9-20-2024

LLMPred: Fine-Tuned Large Language Model Embeddings for Drug Side Effect Frequency Prediction

Siyam Sajnan Chowdhury
University of Windsor

LLMPred: Fine-Tuned Large Language Model Embeddings for Drug Side Effect Frequency Prediction

By

Siyam Sajnan Chowdhury

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2024

LLMPred: Fine-Tuned Large Language Model Embeddings for Drug Side Effect Frequency Prediction

by

Siyam Sajnan Chowdhury

APPROVED BY:

_____

A. Swan
Department of Biomedical Sciences

_____

D. Wu
School of Computer Science

_____

A. Ngom, Advisor
School of Computer Science

September 10, 2024

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Large language models (LLMs) brought about a paradigm shift in the domain of natural language processing, characterized by their large scale, deep architectures, and pre-training on massive amounts of data, enabling them to learn rich and nuanced representations of language. They have demonstrated impressive performance in natural language understanding tasks across different domains. Recent works have started incorporating LLMs in pharmacological domains such as drug discovery and drug interactions.

Drugs play a crucial role in alleviating pain and curing diseases but often come with unintended side effects, which can lead to significant health risks and financial costs. Early detection of these drug side effects during drug development is essential to avoid adverse outcomes. Recent studies have started to focus on a relatively newer problem - predicting the frequencies of given side effects which is an important factor in evaluating therapeutic efficacy. This area, however, remains somewhat underexplored, with only a few studies dedicated to it so far.

In this study, we introduce a novel LLM-based architecture that utilizes LLMs to generate embeddings from drug and side effect attributes in order to predict the frequencies of drug side-effects as well as the high frequency drug side effects. We used Galeano's dataset, a standard benchmark dataset for drug side-effect frequency prediction. Our approach utilized different LLMs to generate embeddings and fine-tune them in order to predict the frequencies. Measuring the frequency of the side effects can help determine the therapeutic efficacy of a drug in clinical settings and help weigh the potential risks and benefits of certain drugs. The key objective of this research is to look into the performance of utilizing large language models for predicting the frequencies of drug side effects.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and gratitude to everyone who has supported me during my course and the completion of my thesis at the University of Windsor.

First and foremost, I wish to extend my heartfelt thanks to my supervisor, Dr. Alioune Ngom, for granting me the opportunity, support, encouragement, and guidance throughout my time at the University of Windsor. Dr. Ngom's unwavering support and exceptional expertise was instrumental in helping me navigate the complexities of a multidisciplinary field. He provided a strong foundation by sharing high-quality and engaging research, and his regular guidance and constructive feedback were crucial in shaping my thesis.

I also want to express my sincere thanks to my thesis committee members, Dr. Dan Wu and Dr. Andrew Swan, for their valuable time and insightful feedback, which greatly contributed to the improvement of my thesis work.

My eternal gratitude goes to my friend Bishwadeep for all his help and constant support. I also wish to extend my heartfelt appreciation to my friend Sudipta for his unwavering assistance and ongoing motivation.

To my parents, Mr. Humayun Chowdhury and Mrs. Nelofar Yasmin Chowdhury, your endless trust, patience, and sacrifices have made this journey possible. To my fiancée, Zaya Mehzabeen, thank you for being my steadfast support throughout this entire journey. To my younger brother, Sahil Sanjan Chowdhury, your constant support has helped me persevere.

Finally, my time at the University of Windsor would not have been the same without the support of my friends—both here and back home. I want to offer my sincere thanks to all my friends, especially Kasra, Kimia, Fahad Bhaiya, and Riaz, who have been essential pillars of encouragement and have made this journey more enjoyable and meaningful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

A molecule consists of two or more atoms chemically bonded together, forming the basic unit of chemical substances, that represents the smallest unit of a pure substance, maintaining the substance's composition and chemical properties even when divided [1]. Drugs are chemical entities composed of specific molecules with pharmacological properties that, when administered to an organism, can bring about changes in the organism's physiology or psychology [2]. These changes do not always bring about the intended therapeutic effects. Sometimes, they can also bring about side effects, which constitute any unintended reactions resulting from drug administration. Side effects can be either therapeutic, providing unexpected benefits [3], or adverse, causing harmful or undesired reactions [4]. Adverse side effects (ASEs) are the undesired reactions of certain drugs and can range from mild symptoms like headaches or nausea to severe complications such as cardiac arrest, brain stroke, organ failure, cancer, or even death [5].

The development and approval of drugs involve a rigorous series of clinical trials that is set in place to ensure safety and efficacy. However, despite these thorough evaluations and tests, absolute drug safety cannot always be guaranteed [6]. Several factors contribute to this uncertainty. The relatively small and controlled nature of clinical trial populations may not adequately represent the diverse real-world populations that will use the drug post-approval [7]. Differences in age, sex,

genetics, pre-existing conditions, and concurrent medications can all influence how a drug affects an individual, potentially leading to unforeseen side effects [6][7]. The limited duration of clinical trials might not capture long-term side effects or rare adverse reactions that only become apparent after extended use or in larger populations [8]. Additionally, an important aspect of clinical trials is the frequency with which an ASE occurs as it is a significant factor in assessing the benefit risk assessment, i.e. the balance between the therapeutic efficacy and the safety risks of a drug [9].

The implications of adverse side effects, especially how frequently they occur, are profound, affecting not only the individuals who experience them but also the broader healthcare system and the drug development process. This is why ASEs have garnered significant attention in recent times as well as looking into the frequencies with which they occur [10].

1.1 Background

ASEs are encompassed by a broader term Adverse Drug Reactions (ADR) which includes other unexpected responses such as allergic reactions or drug interactions besides ASEs. In North America, ADRs are in the top ten causes of fatalities and the annual costs incurred by the Canadian healthcare system is estimated to be over $13 billion (CAD) [11]. Deaths from ASEs rank amongst the fourth to sixth leading cause of fatalities around the world [12].

According to various research studies, it has been estimated that drug-related cases contribute to approximately 1-25% of all hospital admissions and emergency department (ED) visits [13]. It is noted that around two-thirds of these drug-related hospital admissions and ED visits are due to ADRs [14].

Drug discovery involves the innovation of new therapeutic drugs using a combination of computational, experimental, translational, and clinical models [15]. Introducing a new drug to the market has to go through numerous stages of

innovation as well as testing. The average total cost of research and development over many years to introduce a single drug to the market is in the range of approximately $314 million USD to $2.8 billion USD [16]. ASEs pose a significant challenge for pharmaceutical companies, as their occurrence during clinical trials can slow down the drug discovery process and prevent many candidate molecules from being developed into commercial drugs [17]. ASEs are the most common reason for a drug not making it into commercial development, second only to lack of efficacy [18].

With the lengthy and expensive experimental procedures to explore the side effects during the trial phase of a drug in traditional wet experiments, recent researches have started to look into in silico methods, especially those based on machine learning and deep learning, to help expedite the screening process and reduce the expensive trial costs [19].

1.2 Drug Computational Representation

Machine learning models need to convert different forms of data such as texts, images, audio etc. to dense, lower-dimensional numerical vectors to be able to process the data and capture underlying patterns and relationships [59]. These dense vectors are called embeddings.

In order to obtain embeddings for drugs and side effects using Large Language Models (LLM) and predict the frequencies of those side effects using deep learning techniques, we need different computational representation of drugs and side effects. Representation of drugs for computational purposes can be generalized into 4 types – Chemical Structure Representation by strings, Molecular Fingerprint, Drug Descriptors, and Graph Representation as illustrated in Figure 1.2.1.

These representations serve as the foundation for converting complex chemical and biological information into a format that machine learning models can process and analyze. Each of these types captures unique aspects of the drug's chemical

makeup and biological activity, providing a comprehensive dataset for predictive modeling. These diverse computational representations are essential for generating high quality embeddings for downstream tasks such as drug discovery, drug side-effect prediction and drug side-effect frequency prediction.

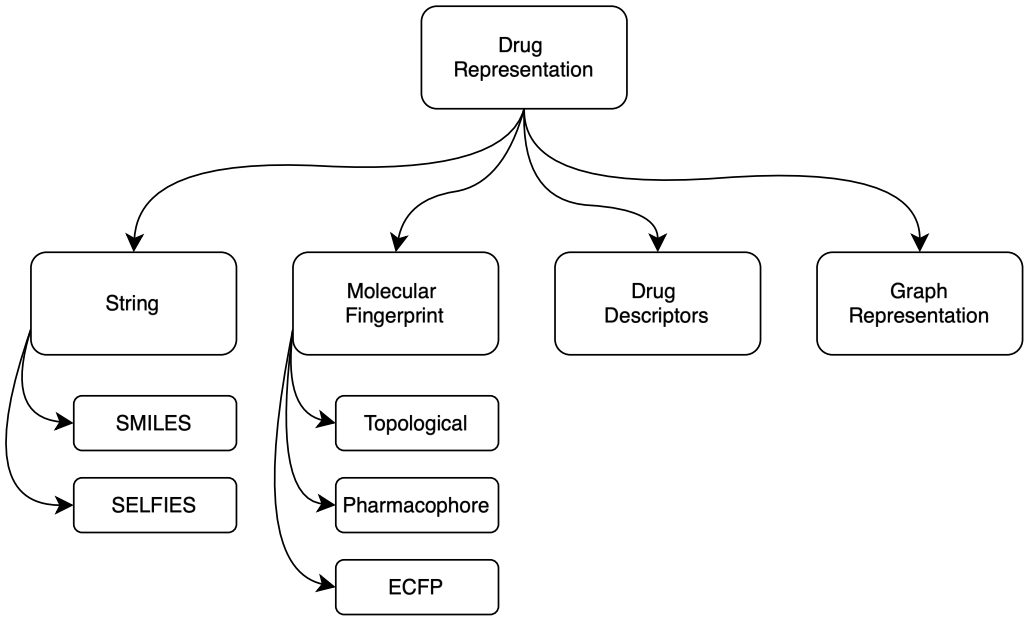


Figure 1.2.1: Category of Drug Representations

1.2.1 String Representation

The simplest string representation that a molecule can be represented by is its molecular formula. Molecular formula consists of just the different atoms that a molecule contains and how many of those molecules are there. For example, the molecular formula of glucose would be represented as $C_6H_{12}O_6$. However, this is a reductive and minimalistic representation of a molecule and lacks information regarding the structure of the molecule. In computational chemistry, this lack of information could hinder the performance of models as it requires representations containing more information such as the bonds and chains of molecules.

