GRAPH TRANSFORMER BASED PERSONALIZED DRUG RECOMMENDATION SYSTEM FOR CARDIOVASCULAR DISEASE

A PROJECT REPORT

Submitted by

ASHWIN PRABHU M

(2023176029)

A report for the phase-I of the project submitted to the Faculty of

INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment

for the award of the degree

of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

SPECIALIZATION IN AI & DS

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600 025

DEC 2024

ANNA UNIVERSITY

CHENNAI - 600 025

BONA FIDE CERTIFICATE

Certified that this project report titled GRAPH TRANSFORMER BASED PERSONALIZED DRUG RECOMMENDATION SYSTEM FOR CARDIOVASCULAR DISEASE is the bona fide work of ASHWIN PRABHU M(2023176029) who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE:
DATE:

Dr. T MALA
PROFESSOR
PROJECT GUIDE
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

DR. S. SWAMYNATHAN
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

This project focuses on enhancing Drug-Target Interaction (DTI) prediction to address the critical problem of identifying novel therapeutic candidates while reducing reliance on lengthy clinical trials. Drug innovation cannot proceed without accurate prediction of DTIs, especially in the public health domain where economical and effective treatments are required. To predict the binding affinities between medicines and target proteins, this study presents a unique computational framework that combines transformer-based protein embeddings with graph-based molecular representations. To process drug compounds that are represented as molecular graphs generated from SMILES strings using RDKit, and the methodology leverages Graph Transformers to capture the complex interconnectedness between atoms and bonds using dynamic edge characteristics. Protein amino acid sequences are subjected to ProtBert embeddings to extract significant structural and sequential features. These embeddings align the drug and protein features to accurately calculate interaction scores when paired with multi-head attention mechanisms. The Adam optimizer, with a Mean Squared Error (MSE) loss function, is used to optimize the model.

The model was trained and evaluated using the benchmark dataset (KIBA) and demonstrated high predictive performance with an MSE of **0.3751** and an MAE of **0.5745**. The proposed method offers significant benefits for early-stage drug discovery by providing reliable and precise DTI predictions. Through the interpretation of attention heatmaps, the model enhances interpretability by offering insights into the crucial atoms and amino acids involved in drug-target binding. The findings show that transformer models and graph neural networks can be effectively combined to improve DTI prediction, making the process more reliable and efficient.

ACKNOWLEDGEMENT

I would like to express my deep sense of appreciation and gratitude to my project guide Dr. T. MALA, Professor, Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai for her invaluable support, supervision, guidance, useful suggestions and encouragement throughout this phase. Her moral support and continuous guidance enabled me to complete my work successfully.

I thank Dr. S. SWAMYNATHAN Professor and Head of the Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai for the prompt and limitless help in providing the excellent computing facilities to do the project and to prepare the thesis.

I am grateful to the project committee members Dr. S. SRIDHAR Professor, Dr. G. GEETHA Associate Professor and Dr. D. NARASHIMAN Teaching Fellow, Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai for their review and valuable guidance throughout the course of my project.

I would also like to thank the faculty member and nonteaching staff members of the Department of Information Science and Technology at College of Engineering Guindy, Anna University, Chennai for their valuable support throughout the course of my project work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

BINDTI	Bi-directional Intention Network for Drug-Target Interaction
CNN	Convolution Neural Network
CVDs	Cardio Vascular Diseases
DTI	Drug-Target Interaction
GCN	Graph Convolution Network
GNN	Graph Neural Network
KIBA	Kinase Inhibitor BioActivity
MAE	Mean-Absoulute Error
MHA	Multi-Head Attention
ML	Machine Learning
MSE	Mean-Squared Error
RNN	Recurrent Neural Network
SMILES	Simplified Molecular Input Line Entry System
TPSA	Topological Polar Surface Area
WHO	World Health Organization

CHAPTER 1

INTRODUCTION

Drug-Target Interactions (DTIs) are essential to the drug development process because they establish the therapeutic potential and biological efficacy of drug candidates against certain protein targets. Finding new drug leads, maximizing treatment results, and confirming target proteins all depend on accurate DTI identification. DTIs have historically been studied through experimental techniques like clinical trials and in vitro experiments. But these methods take a lot of time and money, which slows down the development of new drugs. With its ability to anticipate drug-protein interactions more quickly and affordably while lowering the need for physical tests, computational approaches have become a game-changer.

A large percentage of morbidity and mortality globally are caused by cardiovascular diseases (CVDs), making them one of the biggest public health concerns. The World Health Organization (WHO) estimates that cardiovascular diseases (CVDs) account for around 18 million deaths each year, making them the most common cause of deaths from non-communicable diseases. The multifaceted character of CVDs, which include lifestyle variables, environmental factors, and genetic predispositions, contributes to their complexity. The desire for efficient, individualized therapies and its complexity put a great deal of strain on drug development pipelines and healthcare systems. Furthermore, the need for accurate forecasts of medication safety and efficacy is underscored by the variation in patient reactions to treatments.

Recent developments in computational methods and artificial

intelligence have opened up exciting possibilities for enhancing drug discovery. Graph Neural Networks (GNNs) and Transformer architectures are two examples of machine learning and deep learning models that have shown remarkable effectiveness in simulating the complex interactions between drug compounds and protein structures. These models use transformer-based protein sequence embeddings and graph-based molecular compound representations to capture both local and global properties that are essential to DTIs. Computational techniques also help in drug repurposing, which saves time and money by identifying current medications for novel therapeutic applications.

The goal of this project is to use these computational developments to tackle the difficulties in cardiovascular drug discovery. This study attempts to provide a strong predictive framework for DTIs by combining transformer-based protein embeddings with graph-based molecular learning. By increasing the precision of drug-target interaction predictions, the suggested method aims to reduce expenses and the requirement for lengthy clinical trials while facilitating the quicker identification of promising drug candidates. By doing this, the study hopes to address one of the most important global health issues and aid in the creation of individualized therapies for cardiovascular disorders.

1.1 BACKGROUND

Cardiovascular diseases (CVDs) are a major cause of death globally, contributing to about 17.9 million fatalities per year, or 31% of all deaths worldwide[1]. In order to address the increasing burden of CVDs on public health and healthcare systems, this concerning figure emphasizes the urgent need for novel therapeutic approaches. Finding specific drug-target interactions (DTIs) is often necessary for the effective treatment of cardiovascular disorders. DTIs are essential for drug discovery and personalized medicine. In order to find

novel therapeutic possibilities and lessen side effects, DTIs assist in determining the binding affinity between drug molecules and their target proteins.

Finding drug-target interactions as seen in Figure 1.1 has hitherto mostly depended on expensive and time-consuming experimental techniques like high-throughput screening and clinical trials. Furthermore, these methods frequently fall short when drug candidates and protein targets become more complex[2]. Computational approaches have become a game-changer in recent years, providing effective and affordable substitutes for conventional drug development procedures. Researchers can increase the success rate of virtual screening and early-stage drug development by using artificial intelligence (AI) and machine learning (ML) tools to more accurately anticipate drug-protein interactions[3].

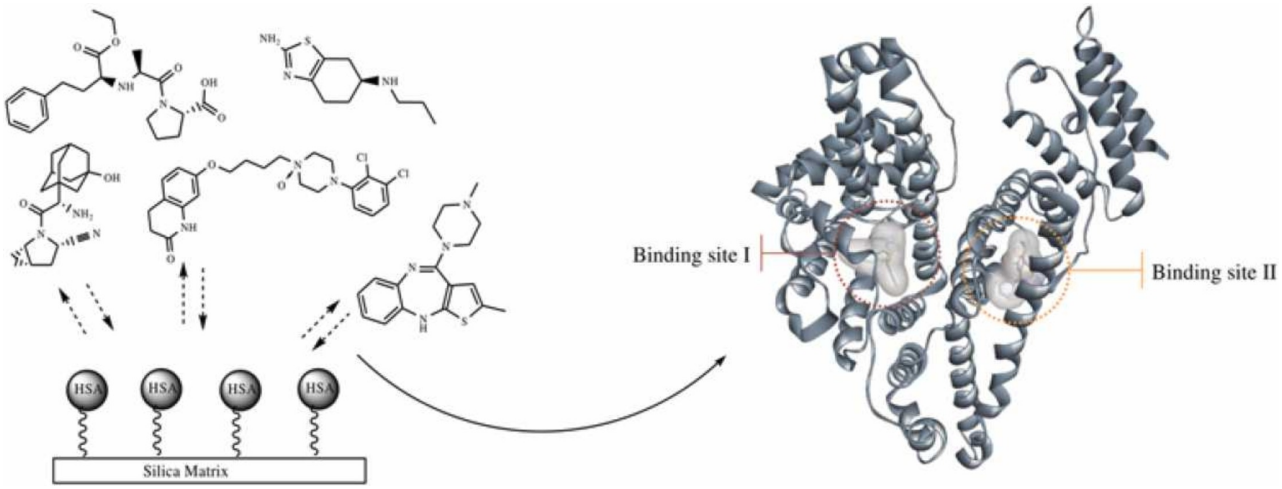

Figure 1.1: Drug-Target Interaction

Graph Neural Networks (GNNs) are one of the computational methods that have drawn the most attention due to its capacity to describe chemical structures as graphs. Such representations allow the model to reflect complex interactions within drug compounds, with nodes representing atoms and edges representing bonds[4]. Comparably, Transformer-based architectures—which were first created for natural language processing—have

proven to be remarkably adept at deriving structural and sequential patterns from protein sequences. By obtaining significant features from both drug compounds and protein targets, GNNs and Transformers work together to create a strong foundation for DTI prediction[5].

Graph-based learning and transformer-based protein embeddings are used in this project to create a computational model that can accurately predict DTIs. The suggested method seeks to reduce the need for expensive clinical studies while speeding up the development of efficient medications for cardiovascular conditions. This paper advances the field of computational drug discovery by tackling the issues of cost, efficiency, and scalability.

1.2 PROBLEM STATEMENT

Graph-based learning and transformer-based protein embeddings are used in this study to create a computational model that can accurately predict DTIs. The suggested method seeks to reduce the need for expensive clinical studies while speeding up the development of efficient medications for cardiovascular conditions. This paper advances the field of computational drug discovery by tackling the issues of cost, efficiency, and scalability.

Targeted therapies that are adapted to particular drug-protein interactions are necessary due to the complex nature of CVDs, which are marked by several biochemical pathways and patient-specific responses. However, the accuracy of typical computer tools in anticipating these interactions is limited. The complex molecular structures of medicinal compounds and the sequential dependencies found in protein sequences are frequently missed by current approaches like matrix factorization and kernel-based models.

