accuracy and generalization of ADR prediction.

Park et al. (2022) [16] against the long-standing challenge of ADR prediction (an issue continuing to hugely impede drug development due to their complex off-target interactions). In vitro experiments and the molecular docking simulations are conventional means for studying drug-target interactions; however, they can only perturbate known proteins, which decreases their predictive scope and efficacy. To overcome these shortcomings, in this study the authors proposed a computational model named as LAPINE (Large-scale ADR-related Proteins Identification with Network Embedding), which was created based on a network approach to enhance the identification accuracy of ADR-related proteins. LAPINE In this paper, combines the data of single-target compounds (STCs) through network embedding algorithms to capture more intricate relationships between drugs with targets/proteins in a higher systemic level. Importantly, it encompasses not only DPIs but also protein-protein interactions (PPIs) and could provide a more comprehensive description in putative ADR-related pathways. By doing so, such a network embedding can encode the biological relatedness in multi-dimensions, in which the representation can reflect not only local but also global topological features. Performance assessment also showed that LAPINE could achieve higher values of PPV and LR than traditional models built on drug intensities-target interactions alone across several large benchmarks of known ADR-related proteins. Their results supported that the model can have better ability in providing predictions of more novel ADR-related proteins, which many of them were then verified by literature evidence. By contrast to the previous models based on drug-only methods, LAPINE showed a better tradeoff between precision and coverage, which made it stand out for practical use in pharmacovigilance studies. Despite the good performance of this study, it also has some limitations. The performance of the model and subsequently its demonstrated validity to predict certain processes is somewhat contingent on the quality and comprehensiveness of interaction networks, which may introduce biases; additionally, other molecular or omics data could increase consideration. The authors propose the application of multi-modal data fusion and deep graph learning approaches to enhance ADR prediction accuracy as well as interpretability.

2.3 Summary of Key Findings

2.3.1 Transition from handcrafted to learned representations

The literature clearly evidences that traditional machine learning algorithms based on manually engineered molecular descriptors are being replaced by deep learning methods, which can learn data-driven features. State-of-the-art CNN- and transformer-based models such as ChemBERTa, DeepDTA[2], and GraphDTA[22] overcome prior methods by learning to automatically mine structural/chemical pat-

terns directly from raw inputs in order to generalize more effectively across wide ranges of chemical/molecular classes.

2.3.2 Emergence of structural and multi-modal modeling

New studies have pointed out the increasing roles of 3D structure for drugs and proteins as well. For GVP-GNN and EGNN, which are based on graph, geometry and topology cues from the real molecular system can probably be expressed more faithfully. This transformation is the transition from structure-agnostic to structure-aware learning in which molecular geometry plays a central role for optimal prediction.

2.3.3 Integration of Pretrained Foundation Models

The use of large-scale pretrained models, like ChemBERTa in the domain of chemicals and ESM in the case of proteins, has made itself a common practice for representation learning. Their contextualized embeddings encode representations that contain rich biochemical information, and fine-tuning them on DTI-related problems leads to better performance and interpretability. This trend highlights a general adoption of foundation models in bioinformatics, where pretrained networks become modularized building blocks for specialized downstream tasks.

2.3.4 Emergence of Multi-Task Learning in Pharmacological Modeling

Recent research increasingly recognizes that drug behavior cannot be adequately explained using a single predictive objective. Pharmacological outcomes such as therapeutic efficacy and adverse drug reactions are inherently interconnected, as they arise from shared molecular interactions and biological pathways. Multi-task learning frameworks have therefore gained attention for jointly modeling related endpoints—such as Drug–Target Interaction (DTI) prediction and Adverse Drug Reaction (ADR) prediction—within a unified architecture. By learning shared representations across tasks, these models enable knowledge transfer between molecular-level interaction patterns and clinical-level outcomes, often resulting in improved generalization and robustness compared to single-task approaches. This joint learning paradigm provides a more holistic view of drug behavior and has become an important direction for building safer and more reliable computational drug discovery systems.

2.3.5 Need for Unified DTI–ADR Frameworks

Despite these achievements, most previous studies treat DTI and ADR prediction independently. This disconnection ignores their biological crosstalk, among which off-target binding is a frequent cause of side reactions. The set of reviewed papers taken as a whole seem to support the need for an integrated multi-task framework that is capable of modeling both therapeutic and side effects at the same time. Such integration may also establish a missing link between molecular-level prediction

and patient-level outcomes, thus improving real-world drug safety and discovery pipelines.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This section presents the functional, technical and methodological needs of the proposed multi-task deep learning framework for predicting Drug-Target Interactions (DTIs) and Adverse Drug Reactions (ADRs). The system is developed as a research-oriented computational system that is intended to support pharmaceutical and biomedical researchers in hypothesis generation and exploratory analysis and not for clinical decision-making.

3.1.1 Functional Requirements

The main functionality requirement of the system is to jointly model molecular interactions and adverse effects in a common learning framework. Given a drug-protein pair, the system has to predict the probability of interaction as a binary choice, while estimating a set of potential adverse drug reactions of the drug in the context of biological interaction. This joint formulation provides the model with the ability to capture shared latent factors between efficacy related interactions and safety related outcomes.

The system must be able to support multi-label ADR prediction, where one drug can be linked to multiple adverse effects, and must be able to scale to a large ADR label space. The model is necessary so that it can work under realistic biomedical constraints such as class imbalance, incomplete interaction coverage and sparsely observed targets. The framework is designed to work entirely as an offline research pipeline to deliver probabilistic outputs to help researchers prioritize interactions or adverse effects for further investigation. No deployment, real-time inference or user-facing clinical interface is assumed.

3.1.2 Technical and Methodological Requirements

From a technical point of view, the hardware needed for the system is a GPU accelerated deep learning environment that can accommodate high dimensional embed-

dings and large multi-label output spaces. The implementation is built on Python and PyTorch is the main deep learning framework used for the implementation, along with other supporting libraries like NumPy and Pandas for data processing and scikit-learn for evaluation metrics and splitting the dataset.

The framework is based on secondary biomedical data datasets to learn both DTI and ADR. Drug-target interaction data are labeled pairs of drugs and proteins and ADR data are associations of drug entities to standardized terms for adverse events. These datasets come together via common drug identifiers and provide the ability to multi-task and supervise the datasets without the requirement of patient-level clinical data. The system assumes that all the datasets have already been preprocessed, cleaned and mapped to common identifier spaces before training takes place.

There are several constraints that affect system performance and design decisions. Model effectiveness is reliant on the availability and quality of external embeddings and selected biomedical data sets, which might contain reporting biases or incomplete coverage. Class imbalance either in terms of interaction labels or in terms of adverse effect frequencies is an inherent limitation, so good loss design and evaluation is required. Additionally, the framework is optimized for methodological soundness and reproducibility and is thus not optimized for deployment efficiency, and instead optimized for research analysis rather than production use.

3.2 Societal Impact

The proposed multi-task deep learning framework has the potential to positively impact the society by supporting more efficient and informed pharmaceutical research. By modeling drug-target interactions as well as adverse drug reactions, the system can help researchers identify potentially unsafe drug candidates earlier in the drug development pipeline, leading to fewer pitfalls in the late stages of the development pipeline and better drug safety. From a public health standpoint, such computational tools can help with the faster discovery of safer therapeutics especially for complex or rare diseases where there is a limited amount of available experimental data. Better prioritization of interactions and adverse effects for further investigation, by improving evidence-based decision making in drug discovery, may lead to decreasing healthcare costs and better patient outcomes.

At the same time, the societal application of predictive models in biomedical research are important issues of concern. There is a risk of misinterpreting probabilistic predictions as definitive medical views (and so, an over-reliance on the model outputs, or inappropriate use of models in contexts for which they were not originally meant to provide support). If such systems are abused, they could be used to indirectly affect unsafe self-medication practices, or unsupported clinical assumptions. Legal and safety considerations therefore dictate explicit communication of the limitations of the system and to restrict the use of the system to non-clinical research situations. Additionally, cultural and societal differences in drug usage, reporting of adverse effects and access to healthcare may not be fully represented in the underlying datasets and thus may limit generalizability across populations. Addressing

these concerns means that models must be framed responsibly, assumptions must be transparent, and results must be interpreted within their appropriate scientific framework.

3.3 Environmental Impact

The proposed framework has an indirect, but meaningful environmental relevance in its contribution to efficiency in pharmaceutical research and drug development. By facilitating early computational screening of drug-target interactions and adverse drug reactions, the system can help to reduce the number of failed experimental studies and unnecessary laboratory trials. This decrease in wet lab experimentation is part of the lower consumption of chemical reagents, biological materials and physical resources, this has a measurable environmental footprint in traditional drug discovery pipelines.

In addition, the framework encourages sustainable research practices and the use of reusable digital datasets and shared learned representations instead of repeated physical experimentation. Once the model is trained, it can be used for several research hypotheses without further material costs, which contributes to the scalability of the model with minimal incremental resource usage. While the system exists in a computational research context, the main contribution to the environment is the support of data-driven decision making, which will lead to less waste, increased efficiency and more sustainable drug development workflows.

3.4 Ethical Issues

The major ethical issue related to this project is the misuse or misinterpretation of model predictions. As the system generates probabilistic estimates of drug-target interactions and adverse drug reactions, there is the risk that the outputs would be over-trusted or misused as clinical evidence.This misuse may lead to conclusions or decisions to engage in self-medication without proper support in the event that the context of the research and the restrictions of the model is not explained properly. That the framework is explicitly placed in the context of research support and not clinical decision is therefore a fundamental professional responsibility.

The main solution to these risks is transparency and responsible practices in AI. The design is aimed at ensuring that the data sources are documented, modeling assumptions and evaluation metrics are recorded so that the results can be interpreted intelligently. No patient-level data is used, which removes privacy and consent issues, but there is potential that biases in curated biomedical data will impact on predictions. Ethical use of the system requires that the researchers be aware of these limitations and avoid overstating conclusions and use the model as a complementary analytical tool and not as a substitute for experimental validation or expert judgement.

3.5 Project Management Plan

The project was being run in accordance to a well organized research based development plan which would be on track with the academic schedule and available resources. The entire plan was split into consecutive steps, which include literature review and problem formulation, data set selection, preprocessing, and embedding integration. The next stages were devoted to the design of models, the formation of multi-task learning, training and validation, comparison with the basis, and analysis of results. The last phase was documentation, refining of the evaluation and writing of thesis. Such a phased direction guaranteed the stable movement and gave an opportunity to improve it repeatedly on the basis of experimental results.

In terms of resource management, the project was mainly based on the effort of the researcher with the help of high-performance GPU resources. The model development and testing were performed with locally available NVIDIA GPUs, such as RTX 5060 Ti, 5070 Ti and 5090, which do not require the expenditure of external cloud computing. The project utilized software tools and libraries that were open-source and thus minimized budgetary requirements. The investments were mainly time, computer resources, and technical skills, and not direct money as the project was performed as an academic research; hence, it can be conducted with great efficiency in a student research setting.

3.6 Risk Management

A number of technical and methodological risks can be found in the development of the proposed multi-task framework. A significant threat was connected to data-related constraints, such as imbalance between classes in drug-target interactions, absence of adverse drug reaction label, and possible inconsistencies between merged data sets. Such risks were handled by preprocessing data carefully, splitting it into controlled sets, and the application of the right loss functions and evaluation metrics that are aimed at the imbalanced learning and multi-label cases.

The other major threat was associated with model complexity and stability in training because of the large dimensionality of embeddings and the concomitant optimization of a set of objectives.In order to deal with this, the pilot experiments were done on smaller datasets, and only after the full-scale training, instability or overfitting could be detected early. Computational limitations like memory constraint and long training sessions were handled through managing batch size accordingly and modular experimentations . Overall, a progressive development approach and methodological validation served as means of minimizing the technical and project implementation risks.

3.7 Economic Analysis

Economically, the suggested framework is a cost-effective way of funding pharmaceutical research and drug discovery at early stages. The project uses mostly the publicly available datasets and open-source software frameworks, which help to save much on the data acquisition and licensing costs. The development and testing of the models were performed based on the locally available GPU resources without the need to incur repetitive costs in using cloud-based computing services. Consequently, the immediate financial expense of the venture is not very high and is mostly confined to hardware and time of researchers.

Regarding the possible benefits, the framework provides economic advantage in the form of early detection of improbable drug-target interactions and possible unsafe adverse effects profile. The system can assist in curbing the occurrence of failed laboratory experiments and late drug development misses which are significant contributors of loss of monies in the pharmaceutical pipelines by backing up the computational pre-screening. Even though the framework is not designed to be a commercial or a clinical product, its application as a research-support tool can further aid in making efficient use of experimental resources, which can eventually provide positive cost/benefit attributes in an academic and industrial research setting.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The proposed framework aims to create a unified deep learning architecture that is capable of jointly predicting Drug-Target Interactions (DTIs) and Adverse Drug Reactions (ADRs) in a single model. The central objective of this methodology is to learn a shared representation that captures both molecular interaction patterns and drug-specific profiles of adverse effects of drugs by integrating heterogeneous biochemical and clinical data modalities.