Some of the more expressive string representations of molecules which is more information-rich are SMILES representation and SELFIES representation of molecules:

1.2.1.1 SMILES

SMILES (Simplified Molecular Input Line Entry System) string representations are a notation system that consists of ASCII characters and follows a set of rules to represent chemical structures in a linear textual format [20]. Some of the key features are:

- Atoms are represented by their atomic symbols. For example, "C" for carbon and "O" for oxygen.
- Single bonds are implied, so they are not shown, and double bonds, triple bonds and aromatic bonds are represented by "=", "#", and ":" respectively.
- Branches are represented using parentheses. For example, "CCC(C)O".
- Rings are represented by numbers at the start and end of a string. For example, cyclohexane would be represented as "C1CCCCC1" with the Cs enclosed by the 1s denoting the ring.
- Aromatic atoms are represented by lowercase letters.

An example of 3-cyanoanisole illustrating the ways in which bonds, branches and rings discussed above are represented by a SMILES string is shown below:

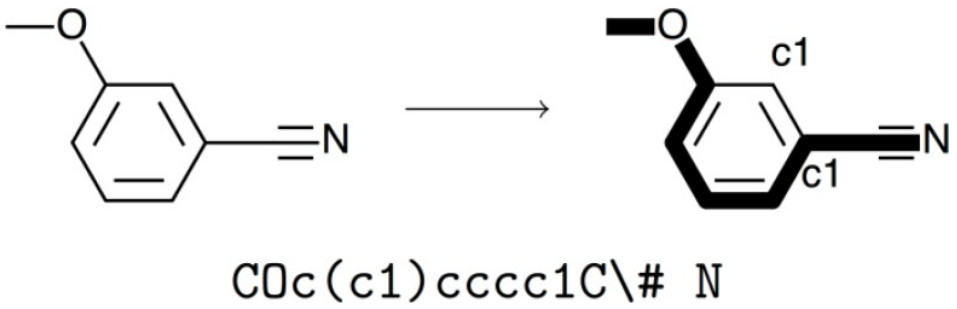


Figure 1.2.2: SMILES representation of 3-cyanoanisole [60]

1.2.1.2 SELFIES

SELFIES (Self-Referencing Embedded Strings) is a novel, machine-readable molecular representation that uses context-free grammar to encode structural information of molecules [21]. SELFIES representation is particularly useful in computational chemistry for representing chemical compounds in a way that ensures syntactical validity, making it more robust for use in machine learning and other computational models. Some of the key features of SELFIES are:

- Self-referencing grammar ensures valid chemical structure thus not requiring post-generation validation checks.
- All SELFIES strings are syntactically valid.
- Its canonical form can uniquely represent molecules.

An example of a SELFIES string representation can be seen in Figure 1.2.3.

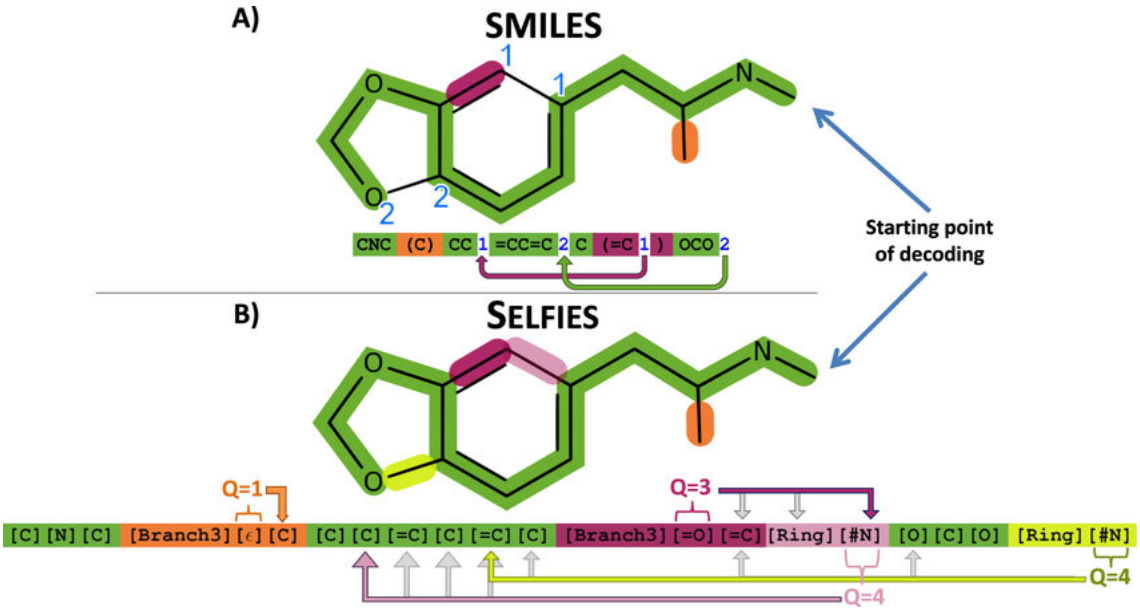


Figure 1.2.3: Molecular Graph Representation as (A) SMILES (B) SELFIES [21]

1.2.1.3 Other String Representations

There are numerous other string representations of the chemical structure of molecules such as InChl [22], Canonical SMILES [23], SMARTS [24], etc.

1.2.2 Molecular Fingerprint

A molecular fingerprint is a vector representation of a molecule's structure, capturing specific information about its atomic and molecular features. These fingerprints capture certain patterns or structures within molecules.

They are also capable of representing certain substructures or properties that enables efficient and convenient comparison and analysis of different compounds in large databases.

Described below are some of the most common fingerprinting techniques:

1.2.2.1 Topological Fingerprint

It captures the molecular connectivity and the arrangement of atoms in a molecule. The name comes from this approach encoding topological information – information regarding how atoms are connected to one another within a molecule into a fixed length binary or integer vector [25].

A widely used example of topological fingerprint is the Daylight Fingerprint [68] method of encoding molecules. In Daylight Fingerprint generation of molecules, various substructures are identified based on different number of bonds in a path – it represents a sequence of *n* number of connected atoms. These substructures are then hashed into specific positions into a bit string. The positions in the fingerprint represent the presence or absence of these substructures in a molecule. Finally, the binary fingerprint formed is a fixed-length binary vector where each position indicates whether a substructure (path) exists in a given molecule.

An example of how daylight fingerprint encodes a molecule into a binary fingerprint vector is illustrated by the encoding of the molecule cyclohexanone in the figure below:

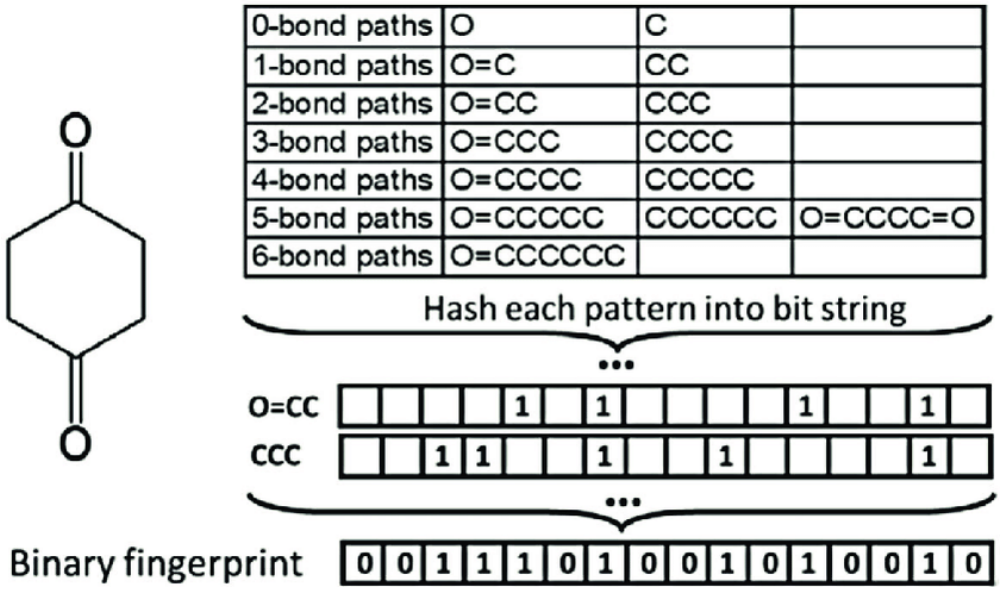


Figure 1.2.4: Daylight Fingerprinting of Cyclohexanone [68]

1.2.2.2 Pharmacophoric Fingerprint

The pharmacophore model is used in drug discovery and design to identify and optimize potential drug candidates by ensuring they possess the necessary features to elicit the desired biological response. They look for pharmacophoric features such as hydrogen bond donors, hydrogen bond acceptors, hydrophobic centers, et needed to bind to target proteins or enzymes.