Furthermore, the predictive accuracy and generalizability of current

deep learning-based models are limited by issues including imbalanced datasets, static edge characteristics in molecular graphs, and an inability to interpret the models. It is challenging to find trustworthy medication options for the efficient treatment of cardiovascular disorders because of these restrictions.

An improved computational framework that can accurately describe intricate drug-target interactions is desperately needed to overcome these issues. A more thorough comprehension of the connections between medicinal ingredients and protein targets would be possible with a model that could combine transformer-based protein embeddings with graph-based chemical representations. Such a system can get around the drawbacks of conventional methods by utilizing strategies like dynamic edge characteristics and attention processes. The ultimate goal of this effort is to improve public health outcomes by speeding up the drug development process, lowering the need for drawn-out clinical trials, and providing tailored therapeutic recommendations, especially for cardiovascular illnesses.

1.3 OBJECTIVE

This project's main goal is to create a Drug-Target Interaction (DTI) Prediction Model utilizing a Graph Transformer Network that uses ProtBert embeddings for protein sequences and dynamic edge computation for molecular graphs. The project's goal is to use cutting-edge computational methods to increase the precision and effectiveness of DTI predictions, particularly when it comes to treating cardiovascular illness.

In order to achieve this, we incorporate dynamic edge computation into a Graph Transformer Network. By using this method, the model may handle drug molecules that are represented as molecular graphs, in which bonds are represented by edges and atoms by nodes. During training, the model may

adaptively understand the interactions between atoms and bonds thanks to the addition of dynamic edge characteristics. In order to enhance the feature set and provide a thorough depiction of pharmacological molecules, the project also includes chemical characteristics including hydrophobicity, molecular weight, and topological polar surface area.

The second goal is to efficiently represent protein sequences using ProtBert embeddings. A transformer-based model that has already been trained, ProtBert, is able to extract significant structural and sequential information from amino acid sequences. The biological characteristics of the target proteins are captured by the model thanks to these embeddings, which is essential for precise interaction predictions.

Calculating interaction scores, which measure the binding affinity of medications and target proteins, is another goal of the study. The model calculates interaction scores that represent the probability of a successful interaction by fusing the protein sequence embeddings with the dynamic edge properties of the drug graphs.

The project's final goal is to assess and compare the model's performance. The predicted accuracy of the model is evaluated and trained using KIBA, a common benchmark dataset. The performance and dependability of the predictions are evaluated using measures like Mean Squared Error (MSE) and Mean Absolute Error (MAE).

In summary, the project's goal is to combine ProtBert embeddings for proteins and dynamic edge-based graph learning for medicines with chemical descriptors into a single framework for DTI prediction. This study enhances computational drug development, especially in the treatment of cardiovascular illnesses, by increasing predictive efficiency and accuracy.

1.4 ORGANIZATION OF REPORT

This report is organized as follows: Chapter 2 provides a literature review, discussing existing work on drug-target interaction (DTI) prediction, with a particular focus on graph-based methods and transformer architectures. Chapter 3 describes the system design, outlining the architectural framework of the proposed Graph Transformer-based Personalized Drug Recommendation System, including its key components and methodologies. Chapter 4 covers the implementation process in detail, including dataset preprocessing, model development, training, evaluation procedures, and results, followed by Chapter 5 with a conclusions and future work.

CHAPTER 2

LITERATURE SURVEY

Drug-target interaction (DTI) prediction plays a pivotal role in drug discovery, personalized medicine, and the identification of effective treatments. Traditional experimental methods are resource-intensive, making computational approaches essential. Graph-based deep learning models have become popular for this task due to their ability to represent molecular structures and protein sequences effectively. This section reviews significant studies that have contributed to advancements in DTI prediction.

2.1 GRAPH-BASED METHODS

Graph-based methods have emerged as powerful tools for predicting Drug-Target Interactions (DTIs) by modeling the intricate structures of drug molecules. Peng et al. (2023) introduced BINDTI, which utilizes Graph Convolutional Networks (GCNs) to encode the molecular structures of drugs, where nodes represent atoms and edges represent bonds. This model employs a bi-directional intention network combined with attention mechanisms, effectively integrating drug graph features with protein sequences for efficient interaction prediction [6]. Gao et al. (2024) proposed GraphormerDTI, which leverages Graph Transformer neural networks to encode molecular structures through Transformer-based message passing, achieving superior DTI prediction by integrating drug and protein features [7]. AttentionSiteDTI by Du et al. (2024) identifies effective protein binding sites, representing drug-protein complexes as structured sentences, enhancing interaction modeling and prediction accuracy [8].

2.2 SEQUENCE-BASED METHODS

Sequence-based methods focus on capturing the sequential nature of protein and drug information. Li et al. (2024) introduced DeepDTA, which uses two 1D Convolutional Neural Networks (CNNs) to encode drug molecules and protein sequences independently, then concatenates these features for DTI prediction, highlighting CNNs' efficacy in sequential data handling [9]. Zhang et al. (2024) demonstrated the effectiveness of Transformers in extracting both structural and sequential patterns from protein sequences, showcasing their relevance for long-sequence modeling in DTI tasks [10]. Dehghan et al. (2024) proposed a hybrid RNN-CNN framework that combines the strengths of RNNs for sequential learning and CNNs for feature extraction, achieving a more robust framework for DTI prediction [11].

2.3 CONTRASTIVE LEARNING-BASED METHODS

Contrastive learning methods have recently gained prominence in DTI prediction. Dehghan et al. (2024) introduced CCL-DTI, which employs contrastive loss functions like triplet loss and NT-Xent loss to enhance feature discrimination, integrating multimodal features including protein sequences, molecular structures, and interaction networks [12]. Monteiro et al. (2024) proposed a co-contrastive learning approach that uses inhomogeneous graph representations to refine feature learning, thereby extracting discriminative characteristics for drug-target pairs [13]. These methods bolster the robustness of DTI models by reducing feature redundancy.

2.4 NETWORK-BASED METHODS

Network-based methods model relationships between drugs and targets using graph-theoretic approaches. Li et al. (2024) highlighted the use

of random walk algorithms and matrix factorization to compute similarities between drugs and proteins, leveraging probabilistic methods for interaction prediction [9]. Monteiro et al. (2024) extended this with heterogeneous information networks, incorporating drug-drug and protein-protein interactions to enhance the predictive framework [13]. Such models excel in capturing large-scale biological relationships and improving our understanding of drug-protein interactions.

2.5 HYBRID METHODS

Hybrid approaches synergize multiple methodologies to leverage their unique strengths. Peng et al. (2023) described attention-based fusion models that integrate graph-based molecular representations with protein features using attention mechanisms, enabling precise alignment of drug structures and protein sequences [6]. Qiu et al. (2024) elaborated on multimodal deep learning frameworks that merge molecular graphs, protein sequences, and biological interaction networks, utilizing attention mechanisms to enhance predictive accuracy [12]. These hybrid models underline the potential for versatile frameworks that address the limitations of single-method approaches.

CHAPTER 3

SYSTEM DESIGN

This chapter focuses on the system design of the project. Graph-based learning for drug compounds and transformer-based embeddings for protein sequences are integrated in the system architecture to create a computational framework for Drug-Target Interaction (DTI) prediction. In order to capture bond characteristics, dynamic edge weights are added to drug SMILES strings before they are converted into molecular graphs using RDKit. Concurrently, ProtBERT is used to embed protein sequences in order to extract sequential and structural properties. The computation of interaction scores is

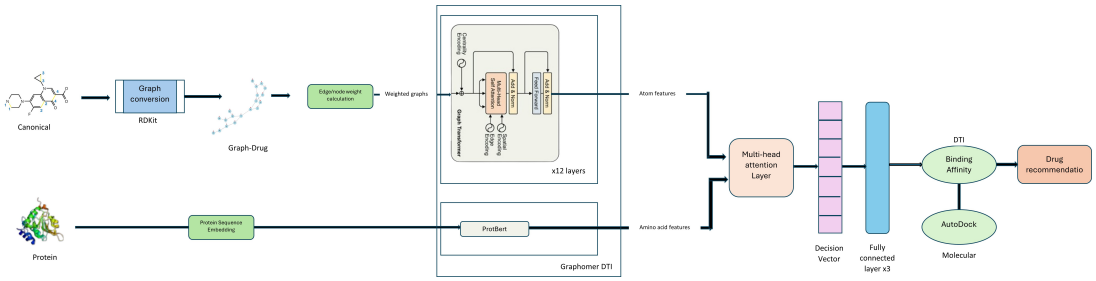

Figure 3.1: System Design Architecture for Drug-Target Interaction Prediction

made possible by the alignment of drug and protein features through the use of a Graph Transformer Network and a multi-head attention mechanism. The output layer makes predictions about the binding affinity, which are then confirmed using AutoDock molecular docking simulations.This system accelerates the drug discovery process by efficiently predicting DTIs, offering a robust and scalable solution for identifying potential drug candidates, particularly for cardiovascular diseases. The following system design focuses on predicting DTIs by aligning drug and protein features. The system for Drug-Target

Interaction (DTI) prediction integrates drug graphs and protein embeddings through a multi-stage process. Figure 3.1 provides an overview of the complete architecture. Each component of the system is described in detail in the subsequent sections.

3.1 DATA DESCRIPTION

The dataset forms the foundation for the system design and development of this Drug-Target Interaction (DTI) prediction model. It encompasses various data types that represent the structural, sequential, and relational characteristics of drugs and targets, as well as their interactions.

3.1.1 KIBA Database

A specialized tool for drug discovery, the KIBA database helps researchers understand how pharmacological molecules interact with biological targets including proteins. It focuses on kinase inhibitors, a class of medication that prevents kinases from doing their job. Enzymes called kinases are in charge of giving proteins a phosphate group, which is necessary for a variety of biological processes, including communication, metabolism, and cell division. Because of their crucial function, aberrant kinase activity is frequently connected to illnesses including cancer and cardiovascular diseases (CVDs). Researchers can create medications that fix these defects and aid in the treatment of such illnesses by focusing on kinases.

KIBA stands for Kinase Inhibitor BioActivity and is a database that integrates and standardizes bioactivity data from multiple sources. Bioactivity data refers to the ability of a molecule (like a drug) to affect a specific target (like a protein or enzyme). This information helps scientists measure how strongly a drug interacts or binds with its target protein.