The overall design takes multimodal representations of drugs and proteins, combining both sequence based and structure aware encodings. Drug molecules are represented in using contextual embeddings based on chemical sequences and geometry-aware molecular graphs, whereas protein targets are represented by means of representations that capture evolutionary sequence information along with three dimensional structural characteristics. Based on empirical evaluation of multiple encoding configurations, the following fusion based representations were selected for the final model:

- ESM + GVP for proteins, and
- ChemBERTa + EGNN for drugs.

Such fused embeddings allow the model to learn complementary semantic and geometric properties that are essential to correct biochemical interaction modeling.

After fusion the combined drug-protein representation is passed through a Variational Latent bottleneck using autoencoders (VAE). The combination feature vector undergoes an encoder network which takes as input a probabilistic latent distribution and predicts its mean and log-variance. A latent variable is then sampled from this distribution according to the reparameterization trick, which allows stochastic representation learning whilst preserving differentiability. This latent space regularization with a VAE helps to regularize the shared representation, promoting compactness and smoothness, which enhances generalization between unseen drug-protein pairs.

The latent embedding sampled is used as common backbone for multi-task prediction. Two task specific heads are connected to this latent representation: one head predicts the probability of drug-protein interaction as a binary classification task while the second head performs multi-label classification of adverse drug reactions using one-hot encoded ADR vectors. Each ADR label is treated independently, and the model can estimate the likelihood of multiple adverse effects of a given drug.

The model is trained end-to-end using a composite objective combination of the DTI classification loss, the ADR multi-label classification loss and a Kullback-Leibler (KL) divergence term based on the VAE. The KL term has the effect of a regularizer on the latent space, while task specific losses move the model towards correct interaction prediction and bad effect profiling. To balance the learning dynamics between tasks, uncertainty based loss weighting is used, so that the model can adaptively change the weight of each task during training.

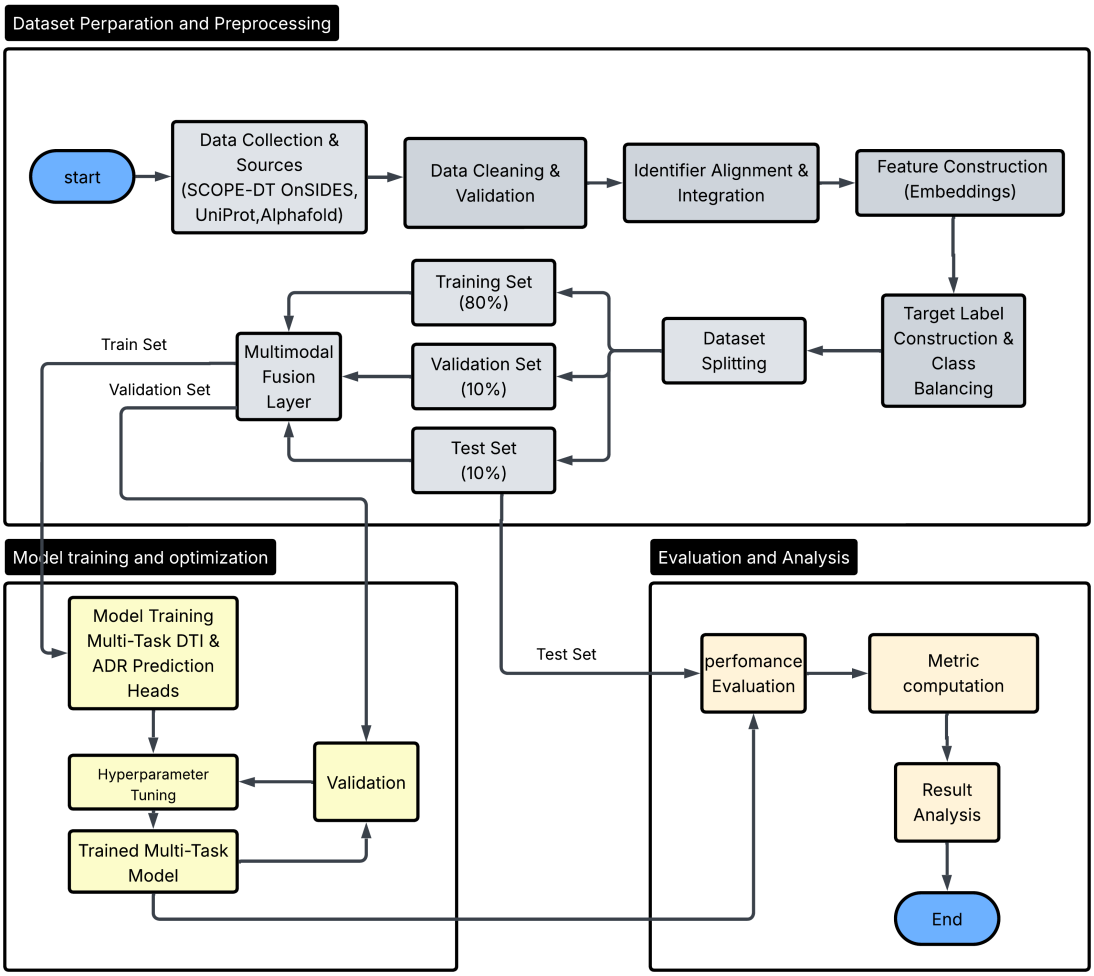


Figure 4.1: Top level overview of the proposed system

In summary, the design process combines multimodal representations of drugs and proteins along with a VAE based latent learning mechanism and multi-task supervision. This unified framework enables the model to simultaneously reason about

the molecular binding behavior and adverse drug reactions, offering a coherent and scalable method for predictive pharmacology.

4.2 Model Specification

This section provides the full specification of the proposed multi-task architecture used to predict Drug-Target Interaction (DTI) and Adverse Drug Reaction (ADR) results simultaneously. The model combines multimodal drug and protein representations through feature-level fusion and learns a regularized shared latent representation using a Variational Autoencoder (VAE). From this shared latent embedding, the model performs (i) binary classification for DTI prediction and (ii) multi-label classification for ADR prediction.

The major stages of the overall architecture are as follows:

- **Drug feature fusion** (ChemBERTa + EGNN)
- **Protein feature fusion** (ESM + GVP)
- **Joint context encoding** of the fused drug–protein pair
- **VAE latent bottleneck** (μ, $\log \sigma^2$, reparameterization)
- **Task-specific output heads** (DTI head and ADR decoder)

This section presents the formal specification of each module and the complete training objective.

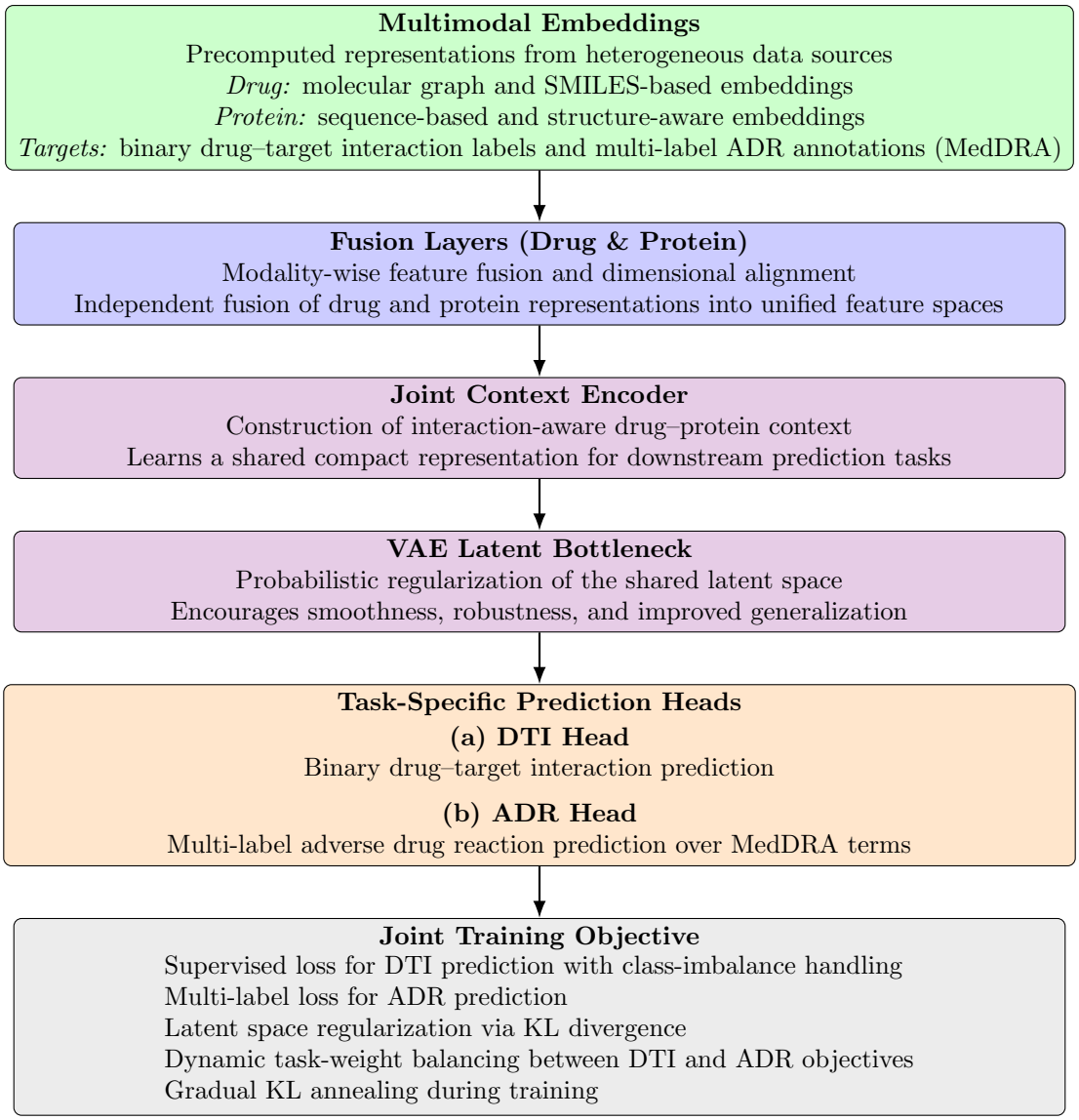


Figure 4.2: Generalized architecture of the proposed multi-task DTI–ADR framework.

4.2.1 Input Feature Construction

For each drug–protein pair, four embedding vectors are provided as input:

- **Drug embedding 1**: $\mathbf{d}_1$ (EGNN-based molecular graph embedding)
- **Drug embedding 2**: $\mathbf{d}_2$ (ChemBERTa SMILES embedding)
- **Protein embedding 1**: $\mathbf{p}_1$ (ESM sequence embedding)
- **Protein embedding 2**: $\mathbf{p}_2$ (GVP structural embedding)

Thus, a single sample is defined as:

$$(\mathbf{d}_1, \mathbf{d}_2, \mathbf{p}_1, \mathbf{p}_2, y_{\text{dti}}, \mathbf{y}_{\text{adr}})$$

where:

- $y_{\text{dti}} \in \{0, 1\}$ denotes the binary interaction label
- $\mathbf{y}_{\text{adr}} \in \{0, 1\}^K$ denotes the ADR target vector
- K is the ADR vocabulary size (in this implementation, $K = 4817$)

The target ADR is represented through multi-hot encoding, where each category of ADR corresponds to one dimension of the vector. A value of 1 denotes that the ADR is present for the corresponding drug.

4.2.2 ADR Multi-Hot Encoding Mechanism

To represent ADR in a machine-readable format, a fixed ADR vocabulary is built using unique MedDRA identifiers. Each MedDRA ID is represented with a unique index, which creates a binary vector space of dimension K.

The list of ADRs associated with a particular drug is encoded as:

$$\mathbf{y}_{\text{adr}}[i] = \begin{cases} 1, & \text{if ADR}_i \text{ is present for the drug} \\ 0, & \text{otherwise} \end{cases}$$

This encoding is appropriate for the multi-label nature of ADR prediction, where a single drug may be associated with multiple adverse reactions simultaneously.

4.2.3 Fusion Layer Design

Since both drugs and proteins are represented using two different embeddings, a fusion operation is needed to combine them in one unified representation for each modality. It is implemented using a fully connected transformation applied to the concatenation of two embeddings. For embeddings $\mathbf{e}_1 \in \mathbb{R}^m$ and $\mathbf{e}_2 \in \mathbb{R}^n$, fusion is defined as:

$$\mathbf{h} = \phi\left(\text{LN}\left(W[\mathbf{e}_1; \mathbf{e}_2] + \mathbf{b}\right)\right) \tag{4.1}$$

where:

- $[\mathbf{e}_1; \mathbf{e}_2]$ denotes concatenation
- $W \in \mathbb{R}^{d \times (m+n)}$ and $\mathbf{b} \in \mathbb{R}^d$ are learnable parameters
- LN denotes layer normalization
- ϕ denotes the ReLU activation function
- Dropout is applied after the activation with probability 0.1

This fusion module is applied separately for drugs and proteins:

- **Drug fusion**: combines $\mathbf{d}_1$ and $\mathbf{d}_2$
- **Protein fusion**: combines $\mathbf{p}_1$ and $\mathbf{p}_2$

The resulting fused representations are:

$$\mathbf{d}_f = f_d(\mathbf{d}_1, \mathbf{d}_2), \qquad \mathbf{p}_f = f_p(\mathbf{p}_1, \mathbf{p}_2) \tag{4.2}$$

In the implementation, the fused embedding dimension d is fixed at 1024.