This type of fingerprint, unlike other types of molecular fingerprinting techniques where the focus is on bond-level or atomic connectivity structure as explain in the previous section, focuses on abstracting the molecule into sets of key functional groups that are significant for drug target interactions.

To generate pharmacophoric fingerprints, 3D structures containing the spatial information of a molecule is considered. The next step is to abstract away the bonds and atoms and just focus on key pharmacophoric features such as hydrogen bond acceptors, donors, cationic and anionic groups, et cetera. This step of isolation of functional group assists in focusing on the features that are more important for target interaction rather than considering the whole atomic structure.

The next step is to connect these abstracted functional groups representing the spatial relationship between them followed by distance between the functional groups being measured. Finally, the distances and relationships are encoded into a binary fingerprint that represents if specific distance or arrangement between these pharmacophoric features are present or absent.

This process of encoding a molecule to its pharmacophoric fingerprint is illustrated by the example below:

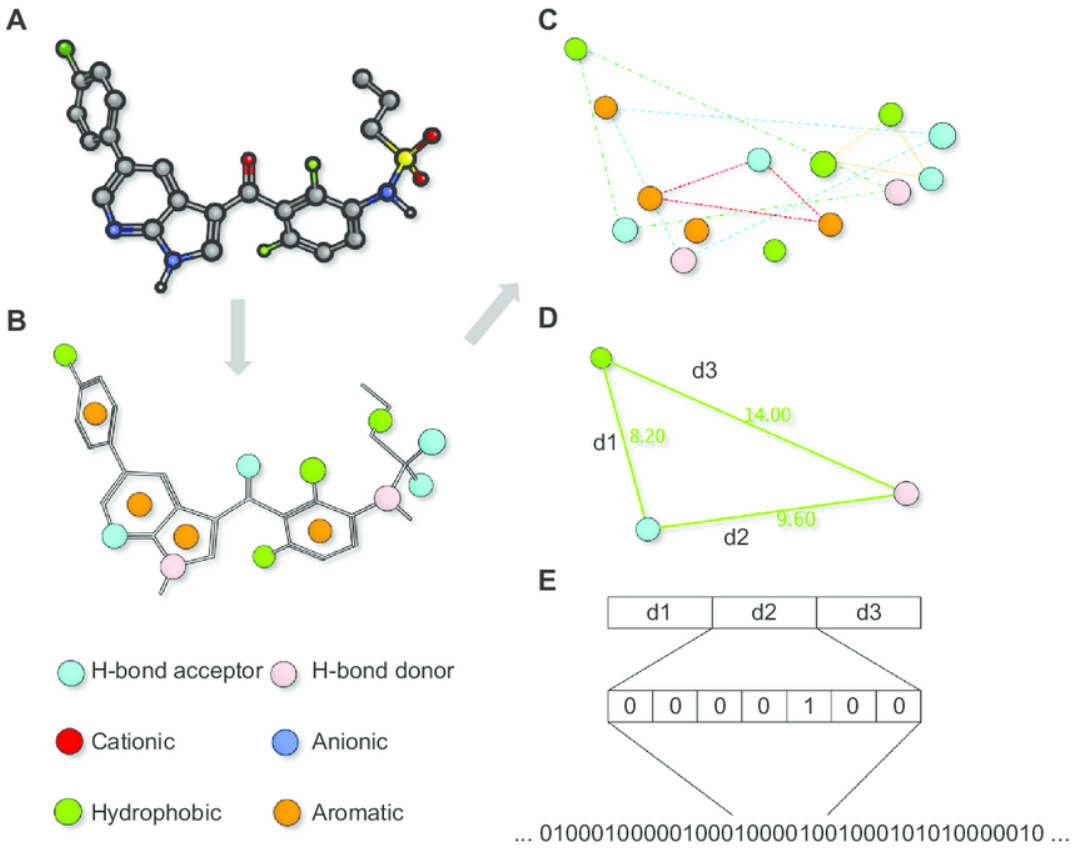


Figure 1.2.5: Pharmacophoric Fingerprint [69]

1.2.2.3 Extended Connectivity Fingerprint (ECFP)

ECFP is a type of molecular fingerprint used in chemical informatics to represent the structural features of molecules [26]. They are designed to capture the presence and arrangement of chemical substructures within a molecule. ECFPs are widely

used in virtual screening, similarity searching, and quantitative structure-activity relationship modeling. They provide a robust and compact representation of molecules, facilitating the comparison and analysis of chemical compounds based on their structural similarities and differences [70]. ECFPs are also referred to as circular fingerprints as they encode the local spatial environment of an atom considered to be the central atom within a specific radius which is usually around 2-4 bonds.

To encode molecules to its ECFP, the first step is to identify a central atom. This is followed by setting specific radii, or the number of bonds from the central atom to be considered. This is illustrated by concentric circles of different radius showing different distance from the central atom. This splits the molecule into atom centered fragments. These fragments are then hashed to specific bits in a fixed-length binary vector. These represent the presence or absence of these fragments in a molecule.

This process is illustrated in the figure below:

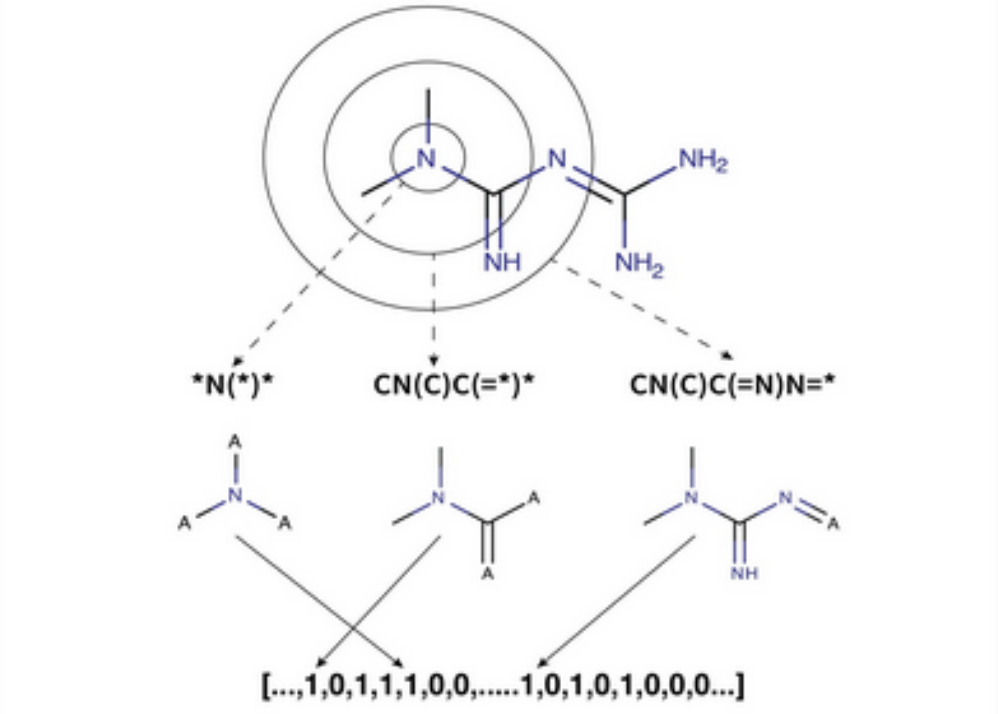


Figure 1.2.6: Extended Connectivity Fingerprint [71]

1.2.3 Molecular Descriptors

Molecular descriptors are quantitative descriptions of a molecule's chemical structure and properties that can be used in cheminformatics and computational chemistry to characterize and analyze chemical compounds. These contain any physicochemical numeric descriptors such as molecular weight, polarity, partition coefficient, solubility, etc.

Molecular descriptors such as molecular weight or partition coefficient are encoded numerically as scalar values. Molecular descriptors such as presence or absence of hydrogen bond donors can be represented as Boolean values. Numerous of these physicochemical molecular descriptors can be combined together to create a feature vector. An example of a feature vector could look like:

[Molecular Weight: 300, Partition Coefficient: 2.5, H-bond Donors: 1, Molar Refractivity: 80.5, Solubility: 0.001]

These vectors are crucial in predicting biological activity of molecules based on their molecular descriptors as well as searching similar molecules based on similarity in their physicochemical properties.

1.2.4 Graph Representation

Molecular graph representation models drugs as graphs, with atoms represented as nodes and bonds as edges. This approach captures the connectivity and spatial relationships between atoms within a molecule, providing a detailed depiction of the molecule's structure. Such representations are highly useful in computational chemistry and drug discovery, as they allow for the application of robust and efficient learning algorithms specifically designed for graph data.

These algorithms can analyze the molecular graph to predict properties, activities, and interactions, making them invaluable tools in the identification and optimization of potential drug candidates.

In this representation, each atom is represented as a node in a graph and the bonds between those atoms are represented by edges in a graph. Different types of atoms can be represented by different types of nodes and the edges can also be of different types based on the different types of bonds between them. However, in the simplest form, all bonds are treated as generic edges.