A number of essential elements make up the KIBA database, which facilitates the investigation of drug-target interactions (DTIs), with a special emphasis on protein kinases and kinase inhibitors. Drug-like substances called kinase inhibitors have been developed specifically to stop or reduce the action of kinase proteins. This suppression is essential for halting aberrant cell signaling pathways that cause cancer and heart disease, among other disorders. Conversely, the target proteins listed in the KIBA database are protein kinases. These kinases are crucial for controlling important cellular functions as metabolism, development, and division. However, they are important targets for therapeutic interventions since their dysregulation can result in the development of illnesses.

The KIBA database uses bioactivity scores to gauge how well kinase inhibitors work against protein kinases. The degree to which a drug binds to and inhibits its target protein is measured by its bioactivity. Multiple experimental measures, including IC_{50}, K_i, and K_d, are integrated to obtain the scores.

$$\text{KIBA Score} = \log\left(\frac{K_i + IC_{50} + K_d}{3}\right), \tag{3.1}$$

From Equation 3.1, the kiba score is calculated. Here, IC_{50} denotes the drug concentration needed to block 50% of the target's activity, whereas K_i quantifies the drug's binding affinity to the target, or how firmly the drug binds. The dissociation constant, K_d, on the other hand, shows how readily a drug unbinds from its target protein.

3.2 DRUG REPRESENTATION

Accurately depicting drug molecules is crucial for modeling their characteristics and forecasting how they will interact with protein targets in computational drug discovery. A number of steps are involved in the drug

representation process, which makes it possible to convert chemical structures into a format that machine learning models can use.

3.2.1 Canonical SMILES

A text-based technique called the Simplified Molecular Input Line Entry System (SMILES) is used to concisely and uniformly depict the chemical structure of molecules. **SMILES** makes it simpler to store, process, and evaluate chemical information computationally by encoding atoms, bonds, and molecular connections as a string of text, in contrast to conventional visual diagrams. A canonical SMILES is a special and standardized form of a SMILES string that guarantees that, regardless matter how its structure is input or drawn, the same molecule always yields a single, consistent representation.

SMILES treats chemical structures as graphs, with bonds between atoms serving as edges and atoms as nodes. Chemical symbols, such as C for carbon and O for oxygen, are used to represent each atom, while specialized symbols, such as $-$ for single bonds, $=$ for double bonds, and # for triple bonds, are used to denote bonds. In molecules, branches are denoted by parenthesis, and ring systems are identified by numbers that represent the relationships between atoms. Chemists and computational tools can effectively interpret chemical structures thanks to this format.

Take the basic yet essential aromatic chemical benzene, for instance. Six carbon atoms are grouped in a ring to produce benzoene, which has a planar, hexagonal structure formed by alternating double bonds. Its canonical SMILES symbol is c1ccccc1, where the numerals 1 signify the beginning and end of the ring structure and the lowercase c indicates aromatic carbons. This notation guarantees that the aromaticity of the molecule is preserved while succinctly expressing the cyclic character of benzene.

The traditional SMILES representation of a more complicated molecule, such as clenbuterol(Equation 3.2), which has several rings, functional groups, and chlorine atoms, is as follows:

$$CC(C)(C)NCC(C_1=CC(=C(C(=C_1)Cl)N)Cl)O, \tag{3.2}$$

Here: (I) CO represents a methoxy group attached to the molecule, (II) c1cc2c(cc1Cl) describes a chlorinated aromatic ring system, (III) C(c1ccc(Cl)c(Cl)c1)=NCC2 represents additional chlorinated benzene rings and a heterocyclic nitrogen-containing structure.

Canonical SMILES are significant because they remove ambiguity and offer consistency. The same molecule may be drawn differently by many users or software programs, but canonicalization makes guarantee that all representations are combined into a common string format. For cheminformatics programs like RDKit and databases like PubChem, which employ SMILES to store and handle enormous volumes of molecular data, this consistency is especially helpful. Canonical SMILES provides a standardized input format for machine learning models in drug development, allowing chemical structures to be transformed into molecular graphs or numerical representations for predictive analysis.

SMILES strings are frequently transformed into molecular graphs, in which bonds are seen as edges and atoms as nodes, for use in computational drug discovery. Advanced models such as Graph Neural Networks (GNNs), which forecast interactions between medicinal compounds and target proteins, use these graphs as input. Researchers can compute molecular characteristics like molecular weight or hydrophobicity, analyze big datasets rapidly, and conduct virtual screening to find promising drug candidates by integrating canonical SMILES into the workflow.

In conclusion, canonical SMILES offers a distinct and condensed representation of molecules, making it a vital tool in cheminformatics and computational chemistry. It makes managing chemical structures easier and forms the foundation of data-driven approaches in drug discovery, where precision and reliability are essential for success.

3.2.2 RD-Kit Module

RDKit is an open-source cheminformatics library that provides robust tools for manipulating, analyzing, and transforming chemical molecules. It is important in computational chemistry, especially in drug development, where predictive models and machine learning depend on chemical representations. SMILES (Simplified Molecular Input Line Entry System) and other textual representations of molecules can be transformed into graph-based representations using RDKit. Atoms are represented as nodes in these graphs, and chemical bonds—such as single, double, or aromatic bonds—are represented by edges. Take the medication clenbuterol, which is used to treat respiratory conditions like asthma. The interactions between its carbon, nitrogen, oxygen, and chlorine atoms are captured by RDKit's efficient conversion of its conventional SMILES string, CC(C)(C)NCC(C1=CC(=C(C(=C1)Cl)N)Cl)O, into a molecular graph.

In addition to converting graphs, RDKit calculates molecular descriptors, which are numerical values that include chemical characteristics crucial for predictive modeling. These characteristics include the number of rotatable bonds, molecular weight, and topological polar surface area (TPSA). For example, RDKit determines a TPSA value of 58.4$Å^2$ and a molecular weight of approximately 277.1 g/mol for clenbuterol. The polarity of the molecule, which affects its capacity to cross cell membranes, a crucial characteristic for drug absorption and bioavailability, is indicated by TPSA. A crucial factor in

determining a molecule's ability to bind to target proteins is its flexibility, which is reflected in the number of rotatable bonds that RDKit calculates.

Additionally, RDKit produces features at the atom and bond levels, which offer more profound understanding of the structure of the molecule. Atom-level characteristics of clenbuterol include atomic charges, hybridization states (sp3 and sp2), and atom types (carbon, nitrogen, chlorine, and oxygen). Stereochemistry and bond types (single and double bonds) are captured by bond-level characteristics, which collectively characterize the spatial organization and chemical connectivity of the molecule. For machine learning algorithms to correctly forecast characteristics such as binding affinity, toxicity, and solubility, these characteristics are essential.

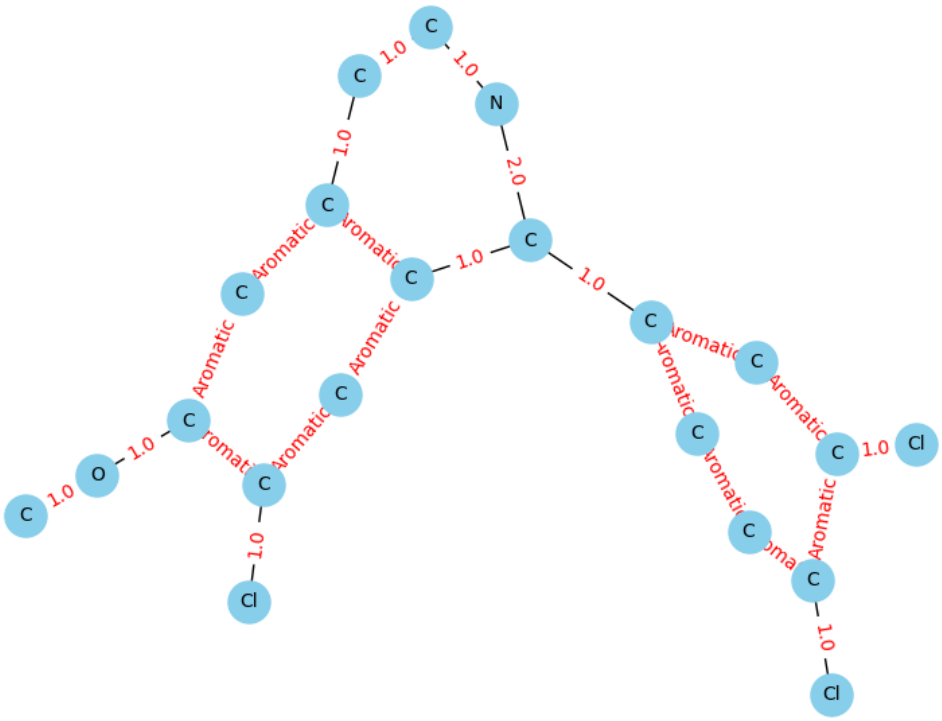

Figure 3.2: Graph Representation of Clenbuterol

The importance of RDKit is found in its capacity to connect

graph-based representations needed for machine learning applications with textual representations (like SMILES). Clenbuterol's SMILES string is transformed into a graph and enhanced with features such as connection information and molecular descriptors using RDKit as shown in Figure 3.2.

In summary, RDKit offers a smooth process for the conversion and analysis of chemical compounds such as clenbuterol. It makes it possible for cheminformatics and machine learning technologies to precisely model molecular behavior by effectively extracting features and descriptors. Large-scale molecular datasets can be processed rapidly because of RDKit's integration with Python-based libraries, which makes it a vital tool for cheminformatics and computational drug discovery.

3.2.3 Molecular Graph Representation

Molecular Graph Representation plays a pivotal role in computational chemistry and drug discovery by enabling the structured representation of drug molecules as graphs.In this model, the nodes correspond to atoms such as carbon, hydrogen, oxygen, nitrogen, or chlorine, while the edges denote the chemical bonds joining these atoms. The types of bonds—single, double, triple, or aromatic—reflect the various atomic interaction topologies and intensities. Molecular graphs, in contrast to linear textual representations such as SMILES, preserve the spatial and topological structure of molecules while capturing their complex interconnection and interactions. For applications like property prediction and Drug-Target Interaction (DTI) modeling, this graph-based method enables machine learning models—in particular, Graph Neural Networks (GNNs)—to more accurately interpret and analyze molecular activity.

The dynamic edge properties constitute a major breakthrough in

molecular graph representations. Dynamic edge features enable bond qualities, including bond type, bond length, and bond angles, to change over time, in contrast to standard graph-based models that employ static bond properties (such as bond type). Because of this flexibility, the model can iteratively improve its comprehension of the molecular structure depending on the training situation. The model can acquire more accurate and expressive representations if, for instance, the importance of a certain bond inside a molecule changes based on how it interacts with other atoms. Therefore, dynamic edge features improve graph-based models' ability to effectively represent intricate chemical interactions.