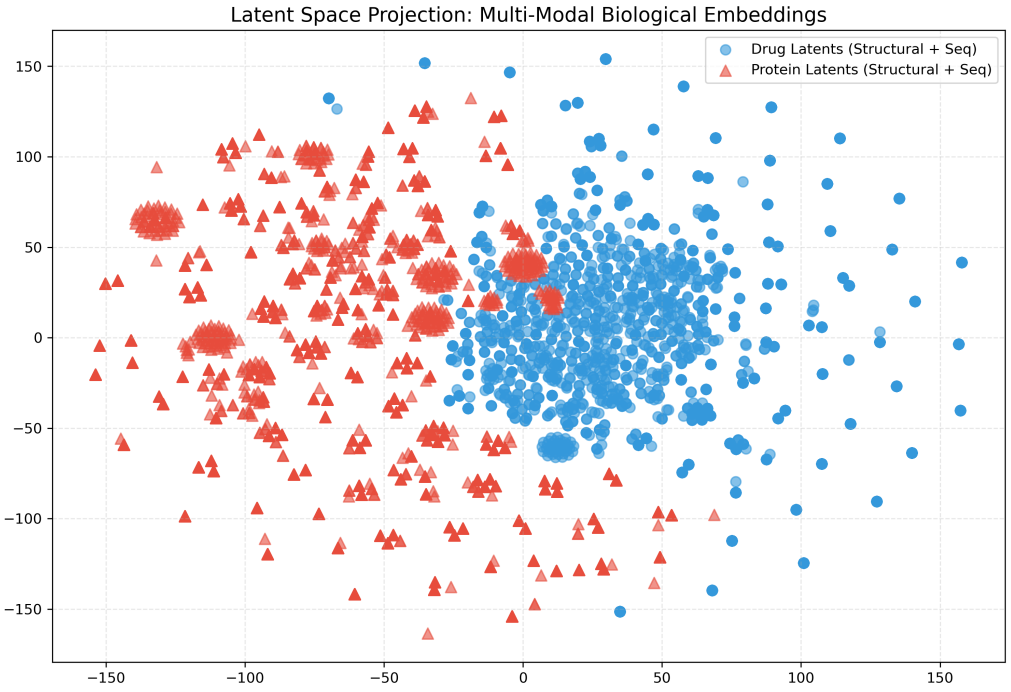


Figure 4.3: Fused Protein and Drug clustering after learning

4.2.4 Joint Context Encoder

The fused drug and protein representations are concatenated to form a paired representation:

$$\mathbf{x}_{dp} = [\mathbf{d}_f; \mathbf{p}_f] \tag{4.3}$$

Since each fused vector is 1024-dimensional, the joint representation lies in:

$$\mathbf{x}_{dp} \in \mathbb{R}^{2048} \tag{4.4}$$

This vector is fed through a multi-layer perceptron (MLP)—termed the *context encoder*—which learns higher-level features of interaction across modalities. The context encoder is composed of:

- Linear($2048 \to 1536$) + ReLU + Dropout(0.2)
- Linear($1536 \to 1024$) + ReLU
- Linear($1024 \to 768$) + ReLU

The resulting context representation is:

$$\mathbf{c} = g(\mathbf{x}_{dp}), \qquad \mathbf{c} \in \mathbb{R}^{768} \tag{4.5}$$

where $g(\cdot)$ denotes the function defined by the context encoder MLP.

4.2.5 Variational Autoencoder (VAE) Bottleneck

In order to regularize the shared representation and to enable compact latent learning, the context vector $\mathbf{c}$ is passed through a VAE-based latent bottleneck.

Latent Distribution Parameterization

Two linear layers compute the parameters of the latent distribution:

$$\boldsymbol{\mu} = W_\mu \mathbf{c} + b_\mu \tag{4.6}$$

$$\log \boldsymbol{\sigma}^2 = W_\sigma \mathbf{c} + b_\sigma \tag{4.7}$$

where:

- $\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2 \in \mathbb{R}^L$
- Latent dimension $L = 768$

Reparameterization Trick

A latent vector $\mathbf{z}$ is sampled using:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I) \tag{4.8}$$

This formulation preserves differentiability and allows end-to-end training.

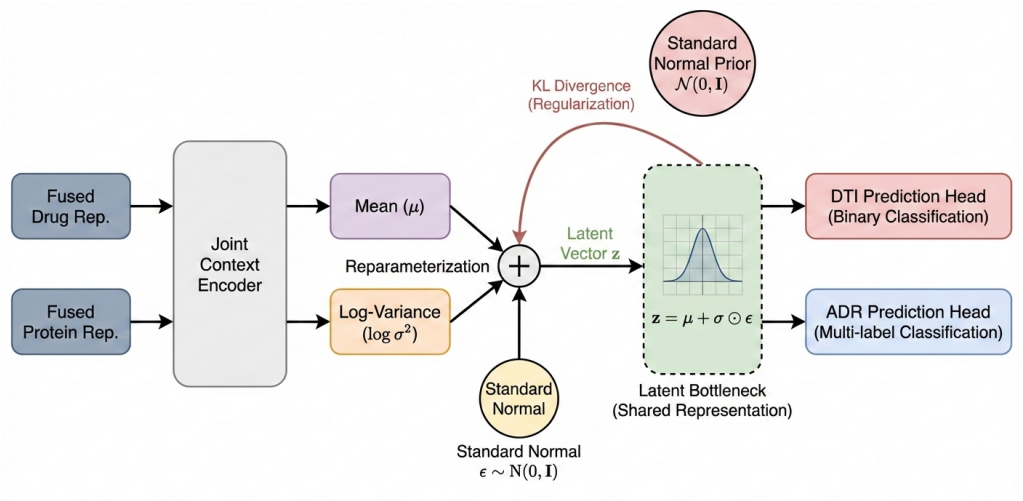


Figure 4.4: VAE-based latent bottleneck for joint DTI and ADR prediction.

4.2.6 Task-Specific Prediction Heads

The latent vector $\mathbf{z}$ serves as a shared backbone for both prediction tasks.

DTI Prediction Head

The DTI head performs binary classification and outputs a single logit:

- Linear($768 \to 384$) + ReLU

- Linear($384 \to 1$)

$$\hat{y}_{\text{dti}} = W_{\text{dti}}\mathbf{z} + b_{\text{dti}} \tag{4.9}$$

During inference, the interaction probability is computed as:

$$P(y_{\text{dti}} = 1) = \sigma(\hat{y}_{\text{dti}}) \tag{4.10}$$

ADR Decoder Head

The ADR head outputs a vector of logits for all ADR categories:

- Linear($768 \to 1536$) + ReLU
- Linear($1536 \to K$)

$$\hat{\mathbf{y}}_{\text{adr}} = W_{\text{adr}}\mathbf{z} + b_{\text{adr}}, \quad \hat{\mathbf{y}}_{\text{adr}} \in \mathbb{R}^K \tag{4.11}$$

Sigmoid activation is applied element-wise:

$$P(\mathbf{y}_{\text{adr}}[i] = 1) = \sigma(\hat{\mathbf{y}}_{\text{adr}}[i]) \tag{4.12}$$

4.2.7 Loss Functions and Optimization Strategy

Training optimizes three main components:

1. **DTI Loss:** Binary cross-entropy with logits:

$$\mathcal{L}_{\text{dti}} = \text{BCEWithLogits}(\hat{y}_{\text{dti}}, y_{\text{dti}}) \tag{4.13}$$

2. **ADR Loss:** Multi-label binary cross-entropy:

$$\mathcal{L}_{\text{adr}} = -\sum_{i=1}^{K} \left[y_{\text{adr}}[i] \log \sigma(\hat{y}_{\text{adr}}[i]) + (1 - y_{\text{adr}}[i]) \log(1 - \sigma(\hat{y}_{\text{adr}}[i])) \right] \tag{4.14}$$

3. **KL Divergence Loss:**

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{j=1}^{L} \left(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2\right) \tag{4.15}$$

The KL weight β is annealed during training:

$$\beta = \min(0.04, 0.001 \times \text{epoch}) \tag{4.16}$$

Uncertainty-based task weighting balances the DTI and ADR losses:

$$\mathcal{L}_{\text{task}} = \exp(-s_1)\mathcal{L}_{\text{dti}} + s_1 + \exp(-s_2)\mathcal{L}_{\text{adr}} + s_2 \tag{4.17}$$

The final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{KL}} \tag{4.18}$$

4.2.8 Training Configuration

- **Batch size:** 64
- **Optimizer:** Adam
- **Learning rate:** 1×10^{-3}
- **Weight decay:** 1×10^{-4}
- **Scheduler:** ReduceLROnPlateau
- **Monitoring metric:** 0.8×Validation DTI AUPRC+0.2×ADR Weighted AUPRC

The training configuration specifies how the model parameters are optimized during learning. Adam optimization is used due to its stability in high-dimensional neural architectures. The ReduceLROnPlateau scheduler automatically adjusts the learning rate when the monitored validation metric stops improving, which helps prevent stagnation and supports stable convergence. Validation DTI AUPRC is chosen as the monitoring signal because DTI prediction datasets are typically class-imbalanced; in such scenarios, AUPRC provides a more reliable performance measure than accuracy.

4.2.9 Output Interpretation

The proposed model produces two outputs for each input drug–protein pair:

1. A DTI prediction (binary interaction probability)
2. An ADR prediction (multi-label adverse reaction vector)

Both outputs are produced based on the shared latent representation $\mathbf{z}$, but they correspond to different clinical and biological interpretations.

DTI Output Meaning

The DTI head produces a single scalar logit $\hat{y}_{dti}$, which shows the model's raw (unnormalized) confidence about whether the given drug interacts with the given protein or not:

$$\hat{y}_{dti} = W_{dti}\mathbf{z} + b_{dti} \tag{4.19}$$

Since logits are not directly interpretable as probabilities, a sigmoid activation has been used during inference:

$$P(y_{dti} = 1) = \sigma(\hat{y}_{dti}) \tag{4.20}$$

This value lies in the range $[0, 1]$ and represents the predicted likelihood of interaction:

- If $P(y_{dti} = 1)$ is close to 1, the model predicts that the drug and protein are highly likely to interact.
- If $P(y_{dti} = 1)$ is close to 0, the model predicts that interaction is unlikely.

In very practical terms, the output of the DTI can be interpreted as a binding/interaction confidence value of score, where higher values indicate stronger predicted association between the drug and the protein target. A classification decision can be generated through the application of a threshold (commonly 0.5), but in most biomedical settings the continuous probability score is more useful since it permits the ranking of drug-target pairs for subsequent validation.

ADR Output Meaning

The ADR head produces a vector of logits $\hat{\mathbf{y}}_{adr} \in \mathbb{R}^K$, where K is the total number of ADR categories:

$$\hat{\mathbf{y}}_{adr} = W_{adr}\mathbf{z} + b_{adr} \tag{4.21}$$

Each element $\hat{\mathbf{y}}_{adr}[i]$ represents one ADR label and is the model's raw confidence of whether that ADR is present or not. Like the output of the DTI, logits are transformed into probabilities by an element-wise sigmoid function:

$$P(\mathbf{y}_{adr}[i] = 1) = \sigma(\hat{\mathbf{y}}_{adr}[i]) \tag{4.22}$$

The resulting vector:

$$\mathbf{p}_{adr} = [P(\mathbf{y}_{adr}[1] = 1),\ P(\mathbf{y}_{adr}[2] = 1),\ ...,\ P(\mathbf{y}_{adr}[K] = 1)] \tag{4.23}$$

represents the predicted probability of each ADR occurring.
This is a multi-label prediction setting, meaning:

- A drug may be associated with multiple ADRs simultaneously
- Each ADR probability is predicted independently
- The model is not forced to choose only one ADR class

Therefore, the ADR output should be interpreted as a risk profile over the ADR space, not a single-class decision.

Converting ADR Probabilities into Predicted ADR Sets

To obtain the final predicted ADRs from the probability vector, one of the following strategies can be used:

1. **Threshold-based selection**

   A probability threshold τ is applied to each ADR label:

   $$\text{Predict ADR}_i = \begin{cases} 1 & \text{if } P(\mathbf{y}_{\text{adr}}[i] = 1) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

   A common choice is $\tau = 0.5$, but in practice τ may be tuned based on validation performance because ADR prediction is typically imbalanced.

2. **Top-k selection**

   Instead of using a fixed threshold, the ADRs with the highest predicted probabilities are selected:

- Predict the top-k ADR labels ranked by $P(\mathbf{y}_{\text{adr}}[i] = 1)$

This approach is useful when the goal is to retrieve the most likely ADR candidates for analysis.

Both approaches make the model output flexible in terms of whether the application needs strict binary decisions or ranked ADR recommendations.

Relationship Between DTI and ADR Outputs

Despite the fact that DTI and ADR outputs are associated with different tasks, they are taught the same latent representation $\mathbf{z}$. Consequently, the latent space captures useful information that helps in both:

- identifying biochemical interaction signals (DTI), and
- capturing drug-related adverse outcomes (ADR).

This common learning facilitates the model to enjoy the benefits of task coupling: interaction patterns affect the latent feature learning process, and ADR supervision promotes the representation to retain clinically relevant drug signatures.