An example of molecular graph representation of the molecules 4-Pentynenitrile and Toluene is illustrated below:

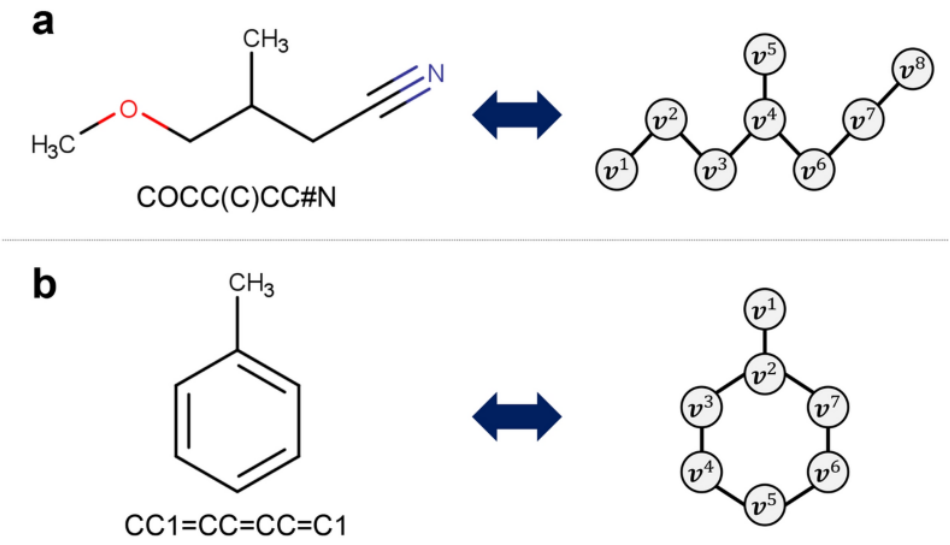


Figure 1.2.6: Molecular Graph Representation of a) 4-Pentynenitrile and b) Toluene

1.3 Large Language Models

Large Language Models (LLMs) are advanced neural network architectures designed to understand, generate, and manipulate human language. These models have very large number of parameters, often in the range of hundreds of millions to tens of billions. These large number of parameters enable these models to capture

the nuances of language and develop comprehension unlike any other machine learning models. These models are built on deep learning techniques, particularly leveraging the Transformer architecture [27], which has revolutionized natural language processing (NLP) due to its ability to handle vast amounts of data and capture complex linguistic patterns.

1.3.1 Advantages

The advantages of using LLMs are:

- They are extremely versatile and can perform a wide array of different tasks such as text generation, translation, summarization, sentiment analysis, etc.
- LLMs have the capacity to generalize very well.
- LLMs excel at context comprehension due to the attention module of the Transformer architecture [27].
- Pre-training of LLMs on huge datasets allows it to come preloaded with context of language thus reducing the need for large, labeled datasets.
- LLMs have demonstrated impressive performance in zero-shot learning (where the model has to categorize or recognize types of a sample without ever having seen any sample of that type) and few-shot learning (where it was exposed to very minimal training data).

1.3.2 Types of LLMs

Different types of LLMs have been discovered that excel in different tasks. However, they could be categorized into 2 main types:

- **GPT (Generative Pre-trained Transformer)** [28]: GPT models are used primarily for generating texts. It utilizes the unidirectional decoder block from the transformer model predicting the next word in a sequence

based on the input sequence. GPTs are pre-trained on a massive corpus of text thus making it capable of very accurate text generation in different domains and can also be fine-tuned to fit various different tasks. One of the most common uses of GPTs are in OpenAI's ChatGPT [32] and Google's Gemini [33].

- **BERT (Bidirectional Encoder Representation from Transformers)** [29]: BERT are primarily used for tasks such as question answering and comprehension of natural language using embeddings.

1.4 Problem Statement

Given a Drug Side-Effect Frequency (DSF) network - consisting of drug $d_m \in D$ where D is a drug set containing m drugs, side effect $s_n \in S$ where S is a side effect set with n side effects, with known frequency values between d_m and s_n; and a new pair $p_{ij} = (d_i, s_j)$ the goal is to predict the frequency $F_{d_i s_j}$ between the new drug and side effect pair d_i and s_j where $F_{d_i s_j} \in [1, 5] \cap \mathbb{Z}$.

For instance,

- DSF – Galeano and Zhao's Dataset [30][31],
- Drug (d) – Gadobutrol
- Side Effect (s) – Headache
- Task - $(d_i, s_j) \rightarrow [1, 5]$

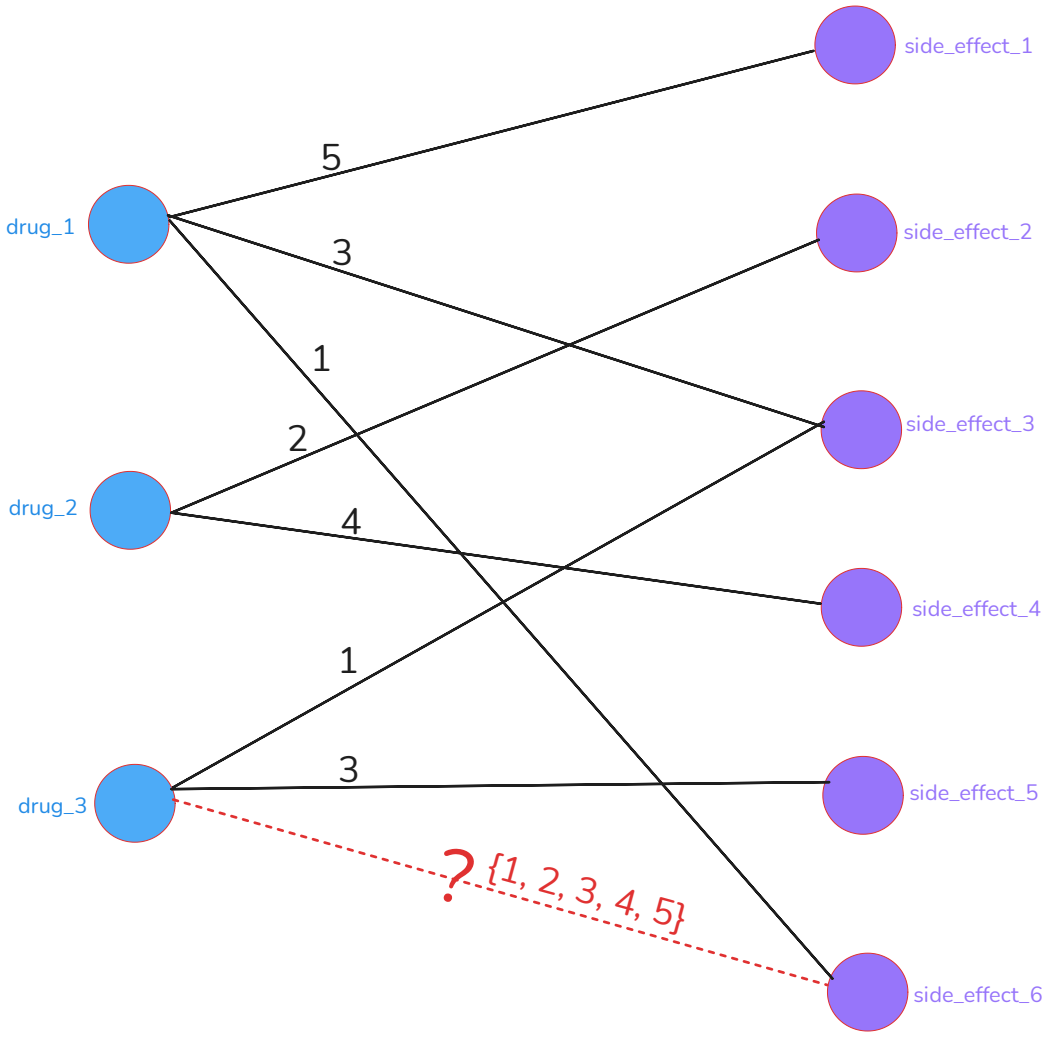


Figure 1.4.1: DSF Network (Sampled)

Furthermore, given a Drug Side-Effect High Frequency (DSHF) network, which is the binarized DSF network with frequency values greater than 3 considered to be high frequency (1) and the rest as 0, consisting of drug set $d_m \in D$, side effect set $s_n \in S$, with known high frequency values between d_m and s_n; and a new pair $p_{ij} = (d_i, s_j)$ we have to predict the high frequency $H_{d_i s_j}$between the new drug and side effect pair d_i and s_j where $H_{d_i s_j} \in [0, 1] \cap \mathbb{Z}$.