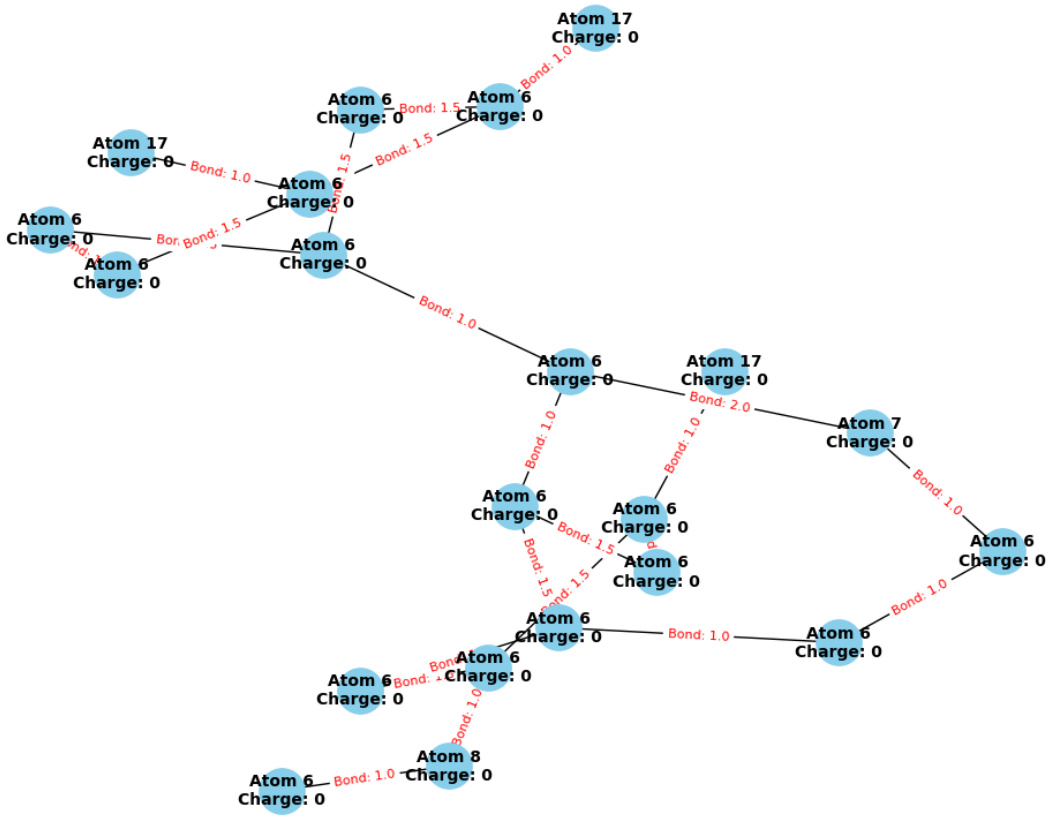

Figure 3.3: Molecular Graph Representation of Clenbuterol

Graphs are crucial for molecular modeling because they can retain the spatial and connectivity details of atoms and links. As seen in Figure 3.3, the

exact arrangement of atoms and bonds can be seen. The arrangement of atoms and bonds within a molecule greatly influences chemical qualities including binding affinity, solubility, and reactivity. For example, a molecule's interaction with target proteins is greatly influenced by the type and location of its functional groups as well as the flexibility of the molecule, which is established via rotatable bonds. Models can accurately capture these characteristics by depicting molecules as graphs, which makes them perfect for applications like forecasting biological activity, toxicity levels, or drug-protein interactions.

Furthermore, graph-based deep learning techniques that can take advantage of the wealth of information contained in the graph structure complement molecular graph representations. By combining data from nearby nodes and edges, Graph Neural Networks (GNNs) interpret molecular networks and enable the model to understand both local and global relationships inside the molecule. For instance, the way a chemical attaches to a protein pocket may be affected by interactions between particular atoms, and GNNs are capable of efficiently capturing these dependencies.

To summarize, drug molecules can be represented as graphs, which offers a strong foundation for simulating molecular structures and how they interact with biological targets. Graph-based models can improve their predictive accuracy by adaptively learning bond characteristics and honing chemical representations through the use of dynamic edge features. In addition to preserving the topological and spatial information essential for comprehending molecular behavior, this method uses cutting-edge deep learning algorithms to identify connections between drug compounds and proteins, leading to advances in computational drug development.

3.2.4 Molecular Descriptors

Molecular descriptors are numerical numbers that give a quantitative picture of a molecule's structural characteristics and chemical characteristics. Because they allow machine learning models to derive valuable insights into the molecular behavior and interaction potential of drug candidates, these descriptors are crucial in cheminformatics and computational drug discovery. Molecular descriptors increase the predictive power of the model for detecting Drug-Target Interactions (DTIs) by storing important physicochemical and structural properties of molecules.

Physicochemical qualities and structural traits are the two main categories of molecular descriptors. The quantifiable chemical characteristics of molecules, which are crucial in defining how they behave in biological systems, are described by their physicochemical properties. For instance, molecular weight (MW), which is determined by Equation 3.3, is the sum of the masses of all the atoms in a molecule.

$$\mathrm{MW} = \sum_{i=1}^{n} (n_i \times M_i), \tag{3.3}$$

where n_i is the number of atoms of type i, and M_i is the atomic weight of atom i. MW influences a molecule's bioavailability, since smaller molecules are often more easily absorbed in biological systems.

LogP (Equation 3.4) is another important metric that assesses the hydrophobicity or lipophilicity of the molecule. The logarithm of the partition coefficient, or logP, is a measure of how a molecule is distributed in an environment that is hydrophilic (water) and hydrophobic (octanol). The

mathematical expression for it is as follows:

$$\text{nLogP} = \log\left(\frac{C_{\text{octanol}}}{C_{\text{water}}}\right), \tag{3.4}$$

where *nLogP* represents the logarithm of the partition coefficient, C_{ocatanol} is the concentration of the compound in octanol and C_{water} is the concentration of the compound in water.

Another physicochemical metric called Topological Polar Surface Area (TPSA) quantifies the surface area of a molecule that contains polar atoms like nitrogen and oxygen together with the hydrogen atoms that are bonded to them. TPSA is used to measure a molecule's capacity to penetrate cell membranes and is notably relevant in medication absorption research.

Apart from physicochemical characteristics, structural factors are also included in molecular descriptors. These characteristics offer information about the flexibility and molecular structure, both of which are essential for drug-target interactions. For example, the number of lone pairs on electronegative atoms (acceptors) and hydrogen atoms attached to electronegative atoms (donors) are used to determine the number of hydrogen bond donors and acceptors. These characteristics are crucial because hydrogen bonds are key to stabilizing molecular interactions with proteins.

Another crucial structural characteristic that measures the molecule's flexibility is the quantity of rotatable bonds. Rotatable bonds refer to single non-ring bonds between atoms, omitting terminal bonds such as methyl groups. Multiple conformations that flexible molecules can take can affect how they attach to proteins. Lastly, the number of aromatic rings indicates if ring systems with conjugated double bonds are present, which enhances the stability and binding affinity of molecules.

Molecular descriptor calculations are effectively carried out with the cheminformatics tool RDKit. RDKit calculates descriptors such as MW, LogP, TPSA, hydrogen bond counts, and rotatable bonds after processing the molecular structure, here from SMILES. In order to provide more context than the graph-based representation of pharmacological molecules, these descriptors are standardized and normalized before being used as input features in machine learning models.

Molecular descriptors are important because they can improve the model's comprehension of medicinal compounds. Descriptors offer further information on the physical characteristics and structural elements of the molecule, whereas molecular graphs depict the spatial arrangement and connectivity of atoms. In tasks like drug-target interaction, bioactivity prediction, and molecular property optimization, this combination helps machine learning models to produce more precise and nuanced predictions, which eventually leads to more effective drug discovery procedures.

3.2.5 Graph Transformer for Drug Molecules

Graph Transformers are used to analyze and represent the chemical structures of pharmacological compounds once molecular graphs and molecular descriptors have been created. Graph Transformers use attention techniques to efficiently simulate both local and long-range dependencies in molecular graphs, in contrast to conventional Graph Neural Networks (GNNs), which rely on confined neighborhood aggregation. This is especially important for precisely describing the links and atom-to-bond connections seen in intricate drug structures.

Each atom in a molecule is represented as a node in Graph Transformers, with bonds between the atoms represented as edges. Dynamic

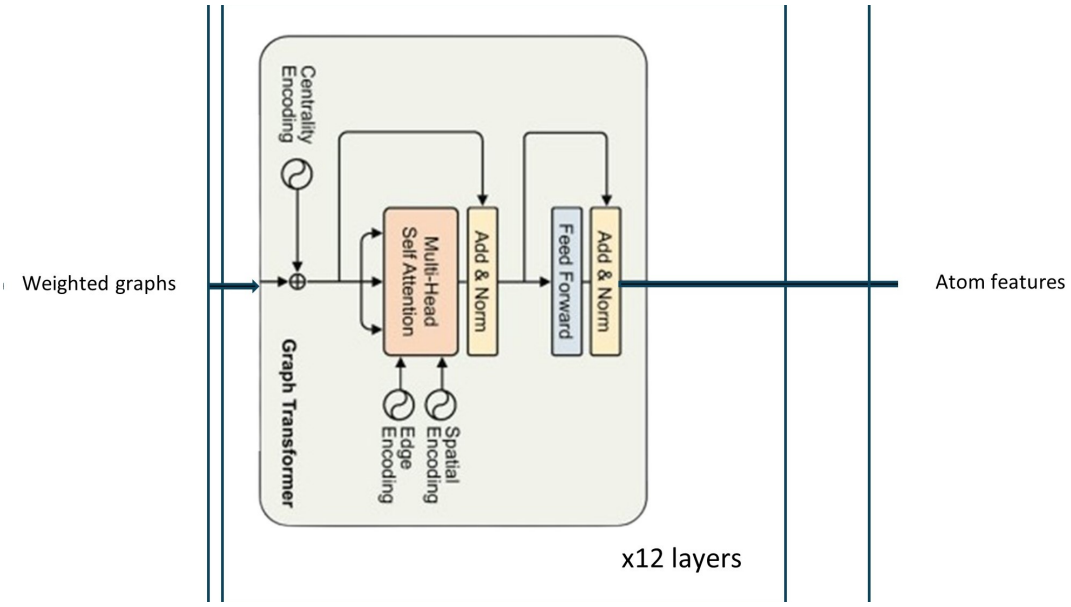

Figure 3.4: Graph Transformer Architecture

edge characteristics are added to the graph to improve the representation. Bond types, bond angles, and other molecular descriptors including molecular weight, hydrophobicity (LogP), and topological polar surface area (TPSA) are examples of these properties, which change over training. The model's capacity to encode molecular connections and attributes is enhanced by its dynamic nature, which enables it to adaptively modify the edge representations.