4.2.10 Summary of the Architecture

The overall model produces two clinically meaningful outputs:

1. an interaction probability for a drug–protein pair
2. a multi-label ADR probability profile for the drug

To create a unified drug and protein representation, the fusion module combines two complementary embeddings and the context encoder learns interaction-aware features. The VAE bottleneck regularizes the latent space and aids generalization. Finally, two task-specific heads translate the shared latent representation into predictions for the DTI and ADR outcomes.

4.3 Data Collection

The data collection process aimed to construct an integrated corpus linking biochemical drug–target interaction (DTI) information with clinical adverse drug reaction (ADR) outcomes, enabling a unified modeling framework for both tasks. To achieve this, two primary biomedical resources were used: SCOPE-DTI (Scalable Computational Prediction of Drug–Target Interactions)[31] for interaction-level biochemical data and OnSIDES (Open Network of Side Effects)[26] for clinical ADR annotations. These sources were selected for their scale, reliability, and standardized identifier systems that facilitate cross-modal alignment.

SCOPE-DTI Dataset

The SCOPE-DTI dataset served as the principal source of biochemical interaction information. It aggregates experimental and high-confidence inferred drug–protein interaction data from multiple public databases including BindingDB, ChEMBL, and DrugBank. The benchmark version used in this research contained approximately 2.2 million candidate drug–protein pairs, spanning a diverse set of compounds and protein families.

Each record in SCOPE-DTI is represented by a structured tuple:

$$(\text{drug\_chembl\_id}, \text{target\_uniprot\_id}, \text{label}, \text{smiles}, \text{sequence}, \text{molfile\_3d})$$

where:

- drug_chembl_id — ChEMBL identifier of the small-molecule drug;
- target_uniprot_id — UniProt accession of the protein target;
- label $\in \{0, 1\}$ — binary indicator of binding (1 = interaction, 0 = non-interaction);
- smiles — simplified molecular-input line-entry string encoding chemical structure;
- sequence — amino-acid sequence of the protein;
- molfile_3d — 3D atomic coordinates of the drug molecule in MOL format;

From the original corpus, only entries containing complete identifiers and valid molecular representations were retained. After removing duplicates and incomplete records, a refined subset of 34,741 unique drug–protein pairs was produced. Within this subset, there were 1,028 unique drugs and 2,385 unique protein targets, with interaction labels distributed as 12,234 positive (binding) and 22,507 negative (non-binding) samples.

This processed DTI dataset provided the biochemical backbone for the multi-task framework, offering both structural and sequence information for each molecular entity. The inclusion of both SMILES strings and 3D coordinates allowed generation of multiple embedding modalities, such as ChemBERTa, Smiles2Vec, and EGNN.

OnSIDES Dataset

The OnSIDES dataset supplied the adverse reaction annotations necessary to model clinical outcomes. It is a standardized resource mapping RxNorm drug identifiers (RxCUIs) to reported adverse drug reactions (ADRs) expressed in Medical Dictionary for Regulatory Activities (MedDRA) terminology. The version employed in this work contained roughly 19,000 RxNorm-linked drug entries with over 4,800 unique ADR terms, including both common and rare clinical manifestations.

Each OnSIDES record consists of:

$$(\text{rxcui}, \text{adr\_term}, \text{standard\_concept\_id}, \text{frequency})$$

where adr_term refers to the standardized MedDRA name of the side effect and frequency denotes its reported prevalence across clinical observations. These data were pre-aggregated across multiple adverse event databases, including SIDER, FAERS, and OFFSIDES, ensuring consistent coverage of approved drugs.

In total, the OnSIDES dataset covered approximately 1,000 unique drugs linked to 4,817 ADR categories, represented in a sparse mapping where each drug could be associated with one or multiple adverse events.

Cross-Dataset Alignment

To connect the biochemical and clinical domains, a robust identifier harmonization process was established using the RxNav REST API. This mapping enabled translation between chemical and clinical namespaces through the following correspondence:

$$\text{ChEMBL ID} \rightarrow \text{PubChem CID} \rightarrow \text{RxNorm CUI} \rightarrow \text{MedDRA Term}_{\text{ADR}}.$$

This procedure ensured that each drug in the SCOPE-DTI dataset was paired with its corresponding ADR profile from OnSIDES. Deprecated or ambiguous mappings were resolved using the historystatus attribute provided by RxNorm, and all cross-reference relationships were logged in version-controlled tab-separated files for reproducibility.

Through this multi-stage linkage, a unified dataset was achieved in which each drug could be simultaneously represented by its chemical structure, protein interaction profile, and clinical side-effect vector. This harmonized dataset served as the foundation for all subsequent preprocessing and embedding stages.

Supporting Protein Data

Protein sequence data were retrieved from the UniProt Knowledgebase (UniProtKB) using the target_uniprot_id field. Structural data were obtained from the AlphaFold Protein Structure Database, which provided predicted 3D atomic coordinates for nearly all human proteins in the dataset. These structures were used to build residue-level graphs processed by the GVP-GNN encoder (details in Section 4.3.2).

This ensured that every protein was represented by both its primary sequence and 3D geometric context, thereby enhancing the fidelity of structural representation in downstream modeling.

Summary of Data Sources

Table 4.1: Summary of integrated data sources used in the multi-modal framework.

Dataset	**Domain**	**Key Identifiers**	**Records / Entities**	**Purpose**
SCOPE-DTI	Biochemical	drug_chembl_id, target_uniprot_id, smiles, molfile_3d	34,741 pairs (1,028 drugs, 2,385 proteins)	DTI binding interactions
OnSIDES	Clinical	rxcui, adr_term, meddra_concept	∼19,000 drug entries, 4,048 ADRs	Adverse drug reaction profiles
UniProt + AlphaFold	Structural	target_uniprot_id, PDB	2,385 proteins	Protein sequences and 3D geometries

The integration of these datasets established a comprehensive foundation for multi-modal learning, where biochemical binding data and clinical outcome data were represented within the same computational framework.

4.3.1 Data Cleaning

Rigorous cleaning ensured completeness and consistency across all modalities. For the DTI subset from SCOPE–DTI, entries lacking valid identifiers or structural fields were removed. Duplicate (drug_chembl_id, target_uniprot_id) pairs were deduplicated, and interaction labels were standardized as $y_{\text{dti}} \in \{0, 1\}$. The final distribution comprised 12,234 positive and 22,507 negative samples. All SMILES strings were canonicalized using RDKit.

For the ADR corpus from OnSIDES, all adverse effect terms were converted to lowercase and mapped to MedDRA Preferred Terms. Rare events with fewer than three observations were removed, and unmapped RxNorm CUIs were discarded. Each remaining record was validated against the activeingredient field to avoid duplicate combinations.

Protein structures from the AlphaFold database were verified with Bio.PDB, ensuring each residue contained atoms N, Cα, C. Graph cache files used by the GVP-GNN encoder were checked for correct tensor shapes and decompression integrity; corrupted files were automatically reconstructed.

Cross–modality validation confirmed that every drug–protein pair referenced valid embeddings and RxNorm links. Non–finite numerical values were replaced with zero, and categorical encodings were verified against fixed vocabularies, producing a structurally consistent dataset of 1,028 drugs, 2,385 proteins, and 4,817 ADRs.

4.3.2 Data Transformation

After cleaning, the heterogeneous biochemical, structural, and textual data were transformed into fixed-dimensional numerical representations suitable for machine-learning pipelines. The transformation procedures differed for drugs, proteins, and adverse reaction terms, each requiring domain-specific encoding methods.

Drug Representation

Drug molecules were represented in three complementary modalities: textual, graph-based, and contextual.

- **ChemBERTa Embeddings (384-D):** SMILES strings were tokenized and passed through the pretrained ChemBERTa-2 model. The contextual hidden states were mean-pooled to yield a 384-dimensional embedding capturing substructure semantics and chemical grammar.
- **Smiles2Vec Embeddings (256-D):** The same SMILES were character-tokenized and encoded using a CNN-BiGRU architecture that learned short-range n-gram features. The final hidden representation captured local atomic patterns and valence motifs.
- **EGNN Embeddings (256-D):** Using RDKit, each molecule's 3D coordinates were extracted from the `molfile_3d` field to construct atomic graphs $G = (V, E)$, where each node v_i carried features $(Z_i, \mathbf{r}_i)$ representing atomic number and position. The E(n)-Equivariant Graph Neural Network (EGNN) propagated messages respecting geometric invariances:

$$h_i' = \phi_h \left(h_i, \sum_{j \in N(i)} \phi_e(h_i, h_j, \|\mathbf{r}_i - \mathbf{r}_j\|^2) \right), \tag{4.25}$$

  producing spatially informed molecular embeddings invariant to rotation and translation.

Protein Representation Protein features were extracted from both primary sequences and predicted 3D structures.

- **ESM-2 Embeddings (1280-D):** Each protein sequence from UniProt was tokenized into amino acids and encoded using Meta AI's Evolutionary Scale Model 2 (ESM-2, 650M parameters). The model output per-residue contextual embeddings, which were averaged to obtain a single 1280-dimensional vector per protein:

$$\mathbf{p}_{\text{ESM}} = \frac{1}{L} \sum_{i=1}^{L} h_i, \tag{4.26}$$

  where L is the sequence length.
- **GVP-GNN Embeddings (1024-D):** AlphaFold-predicted PDB files were converted into residue-level graphs. Nodes represented residues, and edges

captured spatial proximity within 10 Å. The Geometric Vector Perceptron (GVP) encoder integrated scalar and vector features:

$$(s', v') = \text{GVP}(s, v) = (W_s[s; \|v\|],\ W_v v/\|v\|)\,, \tag{4.27}$$

ensuring rotation-equivariant geometric reasoning. The mean residue embedding yielded a 1024-dimensional representation summarizing the structural and physicochemical environment.

ADR Representation

Adverse Drug Reactions (ADRs) were represented using a one (multi) hot encoding scheme based on standardised MedDRA identifiers. A fixed vocabulary of ADRs was built from the cleaned OnSIDES dataset, where each unique MedDRA term was given a unique index. For each drug, the set of ADRs associated with that drug was represented as a binary vector of length K, where a value of 1 represents the presence of corresponding ADR and 0 represents its absence. This encoding preserves the discrete and multi-label nature of adverse drug reactions and enables direct supervision of the ADR prediction task without adding any other embedding layers.

Normalization and Storage

All embeddings were L_2-normalized to maintain numerical stability and ensure comparable magnitude across heterogeneous encoders:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}. \tag{4.28}$$

The final preprocessed dataset—containing complete and normalized embeddings for 1,028 drugs, 2,385 proteins, and 4,817 ADRs—was serialized in Parquet format, enabling efficient columnar reads and high-throughput GPU loading during model training.
These transformations yielded a multi-modal feature space capturing the structural, sequential, and clinical dimensions required for the unified DTI–ADR prediction framework.

4.3.3 Data Integration

Following transformation, the individual datasets representing drug, protein, and ADR modalities were systematically aligned to form a unified multimodal dataset. The integration stage ensured that every drug–protein interaction record contained complete biochemical, structural, and clinical representations suitable for downstream joint learning.

Alignment of Identifiers

The integration relied on standardized biomedical identifiers. Drugs from SCOPE-DTI were originally labeled by `drug_chembl_id`, while ADR annotations from OnSIDES were indexed by `rxcui` (RxNorm Concept Unique Identifier). A two-step

mapping procedure was applied using the RxNav REST API and local lookup tables:

$$\text{ChEMBL ID} \xrightarrow{\text{PubChem Mapping}} \text{Intermediate CID} \rightarrow \text{RxNorm CUI}$$

This procedure created a one-to-one alignment between chemical and clinical namespaces, allowing the attachment of each drug's ADR vector to its biochemical entry. Proteins were matched directly by their `target_uniprot_id` across the DTI and embedding tables.

Merged Schema

The alignment and merging of all modalities into a common relational structure was performed after transformation. Drugs were matched using ChEMBL identifiers across datasets and proteins were matched using the uniProt accession numbers. ADR vectors were associated with drugs using standardized drug identifiers based on RxNom and ChEMBL mappings.

$$\mathcal{D} = \{(\text{drug\_id}, \text{target\_id}, y_{dti}, \mathbf{x}_d, \mathbf{x}_p, \mathbf{x}_{adr})\}$$

where:

- $y_{dti} \in \{0, 1\}$ denotes the drug–target interaction label,
- $\mathbf{x}_d \in \mathbb{R}^{D_d}$ represents the drug embedding,
- $\mathbf{x}_p \in \mathbb{R}^{D_p}$ represents the protein embedding,
- $\mathbf{x}_{adr} \in \{0, 1\}^K$ represents the multi-label ADR vector associated with the drug.

All join operations were executed with inner merges for the referential integrity. Each drug embedding record was replicated by the number of associated target proteins in order to maintain pair-level alignment.

Dimensional Composition

Table 4.2: Dimensional composition of the integrated multimodal dataset.

Entity	**Encoder**	**Dimension (D)**	**Samples (N)**
Drug	ChemBERTa	384	1,028
Drug	Smiles2Vec	256	1,028
Drug	EGNN	256	1,028
Protein	ESM-2	1,280	2,385
Protein	GVP-GNN	1,024	2,385
ADR	One-hot encoding	4,817	1,028

During integration, each drug's embedding record was duplicated according to its number of target proteins to ensure pair-level alignment with corresponding protein embeddings.