For instance,

- DSHF – Galeano and Zhao's Dataset [30][31] binarized,

- Drug (d) – Gadobutrol
- Side Effect (s) – Headache
- Task - $(d_i, s_j) \rightarrow [0, 1]$

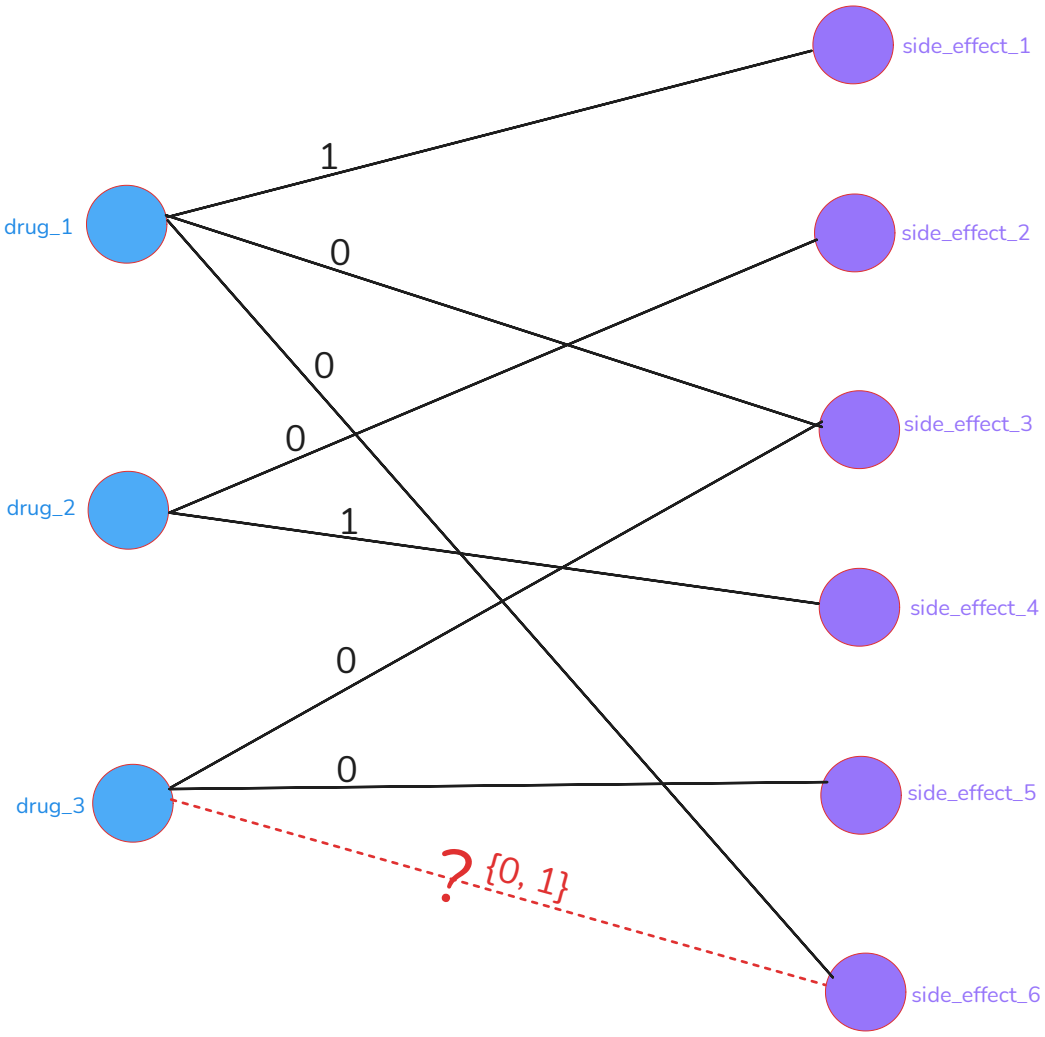


Figure 1.4.2: DSHF Network (Sampled)

This research intends to develop a novel similarity based architecture to generate embeddings using LLMs and fine-tune them using a DSF network curated in previous research [30][31] along with additional descriptions curated from multiple sources to predict the frequency of certain side effects for certain drugs.

We generate the embeddings of the attributes in two concurrent steps. We generate embeddings of the drug chemical structure and side effect names and then generate embeddings of the biomedical text information of the drugs and side effects. We then concatenate these embeddings and perform the final task of drug side effect frequency prediction.

We also use the exact same framework for high frequency drug side effect prediction where we binarized the DSF network.

1.4.1 Thesis Motivation

Understanding the frequency of side effects associated with a drug is crucial for evaluating its overall therapeutic efficacy in clinical settings. By quantifying how often specific adverse reactions occur, healthcare providers and researchers can assess the safety profile of the medication and make informed decisions about its use. This information helps in identifying the threshold at which the benefits of the drug outweigh its potential risks, particularly for drugs that treat serious conditions but may have significant side effects. For instance, a drug that is highly effective in treating a life-threatening illness but has a known side effect that occurs in a very small percentage of patients may still be considered beneficial. On the other hand, if the side effects are more common or severe, the risk-benefit analysis might suggest alternative treatments. Additionally, understanding the frequency and nature of side effects can guide dosage adjustments, patient monitoring strategies, and the development of guidelines for safe use, thereby optimizing patient care and improving therapeutic outcomes. Finding out DSF to ascertain the efficacy of a drug via wet experiments in lab can be quite lengthy and costly. In silico methods could help us filter out potential side effects and have better screening techniques.

LLMs have demonstrated impressive capabilities in NLP tasks, particularly in capturing context comprehending contextually relevant text. Unlike earlier models that struggled with understanding context beyond a few words, LLMs can process

entire sentences or even paragraphs, understanding the nuances of language, idiomatic expressions, and syntactic structures. Also, when these models are fine-tuned on specific tasks, they start from a strong foundational understanding, requiring less task-specific data to achieve high performance.

Despite the overwhelming positives of LLMs, their effectiveness has not been explored in the domain of DSF prediction. This research aims to investigate whether LLMs can enhance predictive performance in this relatively new domain by leveraging their contextual understanding and advanced language processing abilities.

1.5 Thesis Contribution

The key contributions of our paper are as follows:

- We design and train a novel LLMPred model that is based on BERT-based LLMs to predict DSF and DSHF.
- We expand the current benchmark dataset of Galeano [30] and Zhao [31] with biomedical semantic text information of drugs and side effects.

1.6 Thesis Organization

The remainder of the thesis is organized as follows:

Chapter 2 comprehensively reviews the relevant literature in the domain of drug side effect frequency prediction as well as papers on BERT and other LLMs relevant to this research.

Chapter 3 introduces the architecture of our proposed LLM model and explains it in detail.

In Chapter 4, we go through the experimental setup and the evaluation metrics used in this research.

Chapter 5 discusses the performance of our LLMPred model in a 10-fold cross-validation study. We compare the results obtained in our study to the current state-of-the-art methods.

In Chapter 6, we summarize our research findings, share the insights gained throughout the study, and identify potential areas for future exploration.

CHAPTER 2

Literature Review

This section discusses the relevant scholarly research on prediction of frequency of drug side-effects. Large Language Models (LLMs) have been incorporated in this research in order to generate embeddings for the various attributes and descriptors related to drugs and side effects. The domain of pharmacology, especially those on drug research encompasses diverse data sources, including chemical representations of drugs, drug descriptors, side effects, and description of side effects. First, we will discuss the key databases from which the data was sourced, followed by a review of the relevant literature.

2.1 Datasets

2.1.1 PubChem

PubChem [34] is a comprehensive, freely accessible database maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM). PubChem contains three interlinked databases: PubChem Substance, PubChem Compound, and PubChem BioAssay. As of now, PubChem contains over 115 million unique chemical compounds, 309 million substances, and 292 million bioassays bioactivity results, making it one of the largest chemical information repositories available. It provides extensive data such

as chemical structures, identifiers, chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations, and links to other databases. Researchers, educators, and the general public widely use PubChem for drug discovery, chemical research, and educational purposes.

2.1.2 SIDER

SIDER (SIDE Effect Resource) [35] is a comprehensive database of drug-side effect associations, compiled from multiple sources such as package inserts, public databases, and scientific literature. The latest version, SIDER 4.1, includes data on 1430 unique drugs and 5868 unique side effects, encompassing a total of 139,756 known associations.

2.1.3 DrugBank

DrugBank [36] is an online database that offers comprehensive details on drugs, including chemical, pharmacological, and pharmaceutical information, as well as data on drug targets such as protein sequences, structures, and pathways. The database features over 200 data fields, providing extensive coverage. The latest version, DrugBank 5.1.10, includes information on a total of 15,758 approved experimental drugs.

2.1.4 STITCH

STITCH, which stands for Search Tool for Interactions of CHemicals, is a comprehensive resource containing information on chemical interactions including drugs and their association [61]. It also integrates information on chemical-protein interactions. Different types of associations contained in STITCH are Similarity, Experimental, Database, Text-Mining, and Combined Score. The latest version of

STITCH contains 500k chemicals, 9.6 million proteins and 1.7 billion chemical protein interactions.

2.1.5 OnSIDES

OnSIDES (ON-label SIDE effectS resource) is a database that consists of adverse drug events which have been extracted from drug labels by a BERT-based language model [62]. The 200 drug labels from which the adverse drug events were extracted from were manually curated [63]. The latest version of OnSIDES contains 3.6 million drug-adverse drug event pairs for 2,739 drugs extracted from 46,686 labels.

2.2 Drug Side Effect Prediction

Drug Side Effect Prediction involves predicting all the possible side effects associated with a drug. Since a single drug can be associated with multiple side effects, this task is inherently a multi-label, multi-class classification problem. To tackle this, various problem transformation methods can be employed. One of the most key approaches is the binary relevance approach where each drug-side effect pair is treated as a separate binary classification problem.

In this binary classification approach, provided a drug side effect network, the objective of the model is to predict whether a link exists between a given drug and a given side effect pair.