To integrate molecular descriptors with graph-based representations, Graph Transformers as seen in Figure 3.4 processes concatenated node, edge, and molecular-level features. For example, in the case of Clenbuterol, a drug molecule represented by equation 3.2, the molecular graph as in Figure 3.3 consists of nodes representing carbon, nitrogen, and oxygen atoms, while edges represent the bonds between these atoms. The dynamic edge features incorporate descriptors such as bond types and LogP values, allowing the Graph Transformer to learn more expressive and meaningful representations.

The self-attention mechanism, which allows the model to concentrate on the most important atoms and bonds in the graph, is one of Graph

Transformers' unique features. In particular, each node's query (Q), key (K), and value (V) matrices are computed by the attention mechanism and utilized to determine attention scores by the following equation:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q \cdot K^{\top}}{\sqrt{d}}\right)V \tag{3.5}$$

where Q, K, V are linear transformations of the node features, and d is the dimension of the embedding.

The model may weigh the contributions of nearby nodes and edges based on these scores (Equation 3.5), which establish their relative relevance. The self-attention technique allows Graph Transformers to capture long-range dependencies over the whole molecular graph, in contrast to GNNs that use message forwarding to spread information locally. Furthermore, a deeper picture of the molecule is provided by the simultaneous modeling of several interactions using multi-head attention.

The Graph Transformer produces a graph-level embedding as its output, encoding the structure and chemical characteristics of the whole molecule. The revised node embeddings are aggregated using a pooling procedure, like global mean pooling or adaptive average pooling, to produce this embedding. For tasks like predicting Drug-Target Interactions (DTIs), the resultant feature vector is used as the input.

When it comes to drug representation, Graph Transformers have a number of benefits over conventional GNNs. First, they can represent both local and global interactions thanks to the self-attention mechanism, which makes it possible to comprehend the molecular network more comprehensively. Second, the representation of the molecule is enhanced by the use of molecular descriptors and dynamic edge features, which capture both structural and

chemical information. Third, the model's predictive performance is enhanced by the simultaneous learning of several atom-bond interactions made possible by the use of multi-head attention.

Overall, Graph Transformers offer a strong and adaptable framework for drug molecule modeling. They make predictions for tasks like Drug-Target Interaction (DTI) prediction more accurate by fusing attention processes with graph-based representations. This method is especially helpful for drug development, as discovering possible treatment options requires a knowledge of the complex interactions inside molecular structures.

3.3 PROTEIN REPRESENTATION

In this project, the amino acid sequences were transformed into numerical tensors using ProtBERT to create high-dimensional embeddings for protein sequences. Large-scale protein sequence databases were used to pre-train the transformer-based model known as ProtBERT [14]. Its purpose is to extract contextual and meaningful characteristics from protein sequences. A condensed but extremely informative representation of the sequence is given by the resultant embeddings.

ProtBERT, a transformer-based language model particularly trained on biological sequences, to translate protein sequences, which are represented as a string of amino acids, into high-dimensional numerical embeddings. Using a multi-layer transformer architecture, ProtBERT analyzes these sequences and extracts significant patterns from the protein data.

Protein sequences are typically represented as a series of amino acids, such as:

$$MKTIIALSYIFCLVFADYKDDD$$

Here, each letter represents a single amino acid (e.g., M = Methionine, K = Lysine). In natural language processing (NLP), where each amino acid is regarded as a token, these sequences are comparable to words. These sequences are tokenized by ProtBERT, which then processes them to extract valuable representations.

ProtBERT applies the BERT architecture, consisting of multi-head attention and feedforward layers, to encode protein sequences and produce embedding of dimensions 1024. These embeddings are then pooled (averaged or summed) to generate a single high-dimensional tensor that represents the entire protein sequence.

Once the protein embeddings are generated, they are combined with drug representations, such as molecular graphs, to calculate an interaction score. The embeddings enable the model to align the complex biochemical features of the protein with structural features of the drug molecule.

The prediction potential of DTI models is increased when proteins can be represented as high-quality embeddings. ProtBERT gives the model a better knowledge of protein properties by storing contextual, structural, and sequential information. The drug discovery process is greatly accelerated when these embeddings are used in conjunction with graph-based drug representations to allow the model to calculate interaction scores that forecast the binding affinity between a drug and a target protein.

3.4 DRUG-TARGET ITERACTION

In this project, drug graph embeddings and protein sequence embeddings are aligned using a Graph Transformer Network with a Multi-Head Attention (MHA) method (Equation 3.6) to estimate the binding affinity or

KIBA score. To precisely calculate the interaction score, this integration combines the sequential characteristics of the protein with structural data from the drug graph.

The drug graph embeddings and protein sequence embeddings are aligned using the Multi-Head Attention (MHA) mechanism in the Graph Transformer. The MHA mechanism learns to focus on the most relevant interactions between the drug and protein features. Drug graph embeddings h are processed through linear transformations to generate queries (Q) and Protein sequence embeddings p are similarly processed to generate keys $(K)$and values (V).

The attention mechanism, *equation 3.5* computes attention scores, which determine the degree of interaction between drug and protein features [7]. Here, Q is Queries derived from drug embeddings, K is Keys derived from protein embeddings, V is Values derived from protein embeddings and d is Dimensionality of the embeddings (e.g., 1024 for ProtBERT).

Multi-Head Attention splits the embeddings into multiple heads to learn diverse relationships between drug and protein features.The multi-head attention can be expressed as:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{3.6}$$

where:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.7}$$

W_i^Q, W_i^K, W_i^V : Linear transformations for each attention head.
W^O : Final projection to combine outputs of all heads.

Each attention head focuses on different parts of the drug and protein

features, enabling the model to identify subtle patterns and complex interactions.

The output of the multi-headed attention mechanism (MHA) is an aligned representation of the protein sequence embeddings and drug graph characteristics. The complex connections between the drug's molecular structure and the protein's sequential characteristics are captured by this aligned representation. The representation is run through a number of linear layers that project the high-dimensional aligned features into a single scalar value in order to get the final binding affinity score. A prediction layer calculates this scalar value and outputs the binding affinity or KIBA score. The drug's ability to bind to a particular protein is shown by the KIBA score, which quantitatively represents the strength of the interaction between the medication and the protein target. This interaction score offers a reliable way to assess drug-target interactions by utilizing the correlations that have been learnt between drug molecular graphs and protein embeddings. This allows for predictions that are essential for drug discovery.

CHAPTER 4

IMPLEMENTATION AND RESULTS

The Drug-Target Interaction (DTI) prediction framework's technical workflow, tools, and methodologies are explained in depth in this section. This section presents the findings from the evaluation process while striking a balance between conceptual comprehension and real-world application. It guarantees a thorough grasp of the modular system by including crucial elements like pseudocode, important computations, implementation specifics, and performance indicators. Data preprocessing, model creation, training pipeline, and evaluation techniques are all included in the modular approach, which also presents quantitative findings to demonstrate the framework's efficacy.

4.1 DATA PREPROCESSING AND FEATURE REPRESENTATION

The dataset forms the backbone of the system design, encompassing multiple modalities to model Drug-Target Interactions (DTIs). Proper preprocessing and feature representation ensure the data's quality and relevance for predictive modeling.

4.1.1 Drug Representation: Graph Construction

A number of methodical procedures are carried out to convert chemical data into graph-based characteristics and numerical descriptors in order to accurately depict drug molecules for computational modeling. With the help of these procedures, comprehensive molecular data can be extracted and fed

into machine learning models, like the suggested Drug-Target Interaction (DTI) prediction framework. Here we use RDkit for extraction of detailed molecular information.

4.1.1.1 SMILES Conversion

Drug compounds' chemical structures are described using Simplified Molecular Input Line Entry System (SMILES) strings as the input format. A concise, text-based depiction of molecules is offered via SMILES strings. Ethanol, for instance, is written as CCO, where "C" stands for carbon atoms and "O" for oxygen, with the atoms implicitly joined by single bonds. The representation of a more complicated molecule, like a chlorinated derivative, may be COc1cc2c(cc1Cl)... RDKit is used to transform these textual representations into a format that can be processed using graphs.RDKit translates SMILES into molecular graphs, $G(V,E,F_v,F_e)$, where, Nodes(V) correspond to atoms (e.g., Carbon, Oxygen, Nitrogen) and Edges(E) represent chemical bonds (e.g., single, double, aromatic bonds) connecting the atoms, F_v represents the node features, describing atomic properties such as atom types, charges, and hybridization states and F_e represents the edge features, describing bond properties such as bond type, aromaticity, and bond direction.

The basis for graph-based learning is the graph representation, which maintains the spatial relationships and connectivity of atoms. This stage guarantees that molecules' structural characteristics are preserved for subsequent feature extraction.

4.1.1.2 Molecular Descriptors

Molecular descriptors are numerical depictions of a molecule's structural characteristics and chemical makeup. These descriptors improve the predictive power of machine learning models for Drug-Target Interaction (DTI) tasks and offer supplementary data to graph-based representations. Molecular descriptors assist the model better understand the distinct qualities of medicinal compounds by capturing both structural and physicochemical features.In this project, we extracted ten key molecular descriptors, each serving a distinct purpose in understanding drug molecules.

A molecule's molecular weight (MW), which is determined by adding the atomic weights of its constituent atoms, is a representation of its overall mass, as seen in Equation 4.1. This characteristic has a major impact on the permeability, solubility, and bioavailability of drugs.

$$\mathrm{MW} = \sum_{i=1}^{n} (n_i \times M_i) \tag{4.1}$$

where n_i is the number of atoms of type i, and M_i is the atomic weight of atom i.

The partition coefficient (equation 4.2 between octanol and water is measured by the LogP (Hydrophobicity), which indicates whether a molecule is hydrophobic or lipophilic. This characteristic is vital for determining permeability and solubility, which are necessary for drug distribution and absorption.

$$\mathrm{nLogP} = \log\left(\frac{C_{\mathrm{octanol}}}{C_{\mathrm{water}}}\right) \tag{4.2}$$

where *nLogP* represents the logarithm of the partition coefficient, C_{ocatanol} is the

concentration of the compound in octanol and C_{water} is the concentration of the compound in water.