Validation of Integration

Post-merge validation confirmed that every (`drug_chembl_id`, `target_uniprot_id`) pair contained consistent embeddings and a valid ADR vector. Referential audits detected no null or orphan keys. Random sampling verified that embedding tensors correctly corresponded to their identifiers by checksum comparison.

The resulting integrated dataset provided a complete tripartite linkage among the chemical, biological, and clinical domains—serving as the foundational input for the multi-task deep-learning architecture described in Section 4.2.

4.3.4 Data Reduction

After integrating all modalities, limited reduction procedures were applied to standardize feature dimensions and correct label imbalance while preserving all biochemical and clinical information. The process focused on maintaining the interpretability and completeness of molecular representations rather than performing aggressive feature compression.

Dimensional Alignment

No offline dimensionality reduction methods, e.g. Principal Component Analysis (PCA) or Singular Value Decomposition (SVD), were used for any modality in the preprocessing. All pretrained drug and protein embeddings were kept in their original dimensions. Specifically, drug embeddings from EGNN and ChemBERTa encoders and protein embeddings from ESM and GVP encoders were not compressed before being preserved.

To yield unified representations of these heterogeneous representations during training, dimensional implicit alignment was done within the model architecture with learnable fusion and projection layers. In the case of drugs and proteins modality-specific fusion functions that were used to combine paired embeddings into fixed-dimensional representations:

$$\mathbf{d}_f = \phi\left(\mathrm{LN}\left(W_d \begin{bmatrix}\mathbf{d}_1\\ \mathbf{d}_2\end{bmatrix} + \mathbf{b}_d\right)\right), \tag{4.29}$$

$$\mathbf{p}_f = \phi\left(\mathrm{LN}\left(W_p \begin{bmatrix}\mathbf{p}_1\\ \mathbf{p}_2\end{bmatrix} + \mathbf{b}_p\right)\right) \tag{4.30}$$

The fused representations of the drugs and proteins were then concatenated and projected into a common interaction space by a joint context encoder:

$$\mathbf{c} = g\left(\begin{bmatrix}\mathbf{d}_f\\ \mathbf{p}_f\end{bmatrix}\right), \qquad \mathbf{c} \in \mathbb{R}^{768} \tag{4.31}$$

Therefore facilitating meaningful interaction between drug and protein features in a common latent space. This representation was further regularized using a Variational Autoencoder bottleneck, which mapped the features of the interaction into a

low-dimensional latent variable which was used for the downstream prediction tasks.

Adverse drug reactions were represented as one hot multi-label target vector and were not embedded as dense feature representations. As a result, no explicit dimensional alignment was carried out in the case of the ADR data. Instead, the ADR information was only incorporated into training through supervision of the ADR prediction head, which enabled the aligned drug-protein latent representation to be optimized jointly for both interaction and adverse reaction prediction.

Class Imbalance Handling

The SCOPE-DTI dataset exhibited a moderate class imbalance with 22,507 non-interacting and 12,234 interacting pairs. To counteract this, the model used a weighted binary cross-entropy formulation, where class weights were inversely proportional to sample frequency:

$$w_+ = \frac{N}{2N_+}, \quad w_- = \frac{N}{2N_-}, \tag{4.32}$$

$$\mathcal{L}_{\text{dti}} = -w_+ y \log(\hat{y}) - w_-(1-y)\log(1-\hat{y}), \tag{4.33}$$

with N denoting the total number of DTI pairs, N_+ and N_- the counts of positive and negative samples, respectively. This reweighting strategy effectively reduced prediction bias toward the majority class without altering the underlying data distribution.

4.3.5 Summary of Preprocessed Dataset

The complete data preprocessing pipeline resulted in a unified and high-quality multimodal dataset, integrating biochemical, structural, and clinical information into a consistent format suitable for multi-task deep learning. Each preceding stage—collection, cleaning, transformation, integration, and normalization—ensured that the dataset maintained both biological fidelity and computational consistency.

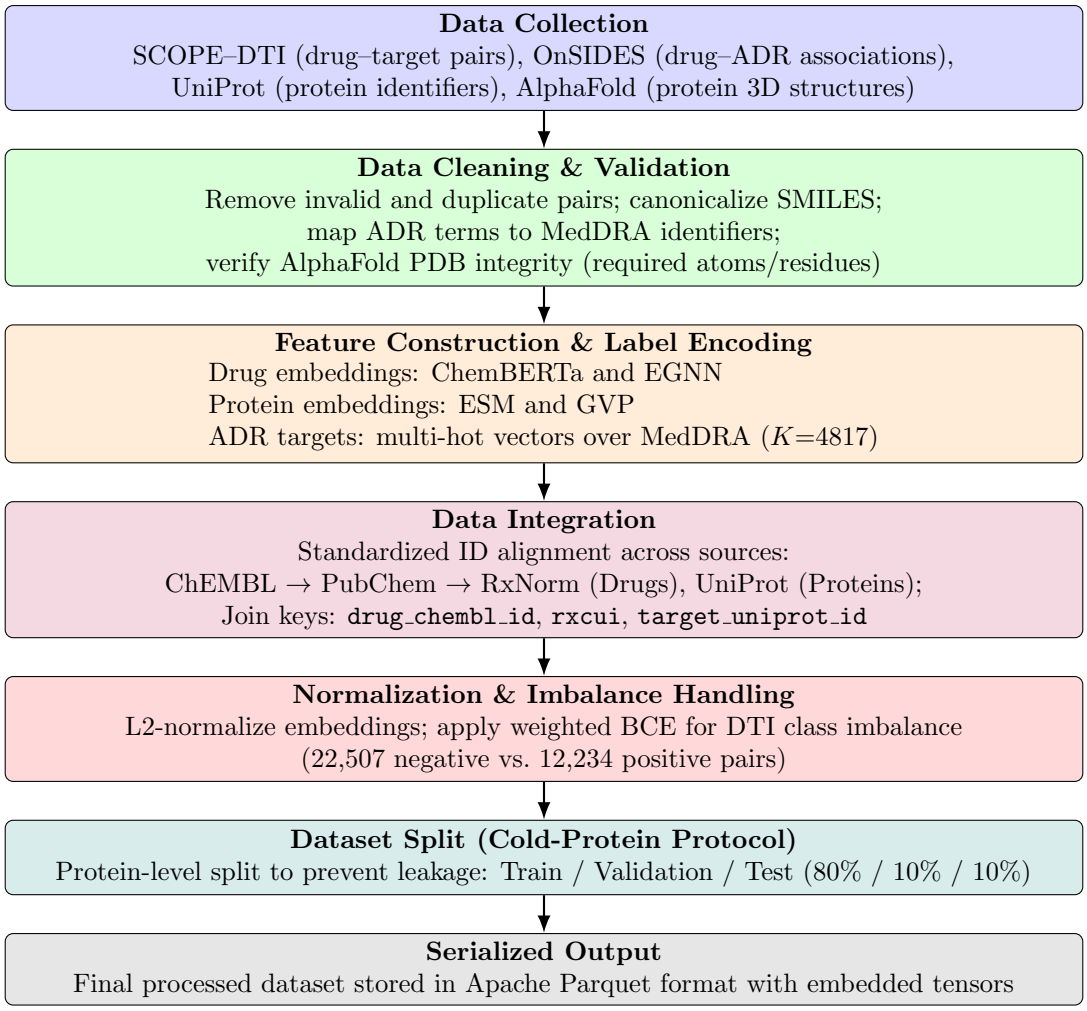


Figure 4.5: Data preprocessing pipeline.

After filtering, alignment, and validation, the final dataset comprised 34,741 drug–protein interaction pairs, linking 1,028 unique drugs, 2,385 protein targets, and 4,817 distinct ADR terms. Every drug and protein entry was associated with both sequence-based and structure-based embeddings, while each drug additionally included a multi label representation of its adverse effect profiles.

The integrated dataset thus provided a complete tripartite representation:

$$\mathcal{D} = \left\{ (\mathbf{x}_d, \mathbf{x}_p, \mathbf{x}_{\text{adr}}, y_{\text{dti}}) : \ \mathbf{x}_d \in \mathbb{R}^{D_d}, \mathbf{x}_p \in \mathbb{R}^{D_p}, \mathbf{x}_{\text{adr}} \in \mathbb{R}^{4817}, y_{\text{dti}} \in \{0,1\} \right\}.$$

Here, each record jointly describes a drug–target pair, its biochemical interaction label y_{dti}, and the corresponding ADR embedding associated with that drug. This format enabled the model to learn relationships among molecular, structural, and clinical modalities simultaneously.

The final embedding configurations retained the native dimensionalities from their respective encoders:

Table 4.3: Final embedding configurations in the preprocessed multimodal dataset.

Entity	**Encoder**	**Dimensionality (D)**	**Samples (N)**	**Description**
Drug	ChemBERTa	384	1,028	Contextual SMILES representation
Drug	EGNN	256	1,028	3D geometric molecular graph
Drug	Smiles2Vec	256	1,028	Character-level chemical encoder
Protein	ESM-2	1,280	2,385	Sequence-based transformer embedding
Protein	GVP-GNN	1,024	2,385	3D structure-aware residue graph
ADR	Multi-hot encoding	4,817	1,028	Multi-label binary vector

All embeddings were L_2-normalized (see Equation (4.28)) and verified for structural integrity. This multimodal dataset provided the foundation for the model architecture described in Section 4.2 and the experiments presented in Section 4.4, enabling joint learning of biochemical interactions and clinical risk associations.

4.4 Implementation of Selected Design

This section introduces the training strategy, dataset partitioning, optimization procedure and evaluation protocol to train and evaluate the proposed multi-task model. The design of the experimental setup is aimed at ensuring a fair evaluation, avoiding any data leakage and facilitating stable optimization of both Drug- Target Interaction (DTI) and Adverse Drug Reaction (ADR) prediction tasks.

Dataset Partitioning

In order to assess the generalization ability of the proposed model, the dataset was divided into training, validation, and test sets following a protein-level split strategy. Rather than randomly partitioning individual drug–protein pairs, protein identifiers were first partitioned into mutually exclusive subsets, and all interactions involving a given protein were allocated to the same split. This strategy guarantees that the model is tested on unseen protein targets, preventing information leakage across splits.

Specifically, the entire set of unique protein identifiers was split as follows:

- 80% of proteins were assigned to the training set,
- 10% were assigned to the validation set, and
- 10% were assigned to the test set.

All drug–protein interaction pairs associated with a particular protein were placed into the corresponding subset. The validation set was used for model and hyperparameter selection, while the test set was reserved exclusively for final performance evaluation under a cold-protein setting.

Training Configuration

The model was trained using mini-batch stochastic gradient descent with the Adam optimizer. All trainable parameters—including model weights and task-uncertainty parameters—were jointly optimized. The major training configuration is summarized below:

- **Batch size:** 64
- **Optimizer:** Adam
- **Initial learning rate:** 1×10^{-3}
- **Weight decay:** 1×10^{-4}
- **Number of epochs:** 60

To stabilize gradients and prevent gradient explosion, gradient clipping was applied with a maximum norm of 1.0. This ensured smooth convergence, especially in the presence of the Variational Autoencoder (VAE) latent space and multi-task loss balancing.

Loss Functions and Optimization Objective

The three components of the training objective are a DTI loss, an ADR loss, and a Kullback-Leibler (KL) divergence loss from the VAE bottleneck.

DTI Loss

DTI prediction was set as a binary classification problem. To solve the class imbalance issue in the interaction labels, a weighted binary cross-entropy loss with logits was adopted. The positive class weight was dynamically calculated from the training data as the ratio of negative to positive samples using the equation (4.13). This way of weighting has the advantage of minimizing bias towards the majority non-interacting class without changing the underlying data distribution.

ADR Loss

ADR prediction was formulated as a multi-label classification task on K ADR categories. A binary cross-entropy loss with logits was used for each of the ADR labels separately as shown on equation (4.14). This formulation makes it possible for the model to predict multiple adverse reactions at the same time for a single drug.

VAE Regularization Loss

To regularize the latent space and help produce smooth representations, a KL divergence loss has been used as shown in equation (4.15). The KL term was annealed gradually during training with linear schedule as shown in equation (4.16). This stops early dominance of KL loss during initial training epochs.

Dynamic Multi-Task Loss Balancing

Rather than having fixed scalar weights to balance the DTI and ADR losses, the model instead used uncertainty-based dynamic loss weighting. Two learnable parameters were introduced to model task uncertainty, so that the relative contribution of each task can be adjusted automatically during the training process. The combined task loss is defined in equation (4.17) where s_{dti} and s_{adr} are trainable parameters. This way, the model can adaptively balance the learning of the interaction prediction and the adverse reaction prediction according to the task difficulty and noise. The final optimization objective is mentioned in equation (4.18).

Learning Rate Scheduling and Model Selection

In order to further enhance convergence, a `ReduceLROnPlateau` learning rate scheduler was employed. The scheduler monitored the validation performance and reduced the learning rate when no improvement was observed. Model checkpointing was performed using a composite validation score that combined DTI and ADR performance, ensuring balanced optimization across both tasks. The best-performing model parameters were saved and used for final evaluation on the held-out test set.