In this approach, the network is a bipartite graph of drug nodes and side effect nodes and if a side effect exists for a certain drug, there will be an edge between the drug node and side effect node. This is illustrated in Figure 2.2.1:

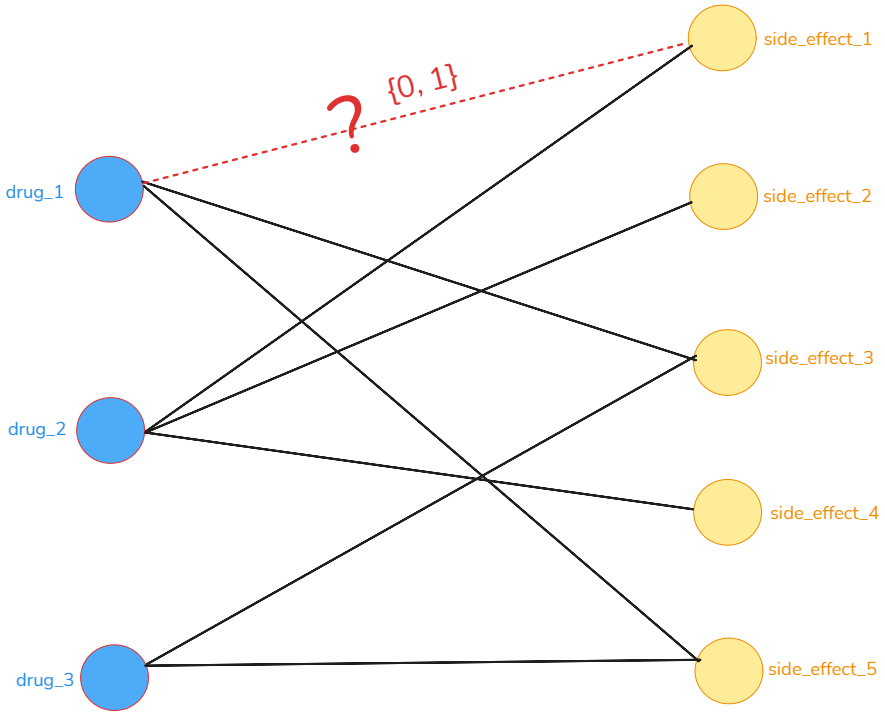


Figure 2.2.1: Drug Side Effect Prediction

The key difference between drug side effect prediction and drug side effect frequency prediction is that in the drug side effect prediction, the association between the drug and side effect nodes, that is, if a certain side effect is associated with a certain drug, is illustrated by the edge between the nodes, and an absence of an edge between the nodes represent no association between that drug side effect pair.

In contrast, drug side effect frequency prediction is where the edges between the drug and side effect nodes represent the presence of an association between a drug and side effect pair, and there are edge weights for those edges that represent how frequently the side effect occurs given that drug is administered.

2.3 Current Literature on Drug Side Effect Prediction

Recent works in the domain of drug side effect prediction can by seen in the work of Yu et al. [64] – IDSE-HE, where they approach the drug side effect prediction problem by framing it as an adjacency matrix reconstruction problem utilizing learned representations through a novel hybrid embedding Graph Neural Network (GNN).

Their method employs dual-view learning with multiple feature perspectives for drugs, specifically using Molecular Fingerprint and Molecular Graph Embedding representations, along with randomly Xavier-initialized vectors for side effects. For molecular graph representation learning, they first designed and trained a simple framework combining a Message Passing Neural Network (MPNN) for the message-passing phase with a set2set model for the readout phase. The trained model was then used to extract the physicochemical properties of drugs.

Subsequently, a graph-level GNN was applied to capture relational information within the overall drug-side effect network. Finally, they multiplied the drug feature matrix and the side effect feature matrix and reconstructed the adjacency matrix.

Their work achieved an F1-score of 0.55 on the SIDER dataset [35] using 10-fold cross-validation.

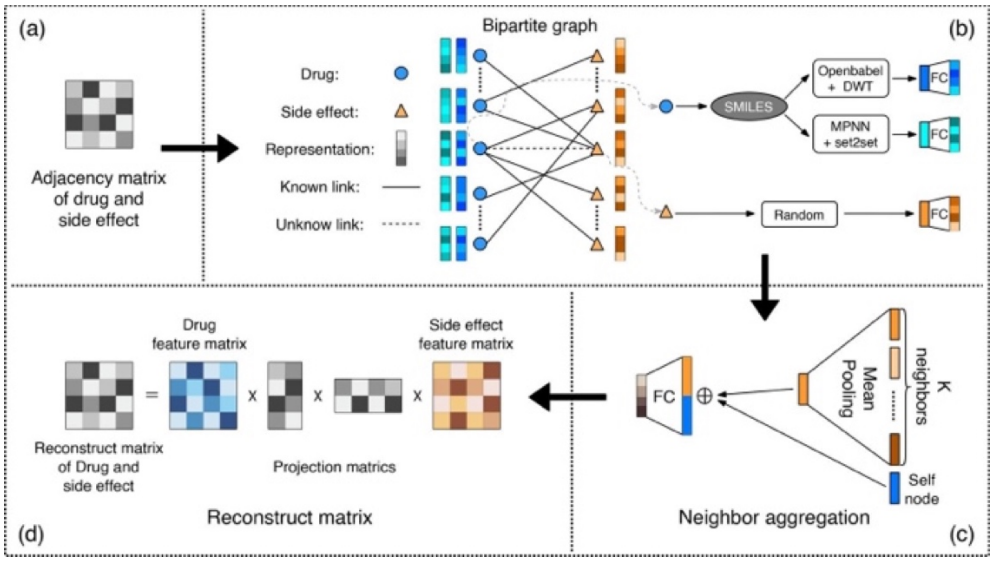


Figure 2.3.1: Architecture of IDSE-HE [64]

2.4 Current Literature on Drug Side Effect Frequency Prediction

2.4.1 Galeano's Model

In Galeano et al.'s study [30], the authors present a framework to predict the frequencies of drug side effects. This paper addresses an elemental part of drug risk-benefit assessment. This assessment is conventionally performed through randomized controlled clinical trials, but this paper goes to challenge that and illustrates the future where a preliminary risk benefit assessment could be ascertained using this framework. They employ a matrix decomposition algorithm that identifies latent, biologically interpretable signatures of drugs and side effects. Their model is trained on a dataset that they curated from the SIDER 4.1 [35] database and Drugbank [36] that has frequency information from 0 to 5 of 759

drugs and 994 side effects. The frequencies are categorized into five classes: very rare, rare, infrequent, frequent, and very frequent.

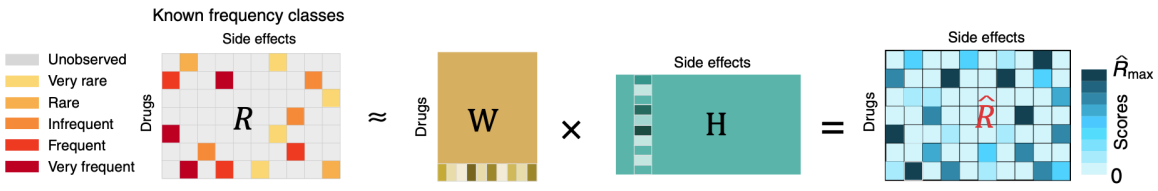


Figure 2.4.1: Architecture of Galeano's Model [30]

The matrix decomposition approach enables the model to learn low-dimensional representations of drugs and side effects. The side effect and drugs are decomposed into drug and side effect vectors represented into a latent space where latent features of drugs capture features related to the drugs' anatomical, therapeutic, and chemical properties, while side effect features relating to the physiological systems affected are captured. The frequency of a drug-side effect pair is predicted by the dot product of their respective matrices.

Galeano's model demonstrates robust performance after performing 10-fold cross validation, achieving a root mean squared error (RMSE) of 1.32 and an area under the receiver-operating curve (AUROC) of 0.932 on the test set. The predictions obtained offer insights into the biological mechanisms underlying drug side effects. The non-negative matrix factorization made the predictions interpretable which enabled the connection from drugs to physiological effect. The approach is especially useful in early clinical trial phases where it can be used to construct a complementary hypothesis for risk assessment.

Galeano et al.'s work introduces a novel, computational method for predicting drug side effect frequencies, potentially enhancing drug safety evaluations, and reducing the reliance on extensive clinical trials.

2.4.2 MGPred

In Zhao et al.'s study [31], the authors propose MGPred, a framework utilizing multi-view data for drug side effect frequency prediction using a novel graph attention model. MGPred uses different types of data such as the similarity of chemical structure for drugs, drug-side effect frequency information, and semantic similarity of the side effects. This multi-view approach provides a wider array of information that allows for a more comprehensive prediction of side effect frequencies.

The study utilizes Galeano's benchmark dataset containing 37,071 drug-side effect frequency pairs extracted from 750 drugs and 994 side effects. MGPred's architecture consists of two segments - a feature extraction part and a prediction part.

In the feature extraction part, separate feature extractors are fed with features from all the different kinds of views. These feature extractors learn latent feature representations of the different attributes by aggregating heterogeneous neighborhood features through a node-level attention module. The representation of drug nodes and side effect nodes are concatenated to get the overall representation vector.