The surface area of a molecule that is occupied by polar atoms, including nitrogen and oxygen, is known as the Topological Polar Surface Area (TPSA). The ability of a molecule to pass through cell membranes, a crucial component of medication bioavailability, is directly tied to TPSA. The sum of the contributions from polar fragments is used to calculate TPSA, usually with cheminformatics tools such as RDKit.

The quantity of hydrogen atoms bonded to electronegative atoms (such as nitrogen or oxygen) that have the ability to make hydrogen bonds is known as hydrogen bond donors (HBD). Likewise, the number of electronegative atoms that can take hydrogen bonds is indicated by Hydrogen Bond Acceptors (HBA). When combined, these descriptions are essential for comprehending binding affinity and drug-target interactions.

The number of single, non-ring bonds that are attached to heavy atoms that are not terminal and permit free rotation is known as the Number of Rotatable Bonds. The flexibility of the molecule is indicated by this characteristic, which affects how well it conforms to target binding sites. The number of rings, which comprises the molecule's aromatic and non-aromatic rings, is another structural characteristic. In drug design, rings are essential scaffolds that affect biological activity and molecular stability. Specifically, the Number of Aromatic Rings counts rings that have aromaticity and conjugated π-electron systems. The stability and specificity of a molecule's binding to proteins are improved by aromaticity.

The percentage of sp3 hybridized carbon atoms in a molecule is shown by the Fraction of sp3 Hybridized Carbons (CSP3). Pharmacokinetics

depends on this descriptor's ability to reveal the molecule's 3D complexity and drug-likeness.

Finally, the Exact Molecular Weight uses the precise isotopic weights of atoms to determine the exact mass of the molecule. This descriptor provides a more precise molecular mass measurement, making it extremely relevant for pharmacokinetics and mass spectrometry.

These descriptors help the model bridge the gap between biological activity and molecular-level properties, improving its accuracy in predicting drug-target interactions.

4.1.2 Protein Represenation

Drug-Target Interaction (DTI) prediction focuses significantly on protein representation since it captures the structural and biochemical characteristics of protein sequences. The ProtBERT model, a transformer-based architecture created especially for biological sequences, is used in this research to process protein sequences, which are made up of amino acid chains. These sequences are represented numerically in a high-dimensional space by ProtBERT.

A protein sequence like "MKYLLPT" is tokenized and processed through several transformer layers by ProtBERT. A tensor representation is produced by mapping each location in the sequence to a high-dimensional vector as seen in Equation 4.3. ProtBERT generates embeddings of shape $(1, 1024)$ for this implementation, where the 1024-dimensional vector represents the extracted features and the single batch dimension corresponds to a single protein sequence.

$$\text{ProtBERT}(\text{Sequence}) \rightarrow \text{Tensor Shape: } (1, 1024) \tag{4.3}$$

These information-rich embeddings capture both structural characteristics (such inferred folding patterns) and sequential interactions (like the proximity and order of amino acids). As a result, the model can learn context-specific protein characteristics that are essential for forecasting interactions with medicinal compounds, such as active sites, binding areas, and general functionality.

4.2 MODEL DESIGN: GRAPH TRANSFORMER NETWORK

The foundation of this study is the Graph Transformer Network, which combines drug graph characteristics with protein embeddings to predict the binding affinity or KIBA score. We go into further detail about each essential element of the model architecture below:

4.2.1 Dynamic Graph Transformer Layer

The Dynamic Graph Transformer Layer is essential when processing drug molecular graphs. Its main objective is to use attention-based message forwarding to improve node embeddings (atom representations) and edge characteristics (bond representations) in the molecular graph. This guarantees that during training, the molecular characteristics and graph structure are precisely recorded. The Dynamic Graph Transformer Layer leverages multi-head attention to compute effective node representations by aligning features across molecular graphs and protein embeddings. The complete pseudocode is shown in Algorithm 4.1. This layer dynamically computes edge features and incorporates them into the message-passing mechanism, improving the model's ability to capture bond-specific interactions.

The Dynamic Graph Transformer Layer's message passing mechanism as seen in Figure 4.1 is what allows the model to communicate

Algorithm 4.1 Dynamic Graph Transformer Layer

Require: Node features x, Edge index $edge\_index$, Edge attributes $edge\_attr$, Molecular features $mol\_features$
Ensure: Updated node features out

Initialize Query Q, Key K, Value V transformations
Initialize Edge MLP and Feedforward Network
Forward Pass:
Compute Q, K, V using linear transformations on x
Reshape Q, K, V to `(num_heads, num_nodes, head_dim)`
Compute Attention Scores: $attention\_scores = \frac{Q \cdot K^T}{\sqrt{head\_dim}}$
Normalize Scores using `softmax`
Apply attention weights to V: $out = attention\_scores \cdot V$
Compute Dynamic Edge Features:
for each edge (src, dst) in $edge\_index$ **do**
    Concatenate $x[src]$, $x[dst]$, and $mol\_features$
    Pass concatenated features through Edge MLP
end for
Propagate messages using updated $edge\_attr$ and x
Combine x and edge information: $out = x + edge\_attr$
Pass out through Feedforward Network
return Updated node features out

across linked nodes in the molecular graph. Node embeddings, which represent atoms, are iteratively updated during message passing by combining data from their neighbors through edges, which represent bonds. By using this approach, the updated node features are guaranteed to reflect not only their own attributes but also those of dynamic edge features and nearby nodes.[15]

Message passing allows nodes in the molecular graph to

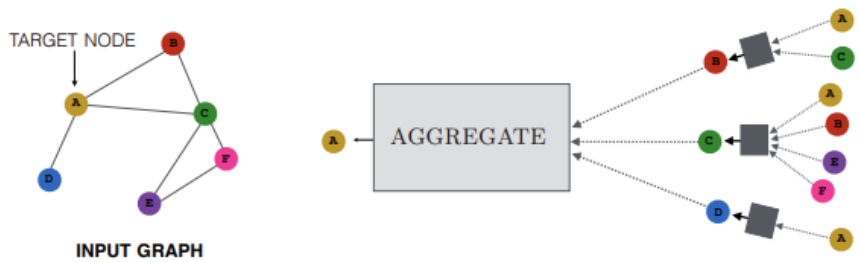

Figure 4.1: Message Passing

communicate with one another through edges, which stand in for bonds, under the framework of the Dynamic Graph Transformer. Information moves from the source node (i) to the destination node (j) for each edge (i,j). Edge factors that are processed during training, such as bond type, aromaticity, and bond-specific parameters, dynamically impact the information flow. A Multi-Layer Perceptron (Edge MLP) is used to improve these features, guaranteeing that bond-specific characteristics are integrated into the message-passing procedure.

The model uses a Multi-Head Attention method, in contrast to conventional GNNs, which aggregate surrounding node information through simple summation or averaging. By calculating attention scores between the source and target nodes, the attention mechanism enables nodes to preferentially focus on significant neighbors during message delivery. This process updates node embeddings $h_j^{(l+1)}$ at layer $l+1$ as:

$$h_j^{(l+1)} = \text{Attention}(Q_i, K_j, V_j) + \text{Edge Features} \tag{4.4}$$

where Q, K, and V are the query, key, and value matrices derived from node embeddings. To improve the representation of molecular structures, the Edge MLP makes sure that the edge features are dimensionally aligned with the node embeddings. To preserve dimensional consistency, the modified node embeddings undergo a linear transformation when message passing is finished. In this stage, the embeddings are ready for further tasks where the improved node and edge information increases prediction accuracy, including predicting chemical characteristics or binding affinity scores.

4.2.1.1 Dynamic Edge Features

Dynamic edge features, which are dynamically calculated during training to adjust the model to bond-specific information, capture bond-specific attributes in the chemical graph. (I) Bond type, that is, whether a bond is single, double, triple, or aromatic, is determined by its nature. This can be expressed mathematically as a categorical variable:

$$b_{type} \in \{single, double, triple, aromatic\} \tag{4.5}$$

where b_{type} is encoded as a one-hot vector during computation to ensure compatibility with neural networks. (II) Aromaticity affects the stability and reactivity of molecules by indicating whether a bond is a member of an aromatic ring. A boolean variable $b_{aromatic} \in \{0,1\}$ is used where:

$$b_{aromatic} = 1 \; if \; bond \; belongs \; to \; an \; aromatic \; ring, \; else \; 0 \tag{4.6}$$

(III) To represent the bond direction, we define the direction vector $\vec{d}_{\text{bond}}$ between two bonded atoms. If atom i is located at $\vec{r}_i$ and atom j is located at $\vec{r}_j$, the direction vector is given by:

$$\vec{d}_{\text{bond}} = \vec{r}_j - \vec{r}_i \tag{4.7}$$

Normalization is applied to scale the vector, resulting in the unit direction vector $\hat{d}_{\text{bond}}$:

$$\hat{d}_{\text{bond}} = \frac{\vec{d}_{\text{bond}}}{\|\vec{d}_{\text{bond}}\|} \tag{4.8}$$

Where:
$\vec{d}_{\text{bond}}$: The raw direction vector.
$\hat{d}_{\text{bond}}$: The normalized direction vector.

$\|\vec{d}_{\text{bond}}\|$: The magnitude (Euclidean norm) of $\vec{d}_{\text{bond}}$.

4.2.1.2 Multi-Head Attention Layer

The layer employs a multi-head attention mechanism, enabling the model to focus on multiple feature subsets simultaneously. Each attention head captures distinct aspects of the molecular graph, and their outputs are concatenated and linearly projected to produce a unified representation. Additionally, a fully connected layer projects the final embeddings to the desired dimensionality, ensuring compatibility with downstream tasks.

For each node feature$X$, the layer computes query (Q), key (K) and value (V) matrices through linear transformations as seen in equation 4.9.

$$Q = W_Q X,\ K = W_K X, V = W_V X \tag{4.9}$$

where W_Q, W_K, W_V are learnable weight matrices. These transformations allow the model to identify which nodes are relevant (Q), provide information (K) and hold associated data (V) for attention computation.

The attention score (equation 4.10 are calculated as the dot product of Q and K, scaled by the square root of the embedding dimension (d) to stabilize gradient, followed by the softmax function:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q \cdot K^\top}{\sqrt{d}}\right) V \tag{4.10}$$

The multi-head attention mechanism processes the same input in parallel using multiple attention heads. Each head captures distinct subsets of features. The

outputs from all heads are concatenated and linearly projected:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{4.11}$$

where Q, K, V are linear transformations of the node features, and d is the dimension of the embedding.