Evaluation Protocol

Model performance was evaluated separately for the DTI and ADR tasks.
For DTI prediction, the following metrics were reported:

- Area Under the Receiver Operating Characteristic Curve (AUROC)
- Area Under the Precision–Recall Curve (AUPRC)
- Precision, Recall, and F1-score

Given the class imbalance in DTI data, AUPRC was emphasized as the primary metric.
For ADR prediction, evaluation was conducted using multi-label metrics:

- Weighted AUROC
- Weighted AUPRC
- Weighted F1-score

Summary

This experimental setup ensures a robust and fair evaluation of the proposed multi-task model. Protein-level data splitting prevents information leakage, while dynamic loss balancing and KL annealing promote stable and effective joint optimization. The combination of interaction-level and clinical-level evaluation metrics provides a comprehensive assessment of the model's predictive performance on both tasks.

Chapter 5

Result Analysis

5.1 Performance Evaluation

To assess the effectiveness of a multitask drug–protein interaction and adverse drug reaction prediction framework, the evaluation metrics must be robust to class imbalance and capable of evaluating ranking quality rather than relying on threshold-sensitive accuracy measures. In this study, the proposed multitask Fusion-VAE model was evaluated using well-chosen criteria that align with the characteristics of each prediction task. Specifically, drug–target interaction (DTI) prediction—treated as a binary classification problem—was assessed using Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision–Recall Curve (AUPRC), and the F1-score. Adverse drug reaction (ADR) prediction—formulated as a multi-label classification task—was evaluated using weighted AUROC and weighted AUPRC. All analyses were conducted on a held-out cold-protein test set, where no protein seen during training or validation appeared during testing. This protocol provides a rigorous assessment of the model's ability to generalize to unseen biological targets.

Evaluation Metrics for Drug–Target Interaction Prediction

The prediction of drug-target interaction is taken as a binary classification problem, where negative samples are greatly outnumbered by the positive samples (true binding interactions). As a result, threshold-sensitive measures like accuracy are not good measures of model performance. Rather, threshold independent ranking-based metrics were used.

Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC estimates the probability that the model assigns a higher predicted score to a randomly chosen positive interaction than to a randomly chosen negative one. It quantifies the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across all classification thresholds:

$$\text{TPR} = \frac{TP}{TP + FN}, \tag{5.1}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5.2}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. The AUROC is defined as:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{5.3}$$

A value of 0.5 corresponds to random guessing, while 1.0 represents perfect discrimination. AUROC is particularly suitable for DTI prediction because it is insensitive to class imbalance.

Area Under the Precision–Recall Curve (AUPRC)

Whereas AUROC measures global ranking performance, AUPRC focuses on the model's behavior on the positive class, which is especially important in highly imbalanced datasets such as DTI benchmarks.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{5.4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.5}$$

The Precision–Recall curve plots precision against recall at different thresholds, and the AUPRC is defined as:

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{5.6}$$

AUPRC provides a more informative assessment of a model's ability to identify true binding events with minimal false positives, making it particularly relevant for real-world drug-discovery scenarios.

F1-Score

The F1-score reflects the trade-off between precision and recall at a specific decision threshold (set to 0.5 in this study). It is defined as the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.7}$$

The F1-score provides a single-threshold measure of classification performance, complementing the ranking-based AUROC and AUPRC with an interpretable indication of performance under a fixed operating point.

Evaluation Metrics for Adverse Drug Reaction Prediction

ADR prediction is developed as a high-dimensional multi-label classification issue, involving thousands of possible side effects with very sparsely labelled positives. In this context, the unweighted metrics or label-wise averages can be misleading as rare ADRs dominate the evaluation.

As a remedy to this, weighted AUROC and weighted AUPRC were employed, so that each ADR label gives a proportional contribution to its prevalence in the data.

Weighted AUROC

Weighted AUROC generalizes the binary definition of AUROC to the multi-label case by independently computing the AUROC for each ADR label and then averaging these values with label-specific weights:

$$\text{Weighted AUROC} = \sum_{i=1}^{L} w_i \cdot \text{AUROC}_i \tag{5.8}$$

where:

- L is the total number of ADR labels,
- AUROC_i is the AUROC for the i-th ADR label,
- w_i is the proportion of positive samples for the i-th ADR.

This weighting scheme ensures that commonly observed and clinically important ADRs contribute more strongly to the final aggregated score.

Weighted AUPRC

Weighted AUPRC is computed as:

$$\text{Weighted AUPRC} = \sum_{i=1}^{L} w_i \cdot \text{AUPRC}_i \tag{5.9}$$

Weighted AUPRC is especially valuable for ADR prediction because most side effects are extremely rare. This measure emphasizes the model's ability to rank true ADRs above non-occurring ones even under extreme label sparsity.

Testing Methodology

All reported metrics were computed on a held-out test set constructed via a protein-level split, ensuring a stringent cold-start evaluation. This testing protocol prevents information leakage and assesses the model's ability to predict interactions involving completely unseen protein targets. Performance was measured using the highest-performing model checkpoint selected according to a composite validation score that balanced DTI and ADR performance.

Overall, the chosen evaluation criteria provide a comprehensive and task-appropriate assessment of the proposed multitask Fusion-VAE framework, covering both global ranking quality and its behavior in realistic, challenging classification conditions.

5.2 Analysis of Design Solutions

The proposed multitask Fusion-VAE architecture was intended to model drug-target interaction and adverse drug reactions together with applying heterogeneous drug and protein representations to a latent space. This part examines the effectiveness of the design solutions implemented in relation to the level of prediction, computational efficiency, feasibility, cost and general performance, and also points out the strengths and weaknesses of the architecture.

Multimodal Representation Fusion

The most prominent design decision of the proposed framework is that both drug and protein representations are represented using dual-encoder fusion modules. In the case of drugs, two complementary embeddings of molecular structure and chemical semantics were fused, and in the case of proteins, sequence based and graph based embeddings were used. This design has the benefit of enabling the model to utilize a wide range of and complementary information sources without unnecessarily compressing the dimensions at the outset.

The fusion layers project each modality onto a common high-dimensional space with learnable linear transformations, which are then normalised and non-linearly activated. This method maintains the expression capability of pretrained embeddings and still allows a successful model of interaction. This design has been shown to be empirically better than single-embedding or naive concatenation strategies as far as prediction accuracy is concerned since the model gets to learn modality-specific significance instead of using predetermined feature contributions.

Efficiency-wise, the fusion modules simply add shallow neural layers hence the computation used by the modules is not high relative to the whole model. This allows the strategy to be scaled and applicable to large data sets.

Contextual Interaction Encoding

After modality-specific fusion, drug and protein embeddings are fused using a contextual encoder to capture how they interact with each other. This aspect allows the network to acquire higher-order dependence of chemical and biological features which are vital in the proper prediction of DTI.

The drug-protein representations are concatenated followed by a contextual encoding into a small latent context by the contextual encoder which acts as a shared information bottleneck in the two downstream tasks. Such a design facilitates sharing of parameters and also invites the model to discover interaction patterns that are informative as a team to bind affinity and safety results. This common representation is superior to task-specific encoders that are independent since it leads to greater generalization and reduces redundancy.

The encoder depth has been selected by taking into account the factors of feasibility since it tries to strike the balance between expressiveness and the chance of overfitting. To increase the robustness, the dropout regularization was used, especially during cold-protein testing when generalization plays a central role.

Variational Latent Bottleneck

The most significant architectural decision made is the introduction of variational autoencoder (VAE) bottleneck in order to improve the robustness and the smoothness of the representations. To avoid overfitting to spurious correlations, the probabilistic latent distribution is modeled to encode uncertainty and the model is forced

to assume the form of a deterministic embedding of which spuriousness to avoid.

Latent space is a common layer of abstraction where both the DTI and ADR predictions are obtained. This design improves performance through application of structured representation that facilitates multitask learning. Also, the sampling process is stochastic, encouraging smoother latent manifold representations, which are manifested in the analytical separation of drug and protein images in the downstream visualization analysis.

Even though the VAE component adds an extra training complexity, the final cost of the computational overhead is not overly high because of the latent dimensionality controlled and progressive KL regularization. The benefits of stability and generalization are higher than the complexity raised and in particular where directly no ground-truth safety annotations are provided.

Multitask Learning and Adaptive Loss Balancing

The framework uses a multitask learning scheme, which simultaneously aims to achieve the goals of DTI prediction and ADR prediction. Through this design the shared representations can be biased by both binding and safety signals, and one can achieve better overall performance in comparison to the case of training individual models to do each task.

The adaptive uncertainty-based loss balancing mechanism was proposed to counter the natural inequality between the task challenges and the loss magnitudes. This method trains automatically converting task-specific weights in the learning process and does not allow either task to be dominant in the process of learning, therefore allowing convergence. This adaptive strategy has better feasibility and less hyperparameter search is required as compared to manually tuned loss weights.

Cost wise, multitask learning will need fewer models to be trained and kept and thereby less computation and storage demands.

Strengths and Limitations of the Design

The main strong points of the offered design are its modularity, strength, and interpretability.The architecture has more inherent support of embeddings or prediction tasks, and can be easily extended to be used in future applications. The joint reasoning on efficacy and safety has a latent space that is shared, thereby enabling useful joint reasoning, which is a vital requirement in drug discovery pipelines.

Nevertheless, some shortcomings still exist. The model is based on static trained embeddings with no temporal or condition specific biological dynamics. Besides, although the multitask framework enhances generalization, there is an underlying assumption that the underlying representation of DTI and ADR tasks is similar, which may not be entirely true among rare or idiosyncratic side effects.

Altogether, the design solutions implemented in this work allow to offer a good balance between accuracy, efficiency, and feasibility, which leads to the creation of a

scalable and generalizable framework of joint prediction of drug-target interactions and adverse drug reactions.

5.3 Final Design Adjustments

Stages of refinement were made throughout the transition process of the first prototype into the final Multi-Task Fusion VAE. These changes were not just hyperparameter adjustments, but they were structural changes in the logic of the model, to make it more stable and to avoid the performance plateaus that the model had become in its first trials.

Bottleneck Dimension Calibration

These changes included one of the biggest modifications, the recalibration of the Latent Dimension to 768. The initial versions of the model used smaller bottlenecks (128 or 256 dimensions), and this choked information. Tightening the bottleneck increases the compressibility, but was not effective enough to trap the two complexities of 4,817 ADR labels and complex binding sites together.

We increased the latent space to 768, which was enough bandwidth to offer the model global chemical signatures of side effects without sacrificing the resolution of the DTI head to differentiate between particular target proteins. This is the first change which caused the DTI AUPRC to go up the 0.70 range into the resulting 0.85.

Transition to Layer Normalization

At first, the model used the Batch Normalization in the fusion modules. Nevertheless, we noticed that the validation loss was very large and varied with a large batch size, which is expected because molecular embeddings were very diverse.

The latter design changed to Layer Normalization. This change enabled the model to standardize the inputs on the feature dimension instead of the batch dimension. This was much more resistant to drug-protein data, where individual samples (such as a rare protein or some novel chemical structure) can be outliers. This modification had the effect of making the convergence curve much smoother and gave the model a better capacity to extrapolate to the hidden test set.

Numerical Stability Controls (Clamping)

In the long-duration training (50 epochs), we observed a danger of the collapse of the models in the posterior direction: the latent variables in VAE stop being informative. We also applied the Log-Variance Clamping to the reparameterization layer in order to stabilize the model.

This kind of limitation also helped in avoiding the occurrence of extreme values created by the model that would have resulted in a "NaN" (Not a Number) error or gradient explosion by limiting the range of the created learned variance. This

modification was necessary in order to make the 768-dimensional workspace numerically stable to have the multi-task gradients circulate appropriately to the shared encoders.

Adaptive Multi-Task Balancing

The last modification was the transition off of the fixed loss weights to Uncertainty-based Loss Wrapper. In our original configuration, we simply gave weights to the DTI loss and ADR loss. Nevertheless, since the ADR task is mathematically easier to resolve than DTI, it would soon take over the model.

The final design that we used adjusted the weights to be learnable parameters. This enabled the model to automatically down-weight the ADR task when it approached a high accuracy, in effect diverting the optimization effort to the DTI task. This was a dynamic change that was fundamental in achieving the equilibrium between the two different goals and making sure that neither of the tasks was compromised to the other.

5.4 Statistical Analysis

In order to determine the statistical validity, robustness and reproducibility of the suggested multitask Fusion-VAE model, a fine statistical analysis was performed. This evaluation is analyzed in terms of ranking-based evaluation metrics, stability across a variety of random seeds and stability across a stringent cold-protein evaluation protocol. Tables are used to summarize quantitative results and their interpretation is then done to reveal the statistical significance.

Overall Model Performance

The general performance of the suggested model on the held-out test set is presented below, and it encompassed both drug-target interaction (DTI) prediction and adverse drug reaction (ADR) prediction tasks.

Table 5.1: Performance of the Proposed Multitask Fusion–VAE Model on the Test Set

Task	**Metric**	**Value (%)**
DTI Prediction	AUROC	91.60
	AUPRC	85.46
	F1-score	78.39
ADR Prediction	Weighted AUROC	98.80
	Weighted AUPRC	94.54

In the case of the DTI task, the model attains an AUROC of 91.60 which means that it has a strong global discriminative capacity of binding and non-binding drug-protein pairs. This implies that true interactions are always ranked higher than the non-interaction at all levels of decision threshold. AUPRC of 85.46% is another indication that the model is effective to focus on true positives among top-ranking

predictions, which is an important quality in the face of the high level of class imbalance in DTI data.