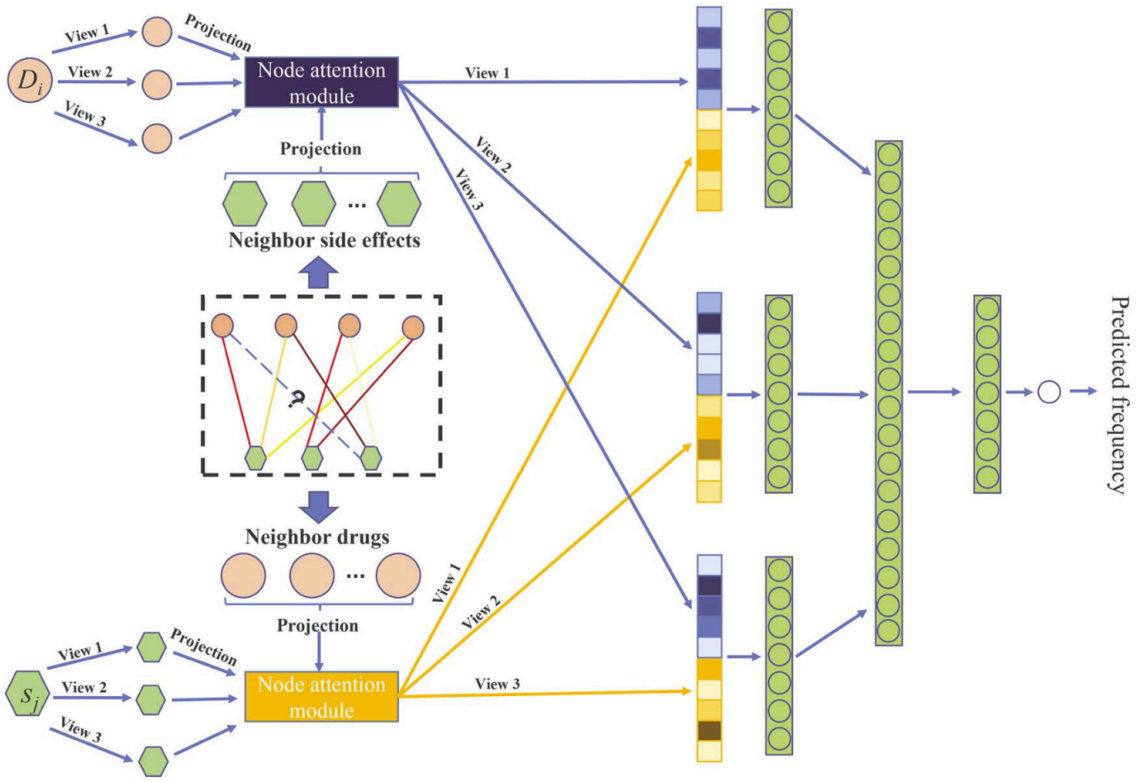


Figure 2.4.2: Architecture of MGPred [31]

In the prediction part, a multilayer perceptron model is used to predict the drug side effect frequency values.

MGPred demonstrates superior performance compared to Galeano's model on the benchmark dataset. The model's effectiveness is validated through 10-fold cross-validation and ablation experiments. MGPred achieved a root mean squared error (RMSE) of 0.6521 and a mean absolute error (MAE) of 0.4905.

This study illustrated the improvement in performance when multiple data sources were incorporated and using deep learning models to improve the prediction of drug side effect frequencies.

2.4.3 SDPred

In Zhao et al.'s study [37] from a year later, the authors came up with SDPred to predict the frequencies of drug side effects using a similarity-based deep learning architecture. Previous methods were limited to predicting frequencies of known drug-side effect, rendering them ineffective for ascertaining the frequency information for new drugs. SDPred, overcomes this limitation by integrating rich features and multi-correlation embeddings, making it applicable to new drugs without any prior data on frequency information.

The SDPred framework builds on Galeano's dataset and expanded it to include 757 drugs and 994 side effects, which contains 37,366 frequency values. The data is taken from the STITCH and DrugBank databases. It consists of chemical structures, drug target proteins, and pre-trained word vectors, and other features. Incorporating these different features model has a robust foundation for predicting drug side effects.

A very crucial element of SDPred is its feature extraction and integration framework. 10 drug similarity matrices are constructed based on different similarity information, chemical structure, target proteins, word embeddings and interaction profiles. The similarity matrices for side effect similarity are developed using semantic similarities and other factors. These matrices are stacked together providing a comprehensive multi-dimensional view of the drugs and their potential side effects and the embeddings are generated followed by the creation of different interaction maps.

SDPred employs a multi-task learning framework where it uses convolutional neural networks (CNNs) to capture high-order correlations between drug and side effect vectors from the interaction map. Multi-layer perceptron (MLP) are then used to combine these embeddings to predict drug-side effect frequencies. This architecture enables SDPred to handle the complex relationships between drugs and their side effects to predict the frequencies.

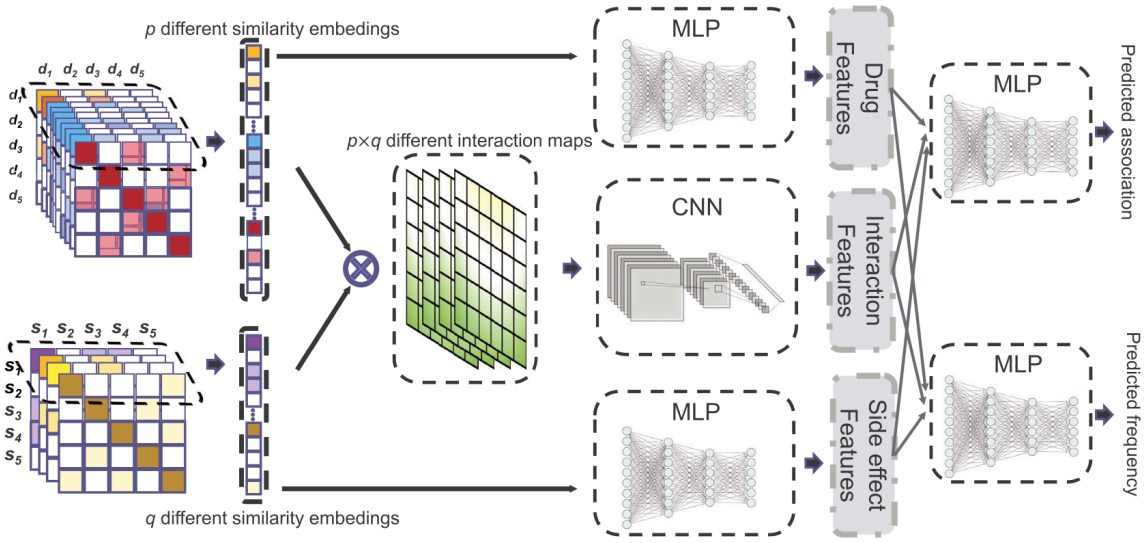


Figure 2.4.3: Architecture of SDPred [37]

The model's effectiveness is rigorously tested as SDPred significantly outperforms existing models in predicting drug side effect frequencies. In 10-fold cross-validation, the model achieves a notable improvement in predictive performance. SDPred demonstrates superior performance across all the metrics used with an MAE of 0.4212 and RMSE of 1.3212, indicating its robustness and reliability in predicting drug side effect frequencies.

SDPred provides significant advantages over traditional models and frameworks for predicting drug side effect frequencies. Its ability to handle new drugs without prior side effect data makes it particularly valuable for drug development and risk assessment.

2.4.4 DSGAT

In Xu et al.'s study [38], the authors propose DSGAT to predict the frequencies of drug side effects. DSGAT introduces a novel deep learning model that utilizes graph attention networks (GAT) for this task. Traditional methods often rely on drug-side effect interaction graphs, which suffer from sparsity and inability to handle cold

start drugs – drugs that do not appear in the training data. DSGAT works on these limitations by employing the molecular graph of drugs instead of interaction graphs. This enables the model to learn embeddings for a cold start drug well.

The DSGAT architecture comprises of an encoder-decoder framework. To learn the embeddings from the molecular graphs of the drugs, a 3-layer GAT is employed by the encoder layer. The encoder uses it to learn representations from the similarity graph of the side effects as well. The decoder part then utilizes matrix factorization to predict the frequency of drug-side effect pairs from the learned representations of drugs and side effects. A novel weighted ε-insensitive loss function is utilized in this paper to tackle the sparsity problem and improve predictive performance.

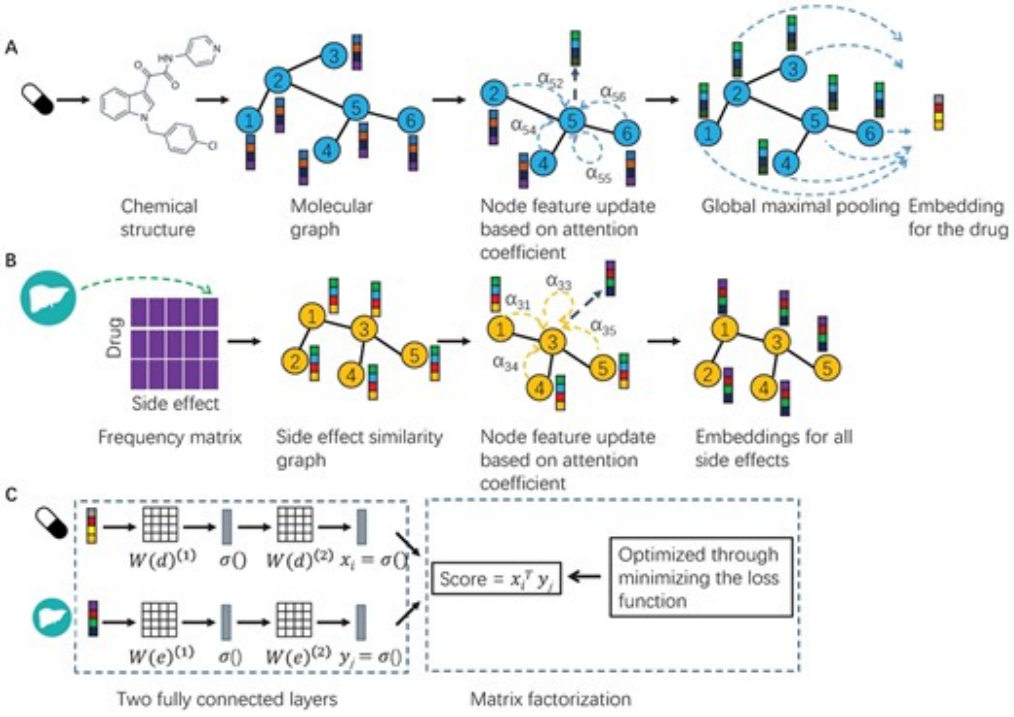


Figure 2.4.4: Architecture of DSGAT [38]

DSGAT worked on Galeano's benchmark dataset comprising 750 drugs, 994 side effects, and 37,071 known frequency items derived from the SIDER database. The experimental results obtained demonstrated performance improvement over most

of the existing models. The model achieved an RMSE score of 1.469 and MAE of 1.175.