4.2.1.3 Edge Feature Refinement

The edge MLP (equation 4.12 dynamically initializes based on the shape of concatenated features from source and destination nodes and molecular descriptors.

$$EdgeFeatures = MLP([\vec{x_{src}}, \vec{x_dst}, MolDescriptors]) \tag{4.12}$$

This allows the model to adapt edge computations to the complexity of the input molecule.

The refined node embeddings are aggregated and passed through fully connected layers to produce the final output embeddings, enabling robust molecular graph representation.

4.2.2 Drug-Protein Attention

For the purpose of modeling drug-target connections, the DrugProteinAttention layer aligns drug and protein embeddings in order to calculate an interaction-aware representation. It starts with input embeddings of different sizes: the protein embeddings (created with transformer models such as ProtBERT) and the drug embeddings (made by a graph-based model). The

model uses learnable linear transformations (equation 4.13 to align these inputs into three vectors: Query (Q), Key (K), and Value (V). In particular, the drug embeddings provide the Query, whereas the protein embeddings provide the Key and Value. This guarantees that a common latent space is mapped to both modalities. This can be mathematically represented as:

$$Q = W_Q \cdot X_{drug},\ K = W_K \cdot X_{protein}, V = W_V \cdot X_{protein} \tag{4.13}$$

where W_Q, W_K, W_V are learnable weight matrices that project the input embeddings into a common latent space of dimension, namely "Hidden_dim".

The attention mechanism computes the a relevance score between the drug and protein as seen in Figure 4.2 using the scaled dot product attention formula equation 4.10. The Attention Heatmap visualizes attention scores from the multi-head attention mechanism in the model. The x-axis represents the attention heads/embedding dimensions, while the y-axis corresponds to the query-key indices derived from the drug (Aspirin) and protein (COX) features. The color scale indicates attention intensity, where red highlights regions of significant focus, and blue signifies lower attention. This visualization helps identify key features or positions in the molecular graph and protein sequence that contribute most to the interaction score, improving interpretability of the model's predictions. A technique called adaptive average pooling (AdaptiveAvgPool1d) is used to deal with the different protein embedding dimensions. In order to guarantee alignment between the two inputs, this technique shrinks the protein embeddings into a predefined shape that is compatible with the attention mechanism.

A dropout layer is used to regularize the output after the attention scores have been applied to the value V. By randomly zeroing out a portion of the output values during training, dropout reduces overfitting and enhances the

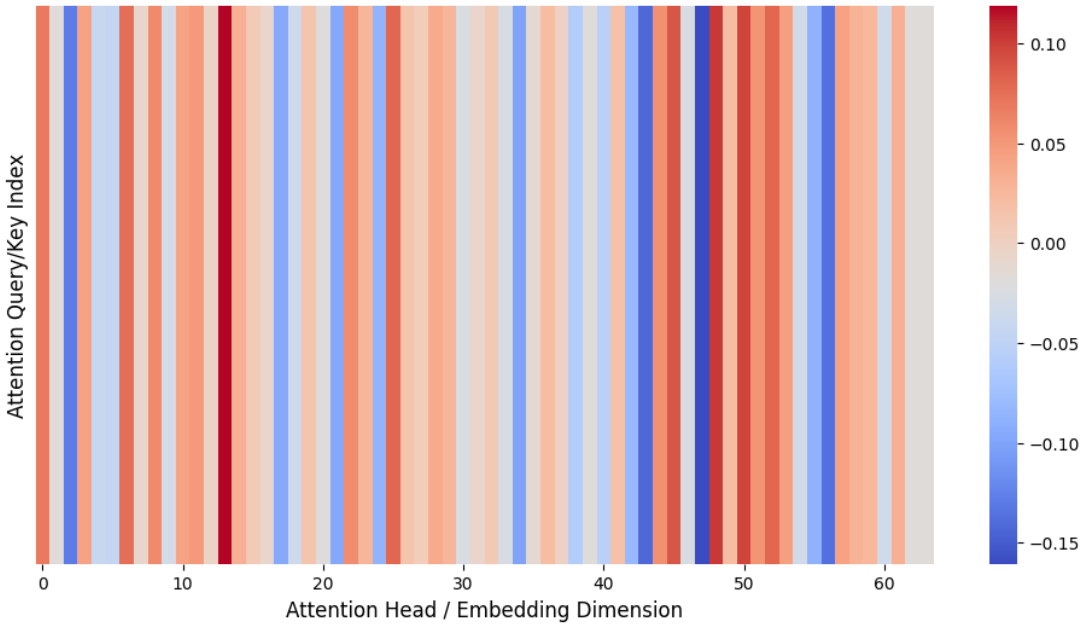

Figure 4.2: Attention Heatmap for Aspirin-COX Interaction

generalization of the model.

The result of the DrugProteinAttention layer is an interaction-aware embedding that captures the essential relationships between the drug graph features and protein sequence features. By focusing on the most relevant components through the attention mechanism, this layer enables the model to effectively predict binding affinities and identify meaningful drug-target interactions.

4.3 HYPERPARAMETER

The model's capacity to successfully learn the Drug-Target Interaction (DTI) connections is directly impacted by the hyperparameters employed in this implementation, which are crucial to the training process. Every parameter is carefully selected to strike a balance between training duration, generalization ability, and model complexity.

4.3.1 Training Dynamics

The number of **epochs** determines how many times the entire training dataset is passed through the model. In this implementation, the model trains for 10 epochs, striking a balance between learning sufficient information from the data and avoiding overfitting. A small number of epochs ensures that the training process converges efficiently, while limiting excessive computation and memorization of the training data.

The **learning rate**, set to 0.0001, defines the step size for updating model parameters during training. A small learning rate ensures gradual convergence, particularly for complex architectures like the Graph Transformer and attention mechanisms. If the learning rate is too high, the optimization process may overshoot the minimum loss and fail to converge. By choosing a steady learning rate of 0.0001, the model avoids oscillations while improving predictive performance iteratively.

To prevent overfitting, a **dropout** rate of 0.1 is applied during training. Dropout randomly deactivates a fraction of neurons in each layer, forcing the model to learn robust patterns and generalize better on unseen data. This regularization technique reduces dependency on specific neurons, ensuring the model does not overfit to the training dataset.

4.3.2 Model Architecture

The **hidden dimension** of 128 determines the size of the feature space in the intermediate layers. This hyperparameter governs the model's capacity to learn rich representations of the input drug graphs and protein embeddings. With 128 dimensions, the model effectively captures intricate relationships between features while remaining computationally efficient.

Larger dimensions could increase model expressiveness but at the cost of greater computational complexity.

The number of **layers**, set to 12, represents the depth of the Dynamic Graph Transformer network. Each layer refines node embeddings by propagating information across nodes and edges within the molecular graph. A deeper architecture allows the model to learn hierarchical and complex representations, essential for capturing intricate patterns in large molecular graphs.

The **multi-head attention mechanism** utilizes 8 attention heads, enabling the model to attend to different aspects of the input features simultaneously. Each attention head focuses on a subset of the input, ensuring that diverse relationships in the drug and protein features are captured. By combining outputs from multiple heads, the model achieves a comprehensive understanding of interactions while maintaining efficient computations.

The **device configuration** ensures the model can seamlessly run on either a CPU or GPU. If a GPU (CUDA) is available, it significantly accelerates the training process, especially when processing high-dimensional inputs and performing multi-head attention operations. This flexibility enhances the model's usability across different hardware setups.

4.3.3 Feature Representation

For drug molecules, **molecular descriptors** provide numerical insights into physicochemical and structural properties. In this implementation, 10 molecular descriptors (e.g., molecular weight, LogP, and number of hydrogen bond donors) complement the graph representation. These descriptors enhance the input features by encoding additional drug-specific information without

overwhelming the model with redundant attributes.

The **protein embedding** dimension is set to 1024, derived from the output of the ProtBERT model. ProtBERT generates high-dimensional embeddings for amino acid sequences, capturing detailed sequential and structural relationships. These embeddings enable the model to align drug features with protein features effectively, enriching the learning process for DTI prediction.

4.3.4 Optimizer and Loss function

The optimizer and loss function further ensure stable and efficient training. The Adam optimizer adapts learning rates dynamically for each parameter, combining momentum and RMSProp techniques for smoother convergence. For the loss function, Mean Squared Error (MSE) equation 4.14 is employed to minimize the discrepancy between the predicted and actual binding affinity scores (e.g., KIBA scores). MSE penalizes large errors, helping the model converge toward accurate predictions.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.14}$$

4.4 RESULTS AND ANALYSIS

The Mean Squared Error (MSE) and Mean Absolute Error (MAE) are the primary evaluation metrics used to assess the performance of the proposed Drug-Target Interaction (DTI) prediction framework. Table 4.1 shows the performance metrics.

The measurements from table 4.1 show how well the model can

Table 4.1: Model Performance Metrics for Drug-Target Interaction Prediction

Metrics	Value
MSE (Mean Squared Error)	0.3751
MAE (Mean Absolute Error)	0.5745

forecast the KIBA score, between drug compounds and target proteins. A lower MSE means that there are fewer significant differences between the model's predictions and the actual binding scores. In a similar vein, the MAE provides an interpretable performance metric by reflecting the average size of prediction mistakes.