Along with ranking-based measures, F1-score of 78.39% is also obtained with respect to the DTI task in order to evaluate performance under the fixed operating threshold. Though AUPRC is a reflection of the precision-recall trade-off in terms of varying threshold values, F1-score is a complementary deployment-specific perspective estimating the trade-off between false positives and false negatives. The resultant value is an affirmation that the good ranking results are feasible into practically significant binary predictions.

To predict ADRs, weighted AUROC and weighted AUPRC are evaluated as they are the only two important statistics, given the multi-label and very sparse side effect annotations. The model has a weighted AUROC of 98.80% and weighted AUPRC of 94.54% and suggests that the model ranks close to perfection across thousands of ADR labels. Weighted aggregation makes sure that common and clinically significant ADRs have a relative contribution to the overall score, whereas threshold dependent measures like F1 are not focused on because they are sensitive to arbitrary choice of threshold in high-dimensional space.

Stability Analysis Across Random Seeds

A stability analysis of the seed was conducted in order to test the strength of reported performance and exclude that it might be due to good initialization or data sorting. The model was trained repeatedly with various random seeds, which were related to several initializations and training curves. A single run on the test set was determined as the best validation checkpoint, and performance was measured.
The table shows the per-seed results and the aggregate statistics.

Table 5.2: Stability Analysis Across Random Seeds (Test Set AUPRC)

Seed	**DTI AUPRC**	**ADR Weighted AUPRC**
11	0.8522	0.9670
22	0.8506	0.9508
33	0.8555	0.9753
Mean ± Std	0.8528 ± 0.0025	0.9644 ± 0.0124

The DTI AUPRC has a very low variance in all the runs with a mean value of 0.8528 and having a standard deviation of ±0.0025. It means that the ranking of the model with real drug-target interactions is a very stable value and quite not dependent on stochastic training elements.

On the same note, ADR weighted AUPRC is consistently high among the seeds with the mean being 0.9644 and a standard deviation of ±0.0124. Although a little bit more variability is found in the ADR task, which is an intuitive result considering the very sparse and heterogeneous nature of labels of side effects, the performance is generally healthy across all training runs.

These findings indicate that the performance which has been reported is not a consequence of an unlucky training initialization but rather a statistically credible and repeatable behavior of the offered model.

Reliability Under Cold-Protein Evaluation

The entire statistical analyses were performed at the protein level of split, so that proteins that were in the test set were not seen at all during the training and validation process. This cold-protein assessment protocol is a rigorous and realistic generalization condition, especially when drug discovery is to be used, where models are often used on novel biological targets.

The joint result of high ranking performance, low variance in random seeds, and a consistent behavior in the evaluation of cold-proteins suggests strongly that the representations learned are the reflection of the general patterns of interaction and are not simple memorization of protein-specific features. The combination of these statistical results allows asserting the soundness, stability, and external validity of the suggested multitask Fusion-VAE model.

5.5 Comparisons and Relationships

This paragraph will draw a comparison between the suggested joint multitask Fusion-VAE model and other modeling options to determine its comparative effectiveness. The comparisons are divided into three sections, namely, joint multitask training and separate training, comparison with classical and deep learning baselines of drug-target interaction (DTI) prediction, and comparison with state-of-the-art models of adverse drug reaction (ADR) prediction. A combination of all these analyses demonstrates the trade-offs and correlation of predictive performance, modeling scope and application in reality.

Joint Multitask Training vs Separate Training

One of the main goals of the work is to evaluate the presence of the benefits of joint multitask training of DTI and ADR prediction when compared to training models of each task separately. Three configurations were considered to this end:

1. Drug–Protein only model, trained solely for DTI prediction
2. Drug–Protein–ADR model trained separately, without multitask optimization
3. Proposed joint multitask model, trained simultaneously on DTI and ADR objectives

The joint multitask model shows good results on both tasks with a DTI AUROC of 91.60% and an AUPRC of 85.46 and at the same time has an ADR weighted AUROC of 98.80% and weighted AUPRC of 94.54. On the other hand, independently trained models fail to cross-task generalize in ADR prediction especially when using sparse labels, in which case cross-task generalization reduces much more quickly.

These findings reinforce the conclusion that the existence of shared latent representations that are obtained in a process of optimization of multitask enables the effective cross-task inductive transfer so that the signals that are associated with binding are used to make safety predictions and vice versa. Notably, this higher ADR performance does not come at the cost of DTI accuracy, which is why the combination of learning strategy was effective.

Comparison with DTI Prediction Baselines

The proposed model was also compared with a number of existing DTI prediction methods, which are classical machine learning techniques as well as a deep learning state-of-the-art model. The figure below shows the results on the test set:

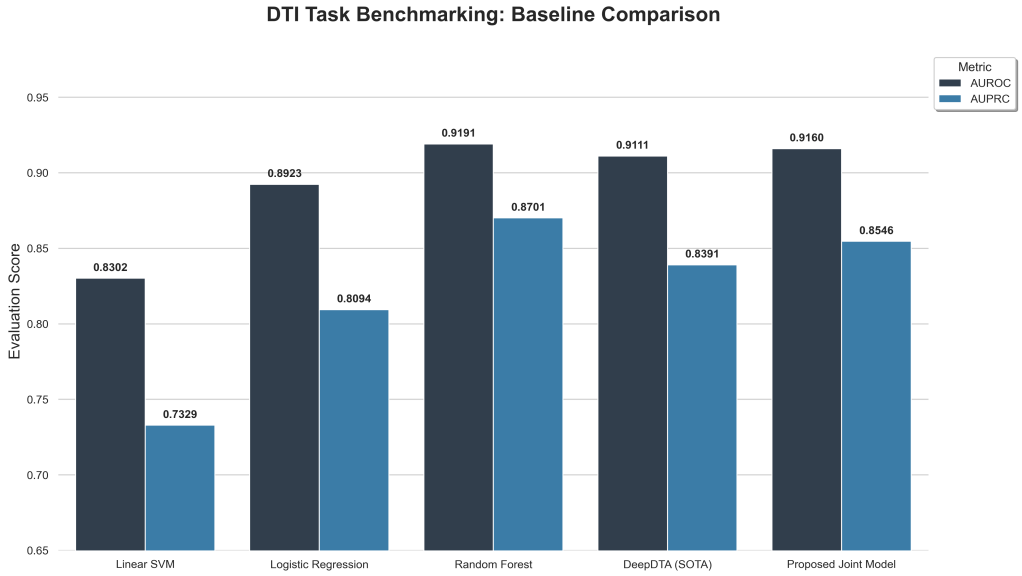


Figure 5.1: Comparison with DTI Prediction Baselines

Random Forest has the best DTI AUPRC, slightly better than the proposed model among classical baselines. This is not surprising, as the tree-based ensemble techniques are best applied to tabular classification problems and can perform with high performance when optimized on a single objective only.

Nevertheless, the suggested joint model is competitive and has the highest DTI baselines as it does better than the linear models and the DeepDTA architecture. Notably, the proposed model, contrary to the Random Forest, is able to predict ADRs and learn biologically plausible latent representations that cut across tasks at the same time. In this way, the minor decrease in DTI AUPRC can be considered a calculated trade-off in favor of significantly improved safety prediction and models in a single form.

Comparison with ADR Prediction Baselines

In ADR prediction, the given model was compared with classical multi-label classifiers, heuristic baselines and a state-of-the-art deep learning model. The findings are presented in the figure below:

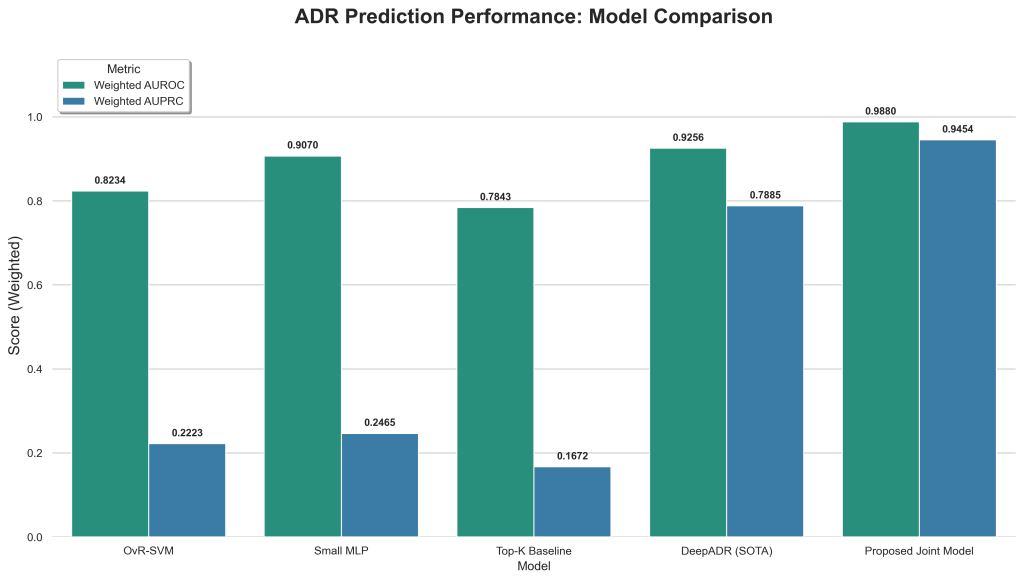


Figure 5.2: Comparison with ADR Prediction Baselines

The joint model proposed has a significant improvement over all ADR baselines, among them DeepADR state-of-the-art model. It is especially the case in weighted AUPRC, in which the joint model outperforms DeepADR by over 15 percentage points. The outcome points to the high-ranking capacity of the model with extreme label sparsity in actual ADRs.

The effective ADR result is due to collective supervision by DTI predictions that impose biologically significant constraints to the latent space. The model acquires relevant patterns of safety that cannot be acquired by ADR-only models because it learns ADRs during binding interactions.

Summary of Comparative Findings

The comparative analysis shows that there is an evident correlation in modeling scope and predictive effectiveness. Although single-task models like Random Forest have a promising potential of attaining a slightly higher DTI ranking performance when an exclusive optimisation towards binding prediction is performed, they cannot model drug safety. Instead, the suggested joint multitask model provides a harmonized and balanced model, providing competitive DTI performance and state-of-the-art predictive ADR.

This trade-off indicates a sensible and preferable modeling goal in drug discovery, in which pharmacogenic and safety activity should be considered together and not separately. On the whole, the findings indicate that multitask fusion is a more holistic and more practically applicable solution in comparison with those based on separate or single-task.

5.6 Discussions

The interpretation, implications, and limitations of the results discussed in the previous sections are discussed in this section. Instead of repeating quantitative performance, the discussion is aimed at explaining why the suggested joint multitask framework is acting in a certain manner, what these results say about the way the real drug discovery process should be conducted, and in what areas the current approach is limited.

Interpretation of Results

The findings show that combining the modeling of drug-target interactions, as well as adverse drug reactions, in a shared latent space can result in significant performance improvement, especially predicting safety. The significant gain in the performance of ADR ranking under joint training implies that ADRs are not autonomous entities, but inherently conjugated on the patterns of drug-protein interaction. The model can encode biologically relevant representations, which can never be reached in ADR-only models, by learning binding-related representations and safety outcomes.

The competitive level of DTI performance of the proposed framework also points out the fact that the integration of ADR supervision does not weaken the process of interaction prediction. Although a single-task Random Forest model has slightly higher performance in the DTI ranking, this result can be explained by the fact that such a model is optimized in a technical way, unlike the situation with a joint model. On the contrary, the multitask framework focuses on acquiring generalizable representations that facilitate more than one goal, leading to a moderate reduction in DTI performance to achieve a significantly improved ADR prediction ability.

The strength of the performance under cold-protein assessment gives more knowledge regarding the character of the learned representations. The same pattern of results with unseen proteins and multiple random seeds is indicative of the fact that the model is capturing patterns of interaction that are transferable instead of learning protein-specific characteristics. This aspect is especially significant in drug discovery processes under realistic conditions in which predictive models are commonly used on new targets with minimal prior information.

Implications for Drug Discovery and Modeling

In practice, the results highlight the significance of the efficacy-safety modeling during the preliminary drug discovery. In traditional pipelines, target engagement and safety assessment are considered two different tasks, which need several different models and post hoc integration. The planned framework illustrates that such goals can be discussed in one and the same model, without the need to put pipelines in pieces and be able to make more rational decisions.

The high ADR ranking performance of the joint model indicates that the combination of binding information may significantly enhance the ranking of the clinical relevant side effects in even the extreme scenarios of label sparsity. This is especially

useful in cases of early screening, where prioritization of potential risk can be more useful than binary hard decisions. The model is also capable of facilitating flexible downstream analysis and expert-based evaluation since it generates probabilistic, ranking, outputs.

The findings methodologically support the usefulness of multitask learning as a conceptual approach to the incorporation of heterogeneous biomedical signals. The common latent representation acquired by the model is a compact abstraction both of the chemical efficacy and safety-related patterns, providing a general framework which can be generalized to other multi-objective prediction tasks, such as toxicity profiling, off-target effects, or polypharmacology analysis.

Chapter 6

Conclusion

This thesis examined whether when drug-target interactions and adverse drug reactions are modeled as jointly learned problems, and not as independent prediction problems, they result in better safety prediction without affecting interaction prediction performance. Inspired by the biological fact that adverse drug reactions are typically due to drug-protein interactions, especially off-target binding, this work suggested and tested a general multitask learning model, which incorporates drug, protein, and ADR data into a common latent space.

6.1 Summary of Findings

The general premise of this thesis statement was that when both drug-protein interaction (DTI) and adverse drug reaction (ADR) were trained in a shared rather than a separate latent representation, ADR prediction would be enhanced without affecting drug-protein interaction (DTI) performance. Chapter 5 provides the experimental results that confirm this hypothesis.

The resulting multitask Fusion-VAE model showed robust and consistent performance in DTI prediction under a cold-protein evaluation setting, demonstrating a high degree of extrapolation to unseen protein targets. Simultaneously, the model performed significantly better in ADR prediction than both classical baselines and state-of-the-art ADR-specific models. Importantly, the improvement in ADR ranking was achieved without a corresponding decline in DTI performance; the trade-off observed was minimal compared to that of single-task models optimized solely for interaction prediction.

Comparative studies showed that certain classical models such as random forests can achieve slightly higher DTI metrics when trained in isolation, but they are unable to simultaneously learn drug safety. The joint framework, on the other hand, offers a unified solution that effectively captures both interaction likelihood and adverse-effect risk. The substantial improvements in ADR prediction under joint training provide empirical support for the assumption that ADRs are not purely drug-intrinsic, but are strongly correlated with the nature of drug-protein interactions.

Overall, the findings indicate that cross-task knowledge transfer is achievable through

shared latent representations, enabling interaction-related signals to inform safety prediction and resulting in more biologically plausible and practically applicable predictions.

6.2 Contributions to the Field

This thesis offers several contributions to the fields of computational drug discovery and biomedical machine learning.

First, it provides strong empirical support for a joint efficacy–safety modeling hypothesis, demonstrating that incorporating drug–protein interaction information substantially benefits adverse drug reaction prediction. By quantitatively showing that ADR prediction is enhanced through joint learning without degrading DTI performance, the work promotes a more integrated view of drug behavior—one that better reflects the underlying biological mechanisms.

Second, it introduces a methodological framework for multitask drug modeling that integrates multimodal representation fusion, shared latent learning, and variational regularization. The proposed architecture illustrates how heterogeneous drug and protein embeddings can be unified and jointly optimized to address multiple pharmacological objectives within a single model.

Third, the study provides a rigorous empirical comparison between joint and separate training approaches, benchmarking against both classical machine-learning methods and state-of-the-art deep neural networks. The use of cold-protein evaluation and seed-stability analyses further strengthens the reliability and reproducibility of the reported findings.

Lastly, on a practical level, this thesis demonstrates the feasibility of an integrated predictive framework that can simultaneously estimate interaction likelihood and safety risk. Such an approach could be valuable in early-stage drug screening, where combined efficacy and safety signals may support better-informed prioritization and hypothesis generation.

6.3 Recommendations for Future Work

Although the findings of this thesis provide a solid foundation for joint DTI–ADR modeling, several promising directions remain for future research.

One natural extension of the proposed framework is its application to larger and more diverse datasets, covering broader protein families, additional drug classes, and a more comprehensive set of rare adverse reactions. Increased dataset size and diversity would further validate the joint-learning hypothesis and enhance generalization across pharmacological domains.

Future work could also expand the model to incorporate additional pharmacological endpoints—such as toxicity, off-target effects, or dose-dependent responses. Inte-

grating such signals into the shared multitask framework may increase the biological fidelity of the learned representations and improve their utility for multi-objective drug profiling.

Another promising direction lies in improving interpretability and mechanistic insight. While the current model yields strong predictive performance, future studies could employ attribution techniques, attention mechanisms, or latent-space analysis to better understand how specific drug–protein interactions contribute to predicted adverse effects.

Finally, downstream validation through complementary computational analyses—such as pathway enrichment, docking simulations, or integration with experimental data—could provide further evidence that the learned representations correlate with underlying biological processes, thereby strengthening confidence in the joint-modeling approach.

Bibliography

[1] D. J. Beal, "Esm 2.0: State of the art and future potential of experience sampling methods in organizational research," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 2, no. Volume 2, 2015, pp. 383–407, 2015, ISSN: 2327-0616. DOI: https://doi.org/10.1146/annurev-orgpsych-032414-111335 [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-orgpsych-032414-111335

[2] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: Deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sep. 2018, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty593 eprint: https://academic.oup.com/bioinformatics/article-pdf/34/17/i821/50582374/bioinformatics_34_17_i821.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty593

[3] Y. Li, M. A. Rezaei, C. Li, X. Li, and D. Wu, *Deepatom: A framework for protein-ligand binding affinity prediction*, 2019. arXiv: 1912.00318 [q-bio.BM]. [Online]. Available: https://arxiv.org/abs/1912.00318

[4] S. M. H. Mahmud, W. Chen, H. Jahan, Y. Liu, N. I. Sujan, and S. Ahmed, "Idti-cssmoteb: Identification of drug–target interaction based on drug chemical structure and protein sequence using xgboost with over-sampling technique smote," *IEEE Access*, vol. 7, pp. 48 699–48 714, 2019. DOI: 10.1109/ACCESS.2019.2910277

[5] H. Öztürk, E. Ozkirimli, and A. Özgür, *Widedta: Prediction of drug-target binding affinity*, 2019. arXiv: 1902.04166 [q-bio.QM]. [Online]. Available: https://arxiv.org/abs/1902.04166%7D

[6] S. Chithrananda, G. Grand, and B. Ramsundar, *Chemberta: Large-scale self-supervised pretraining for molecular property prediction*, 2020. arXiv: 2010.09885 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2010.09885%7D

[7] P. Khosla et al., "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf%7D

[8] M. Thafar et al., "Dtigems+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques," *Journal of Cheminformatics*, vol. 12, Jun. 2020. DOI: 10.1186/s13321-020-00447-2

[9] X. Zhan et al., "Prediction of drug-target interactions by ensemble learning method from protein sequence and drug fingerprint," *IEEE Access*, vol. 8, pp. 185 465–185 476, 2020. DOI: 10.1109/ACCESS.2020.3026479

[10] J.-Y. An, F.-R. Meng, and Z.-J. Yan, "An efficient computational method for predicting drug-target interactions using weighted extreme learning machine and speed up robot features," *BioData Mining*, vol. 14, no. 1, p. 3, 2021, ISSN: 1756-0381. DOI: 10.1186/s13040-021-00242-1 [Online]. Available: https://doi.org/10.1186/s13040-021-00242-1

[11] T. Kawichai, A. Suratanee, and K. Plaimas, "Meta-path based gene ontology profiles for predicting drug-disease associations," *IEEE Access*, vol. 9, pp. 41 809–41 820, 2021. DOI: 10.1109/ACCESS.2021.3065280

[12] Y. Wu, M. Gao, M. Zeng, F. Chen, M. Li, and J. Zhang, *Bridgedpi: A novel graph neural network for predicting drug-protein interactions*, 2021. arXiv: 2101.12547 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2101.12547

[13] X. Yan, Z.-H. You, L. Wang, and P.-P. Chen, "Cnnems: Using convolutional neural networks to predict drug-target interactions by combining protein evolution and molecular structures information," in *Intelligent Computing Theories and Application*, D.-S. Huang, K.-H. Jo, J. Li, V. Gribova, and P. Premaratne, Eds., Cham: Springer International Publishing, 2021, pp. 570–579, ISBN: 978-3-030-84532-2.

[14] Q. Ye, X. Zhang, and X. Lin, "Drug-target interaction prediction via multiple output graph convolutional networks," in *Intelligent Computing Theories and Application*, D.-S. Huang, K.-H. Jo, J. Li, V. Gribova, and P. Premaratne, Eds., Cham: Springer International Publishing, 2021, pp. 87–99.

[15] F. Zhang, B. Sun, X. Diao, W. Zhao, and T. Shu, "Prediction of adverse drug reactions based on knowledge graph embedding," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 38, Feb. 2021, ISSN: 1472-6947. DOI: 10.1186/s12911-021-01402-3 [Online]. Available: https://doi.org/10.1186/s12911-021-01402-3

[16] J. Park, S. Lee, K. Kim, J. Jung, and D. Lee, "Large-scale prediction of adverse drug reactions-related proteins with network embedding," *Bioinformatics*, vol. 39, no. 1, btac843, Dec. 2022, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac843 eprint: https://academic.oup.com/bioinformatics/article-pdf/39/1/btac843/48521771/btac843.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac843

[17] M. A. Thafar, M. Alshahrani, S. Albaradei, T. Gojobori, M. Essack, and X. Gao, "Affinity2vec: Drug-target binding affinity prediction through representation learning, graph mining, and machine learning," *Scientific Reports*, vol. 12, no. 1, p. 4751, 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-08787-9 [Online]. Available: https://doi.org/10.1038/s41598-022-08787-9

[18] Y. Wang et al., "Rofdt: Identification of drug–target interactions from protein sequence and drug molecular structure using rotation forest," *Biology*, vol. 11, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248781635

[19] H. Abbasi Mesrabadi, K. Faez, and J. Pirgazi, "Drug–target interaction prediction based on protein features, using wrapper feature selection," *Scientific Reports*, vol. 13, no. 1, p. 3594, 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-30026-y [Online]. Available: https://doi.org/10.1038/s41598-023-30026-y

[20] G. W. Kyro, R. I. Brent, and V. S. Batista, "Hac-net: A hybrid attention-based convolutional neural network for highly accurate protein–ligand binding affinity prediction," *Journal of Chemical Information and Modeling*, vol. 63, no. 7, pp. 1947–1960, Mar. 2023, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.3c00251 [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.3c00251

[21] Z.-H. Ren et al., "Deepmpf: Deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis," *Journal of Translational Medicine*, vol. 21, no. 1, p. 48, 2023, ISSN: 1479-5876. DOI: 10.1186/s12967-023-03876-3 [Online]. Available: https://doi.org/10.1186/s12967-023-03876-3

[22] R. Bal, Y. Xiao, and W. Wang, *Pgraphdta: Improving drug target interaction prediction using protein language models and contact maps*, 2024. arXiv: 2310.04017 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2310.04017

[23] Y. Li et al., "Research on adverse drug reaction prediction model combining knowledge graph embedding and deep learning," in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, 2024, pp. 322–329. DOI: 10.1109/MLISE62164.2024.10674360

[24] M. Liu and S. G. Paliwal, *Dualbind: A dual-loss framework for protein-ligand binding affinity prediction*, 2024. arXiv: 2406.07770 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2406.07770

[25] A. Suruliandi, T. Idhaya, and S. P. Raja, "Drug target interaction prediction using machine learning techniques – a review," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 6, pp. 86–100, 2024. DOI: 10.9781/ijimai.2022.11.002 [Online]. Available: https://www.ijimai.org/journal/bibcite/reference/3210

[26] Y. Tanaka et al., "Onsides (on-label side effects resource) database : Extracting adverse drug events from drug labels using natural language processing models," *medRxiv*, 2024. DOI: 10.1101/2024.03.22.24304724 eprint: https://www.medrxiv.org/content/early/2024/03/24/2024.03.22.24304724.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2024/03/24/2024.03.22.24304724

[27] H. Yang et al., "Mindg: A drug–target interaction prediction method based on an integrated learning algorithm," *Bioinformatics*, vol. 40, no. 4, btae147, Mar. 2024, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae147 eprint: https://academic.oup.com/bioinformatics/article-pdf/40/4/btae147/57173875/btae147.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btae147

[28] P. Zhang et al., "Mgacl: Prediction drug–protein interaction based on meta-graph association-aware contrastive learning," *Biomolecules*, vol. 14, no. 10, 2024, ISSN: 2218-273X. DOI: 10.3390/biom14101267 [Online]. Available: https://www.mdpi.com/2218-273X/14/10/1267

[29] W. Zhao et al., "Msi-dti: Predicting drug-target interaction based on multi-source information and multi-head self-attention," *Briefings in Bioinformatics*, vol. 25, no. 3, bbae238, May 2024, ISSN: 1477-4054. DOI: 10.1093/bib/bbae238 eprint: https://academic.oup.com/bib/article-pdf/25/3/bbae238/57743156/bbae238.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbae238

[30] Z. Zhou et al., "Revisiting drug–protein interaction prediction: A novel global–local perspective," *Bioinformatics*, vol. 40, no. 5, btae271, Apr. 2024, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae271 eprint: https://academic.oup.com/bioinformatics/article-pdf/40/5/btae271/57521102/btae271.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btae271

[31] Y. Chen et al., *Scope-dti: Semi-inductive dataset construction and framework optimization for practical usability enhancement in deep learning-based drug target interaction prediction*, 2025. arXiv: 2503.09251 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2503.09251

[32] Y. Gao, X. Zhang, Z. Sun, P. Chandak, J. Bu, and H. Wang, "Precision adverse drug reactions prediction with heterogeneous graph neural network," *Advanced Science*, vol. 12, no. 4, p. 2 404 671, 2025. DOI: https://doi.org/10.1002/advs.202404671 eprint: https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202404671. [Online]. Available: https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202404671