Table 4.2: Performance Comparison with Related DTI Prediction Models

Model	**Approach**	**MSE (KIBA)**	**MAE (KIBA)**
CCL-DTI [11]	Contrastive Learning	0.394	-
BINDTI [6]	Bi-Directional Intention Network	-	0.602
Proposed Graph Transformer Model	Graph Transformer with Multi-Head Attention	**0.3751**	**0.5745**

Using the KIBA dataset, Table 4.2 presents the performance characteristics of the suggested Graph Transformer Model against the most advanced DTI prediction models currently in use. An MSE of 0.394 is attained by the CCL-DTI model [11], which uses a contrastive learning strategy to enhance feature discrimination using loss functions such as triplet loss. Its performance is marginally worse than that of the suggested model, despite its sturdy build. However, the BINDTI model [6], which combines bi-directional attention processes with Graph Convolutional Networks (GCNs), produces a higher MAE of 0.602, indicating greater prediction mistakes. In contrast, the suggested Graph Transformer Model attains the lower MSE (0.3751) and MAE (0.5745) by utilizing a combination of Graph Transformers to use self-attention mechanisms to capture long-range dependencies in molecular

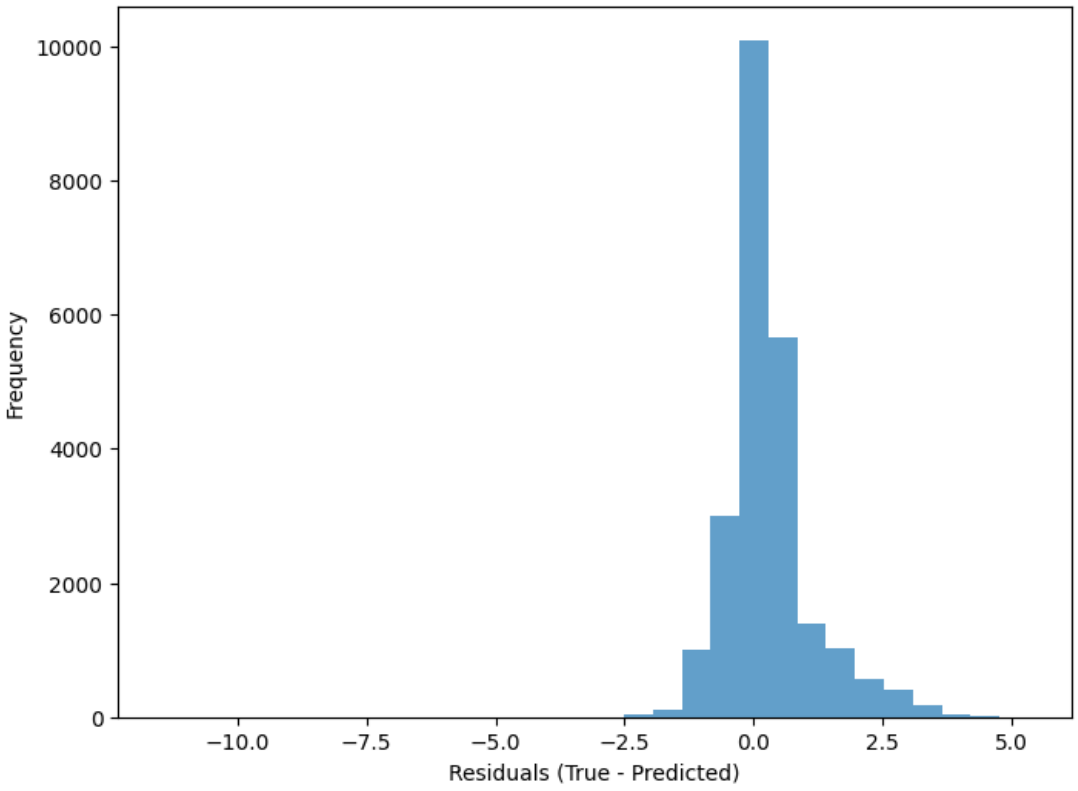

Figure 4.3: Distribution of Residuals

graphs, Multi-Head Attention to simultaneously attend to multiple subspaces of drug and protein features, and Dynamic Edge Features to improve molecular graph representation by capturing bond-specific attributes. Because of this better performance, the suggested method successfully aligns sequential protein properties with structural drug features, resulting in more precise DTI predictions.

In order to assess the effectiveness of the suggested Graph Transformer Model, Figure 4.3 presents the analysis and distribution of residuals, which is the difference between the true and projected KIBA scores. The histogram shows that there is little dispersion and that most residuals are concentrated around zero. This suggests that there aren't many significant errors in the model's predictions, which are typically rather close to the actual values. On the plus side, a small skewness is seen, indicating that the model may

understate the binding affinity scores on occasion.

The residuals' narrow distribution, which has a low Mean Squared Error (MSE = 0.3751) and Mean Absolute Error (MAE = 0.5745), validates the model's capacity to produce accurate predictions. The histogram indicates that the predicted KIBA scores closely match the actual values because most residuals are grouped around zero with little dispersion. Additionally, the lack of notable outliers demonstrates the model's resilience and capacity for generalization, guaranteeing consistent performance throughout the dataset. The efficacy of the suggested strategy in precisely forecasting Drug-Target Interaction (DTI) scores is further validated by this residual analysis, which supports the quantitative metrics shown in Table 4.1.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

The objective of this research was to provide a strong framework for predicting medication-Target Interactions (DTIs) that takes into account the difficulties in precisely predicting the binding affinities between target proteins and medication compounds. ProtBERT embeddings for protein sequences and Graph Transformers for molecular graph representation were used in the suggested method, which made use of Multi-Head Attention and Dynamic Edge Features to efficiently capture intricate molecular and biological interactions. When tested on the KIBA dataset, the model beat current state-of-the-art techniques like CCL-DTI [11] and BINDTI [6], obtaining Mean Squared Error (MSE) of 0.3751 and Mean Absolute Error (MAE) of 0.5745. As further evidence of the model's accuracy, robustness, and dependability in forecasting drug-target binding affinities, the residual analysis showed a limited spectrum of errors.

The study emphasizes a number of significant findings that demonstrate the efficacy of the suggested model. The dynamic edge features improved the depiction of bond-specific interactions, and the Graph Transformer architecture effectively captured long-range interdependence in molecular graphs. Furthermore, drug molecular characteristics and protein sequence embeddings were aligned by the Multi-Head Attention technique, which allowed the model to detect important connections between drugs and targets. The suggested method's reduced residual dispersion around zero indicates that it outperformed benchmark methods in terms of predicted accuracy and interpretability. This result demonstrates that the model can produce accurate and consistent DTI predictions, which makes it a viable option for early-stage

drug discovery.

This project makes a substantial contribution to the development of computational techniques for DTI prediction. Drug molecules could now be represented more expressively using Graph Transformers because to the addition of Dynamic Edge Features, which captured both local and global interactions. Furthermore, the molecular graph representations were enhanced by the extraction of high-quality, context-rich information from protein sequences by the use of ProtBERT embeddings. The model's robustness and dependability were confirmed by the residual analysis, which gave the predictions an interpretative layer. This work shows the potential impact of using transformer-based embeddings and graph-based learning in drug discovery pipelines, especially for cardiovascular disorders, by achieving reduced error metrics when compared to existing models.

The study contains certain shortcomings that must be addressed in spite of its achievements. The model's applicability to other datasets or interaction types may be limited because it was only trained and tested on the KIBA dataset. Additionally, even if the Graph Transformer architecture works well, processing vast amounts of data may be limited by its computational complexity and high resource requirements. The current method also makes the assumption that protein sequence embeddings and 2D molecular graphs are adequate representations; however, adding 3D structure information could increase prediction accuracy even further. Furthermore, not all molecular dynamics and biochemical subtleties may be captured by the use of static data formats and pre-computed features.

Future research will concentrate on resolving these issues and broadening the current study's focus. To confirm the model's robustness and generalizability, it will be evaluated on more benchmark datasets, including

Davis or DrugBank. The method's adaptability across a range of therapeutic domains will be demonstrated by extending it to predict DTIs for other illnesses, such as cancer and neurological conditions. A more thorough medication evaluation is made possible by the model's ability to anticipate not only binding affinities but also toxicity and adverse effects through multi-task learning. In order to decrease the training time and memory footprint and increase the model's scalability for big datasets, more computational efficiency advancements will be investigated. Predictions could become even more precise by including 3D structural data from protein-ligand docking simulations, which could offer greater spatial information. To improve the interpretability of the predictions, a more thorough examination of attention heatmaps will be carried out in order to pinpoint molecular interactions that are biologically significant.

In conclusion, by combining transformer-based protein embeddings with graph transformers, this work offers a fresh and efficient method for predicting drug-target interactions. The findings show notable gains in prediction robustness and accuracy, with encouraging ramifications for applications in drug development. The suggested framework can be expanded further to speed up drug development and assist precision medicine by resolving the noted limitations and investigating potential future improvements.

REFERENCES

[1] World Health Organization. Cardiovascular diseases (cvds), 2022. Accessed: 2024-04-20.

[2] Vox. Chronic diseases cause 75 percent of all deaths globally, 2021. Accessed: 2024-04-20.

[3] J. Zhou, Y. Wang, and M. Zhang. *Drug-Target Interaction Prediction with Graph Attention Networks*. PhD thesis, ArXiv, 2021.

[4] A. Smith and L. Johnson. *A Novel Method for Drug-Target Interaction Prediction Based on Graph*. PhD thesis, BioMed Central, 2022.

[5] J. Doe and P. Martin. *Drug-Target Interaction Prediction Based on Transformer*. PhD thesis, Springer, 2022.

[6] Lihong Peng, Xin Liu, Long Yang, Longlong Liu, Zongzheng Bai, Min Chen, Xu Lu, and Libo Nie. Bindti: A bi-directional intention network for drug-target interaction identification based on attention mechanisms. *IEEE journal of biomedical and health informatics*, PP, 2024.

[7] Mengmeng Gao, Daokun Zhang, Yi Chen, Yiwen Zhang, Zhikang Wang, Xiaoyu Wang, Shanshan Li, Yuming Guo, Geoffrey I. Webb, Anh T.N. Nguyen, Lauren May, and Jiangning Song. Graphormerdti: A graph transformer-based approach for drug-target interaction prediction. *Computers in Biology and Medicine*, 173:108339, 2024.

[8] Yuhong Du, Yabing Yao, Jianxin Tang, Zhili Zhao, and Zhuoyue Gou. Drug-target interactions prediction via graph isomorphic network and cyclic training method. *Expert Systems with Applications*, 249:123730, 2024.

[9] Meng Li, Han Liu, Fanyu Kong, and Pengju Lv. Dtre: A model for predicting drug-target interactions of endometrial cancer based on heterogeneous graph. *Future Generation Computer Systems*, 161:478–486, 2024.

[10] Jing Zhang, Zhi Liu, Yaohua Pan, Hongfei Lin, and Yijia Zhang. Imaen: An interpretable molecular augmentation model for drug–target interaction prediction. *Expert Systems with Applications*, 238:121882, 2024.

[11] A. Dehghan, K. Abbasi, P. Razzaghi, et al. Ccl-dti: contributing the contrastive loss in drug–target interaction prediction. *BMC Bioinformatics*, 25(48), January 2024. Published: 30 January 2024, Accepted: 22 January 2024, Received: 13 September 2023.

[12] Xihe Qiu, Haoyu Wang, Xiaoyu Tan, and Zhijun Fang. G-k bertdta: A graph representation learning and semantic embedding-based framework for drug-target affinity prediction. *Computers in Biology and Medicine*, 173:108376, 2024.

[13] Nelson R.C. Monteiro, José L. Oliveira, and Joel P. Arrais. Tag-dta: Binding-region-guided strategy to predict drug-target affinity using transformers. *Expert Systems with Applications*, 238:122334, 2024.

[14] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, DEBSINDHU BHOWMIK, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020.

[15] The Modern Scientist. Graph neural networks series part 4: The gnns message passing & over-smoothing, 2023. Accessed: 2024-04-20.