DSGAT illustrated its robustness to alterations in chemical similarity between training and test sets, confirming its generalization power. The independent test on post-marketing side effects further validated its predictive power, indicating practical utility for drug risk-benefit evaluation.

2.4.5 NRFSE

In Wang et al.'s study [39], the authors introduced NRFSE, a new method based on neighborhood regularization to predict the frequencies of drug side effects using multi-view data. NRFSE takes the drug-side effect frequency matrix and decomposes it leveraging a class-weighted non-negative matrix factorization and unknown drug-side effect frequency pairs are modelled using drug-side effect pairs. NRFSE uses multi-view neighborhood regularization to integrate 3 drug attributes (side effect frequency, chemical structure, and Gene Ontology (GO) annotations) and two side effect attributes (frequencies across drugs and MedDRA terms). This method works by modelling certain similar drug and side effect pairs to have similar latent signatures.

The study utilizes a modified version of Galeano's benchmark dataset containing 34,604 known frequency items across 664 drugs and 994 side effects. It has less drugs because the drugs for which GO target information were not available were dropped. NRFSE's architecture includes non-negative matrix factorization on the drug side effect frequency matrix to predict the frequency scores and multi-view neighborhood regularization to refine embeddings for new drugs and side effects based on the three drug views and two side effect views.

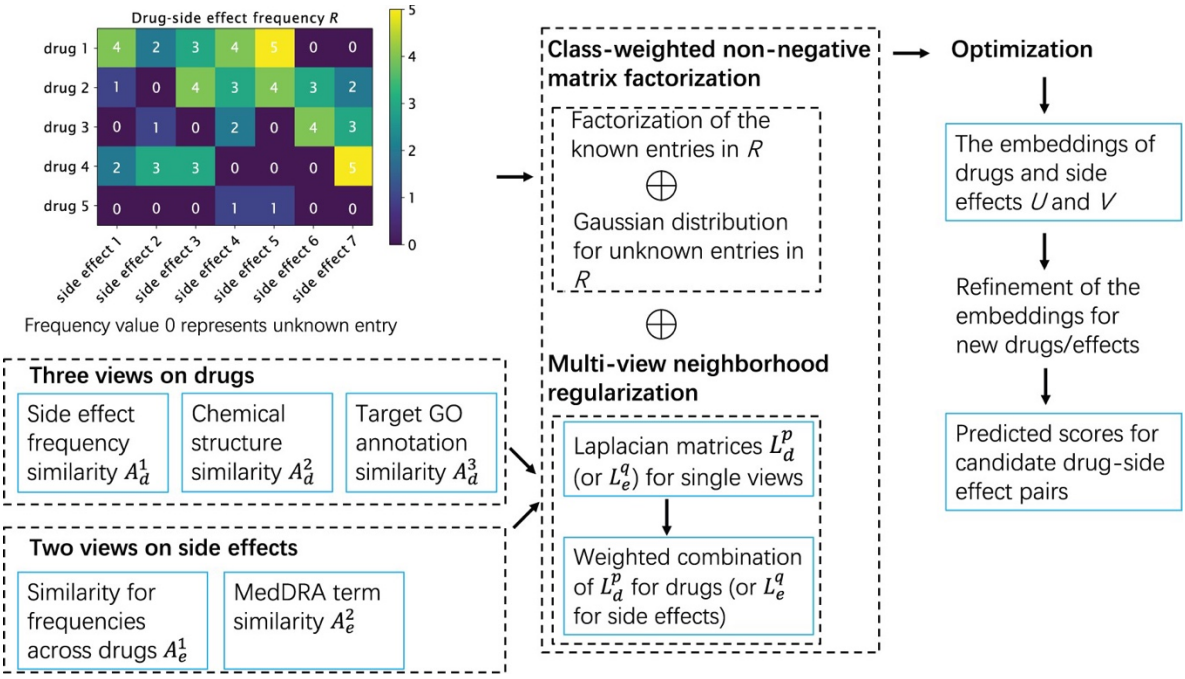


Figure 2.4.5: Architecture of NRFSE [39]

The model's effectiveness is validated through 10-fold cross-validation under different scenarios such as warm-start and cold-start. The results show that NRFSE significantly outperforms previous approaches in AUC and AUPR values. They got AUC scores of 0.898 and AUPR scores of 0.442. They also achieved RMSE scores of 0.1.378 and MAE scores of 1.142.

Furthermore, NRFSE ran an independent test on post-marketing side effects which illustrates the model's practical usefulness, accurately predicting the frequencies of side effects not included in the training data. Ablation experiments highlight the importance of integrating multiple data sources, and sensitivity analysis validates the robustness of NRFSE's hyperparameter settings. The study concludes that NRFSE provides a reliable tool for predicting drug side effect frequencies, with potential applications in guiding randomized controlled trials and enhancing drug safety assessments.

2.4.6 Park's Dual Representation Learning Model

In the recent Park et al.'s study [40], the authors came up with a novel deep learning model for drug side effect frequency prediction, Dual Representation Learning. They utilize protein target information. Previous methods used structural and chemical properties of drugs and drug-side effect associations for predicting frequencies of those pairs. Although the previous models performed well, the fact that rich information contained in drug target proteins were not utilized left room for improvement in the research. Park's model addresses these shortcomings by leveraging numerous features, including molecular graphs, fingerprints, chemical similarities, and especially, protein target information.

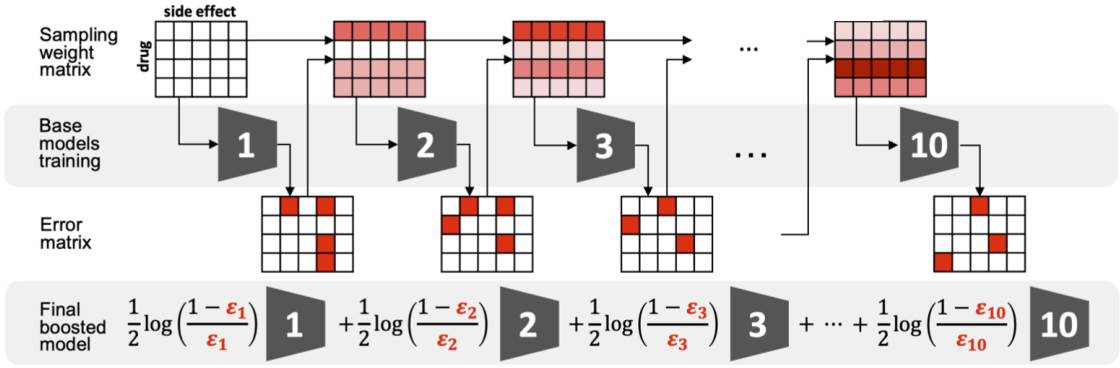


Figure 2.4.6: Architecture of Park's Model [40]

The study utilizes Galeano's benchmark dataset containing 37,071 drug-side effect pairs, including 750 drugs and 994 side effects. The proposed model takes all these heterogeneous features and integrates them to construct embeddings for the drugs and side effects, representing them in a common feature space. This is where the name dual representation learning comes from. This dual representation facilitates the prediction of frequencies of drug side effect pair of both known and unknown pairs.

The architecture of the model uses different techniques to encode different drug features. It utilizes a graph attention network (GAT) to encode the molecular graphs,

fully connected multi-layer perceptron for similarity matrices of drugs and protein target information, and network propagation to simulate the downstream effects of drug-target interactions. The side effect features MedDRA categorical vectors and Glove word embeddings are first concatenated and then embedded using a fully connected multi-layer perceptron. The embeddings are then combined using an Adaboost framework to improve predictive performance.

The model's effectiveness is validated through tenfold cross-validation, showing superior performance compared to existing models MGPred, SDPred and DSGAT. The study also demonstrates that incorporating drug protein target information with explicit targets achieve superior predictive performance. Ablation studies further confirm the value of each feature in the model, and independent tests performed on additional drugs validate the robustness and generalizability of the model.

2.4.7 HMMF

The paper by Liu et al. [53] introduces the Hybrid Multi-Modal Fusion (HMMF) framework, designed to predict the frequency of drug side effects. The HMMF framework leverages a multi-modal approach by integrating various types of data, including molecular structures, biomedical textual information, and attribute similarities of drugs and side effects. To achieve this, the model employs multiple encoders that understand these diverse data types and uses both coarse and fine-grained fusion strategies to integrate the multi-modal features effectively.

The methodology begins with biomedical semantic representation learning, where biomedical text information for drugs and side effects is collected from sources like Wikipedia and PubChem. A KV-PLM model is then used to learn contextual representations of these texts. Concurrently, molecular structure representation learning is carried out by converting SMILES sequences into undirected molecule graphs using RDKit, followed by the application of a graph

attention network (GAT) to extract representations from these molecular graphs. Additionally, attribute similarity learning is conducted by gathering drug-related data from databases like STITCH and CTD, constructing similarity matrices based on chemical structures and drug-disease associations.

The HMMF framework employs a sophisticated multi-modal fusion strategy to integrate the various representations. This involves projecting representations from the three modalities into a unified space, using coarse-grained fusion through element-wise product, and fine-grained fusion via outer product and convolutional neural networks (CNNs).

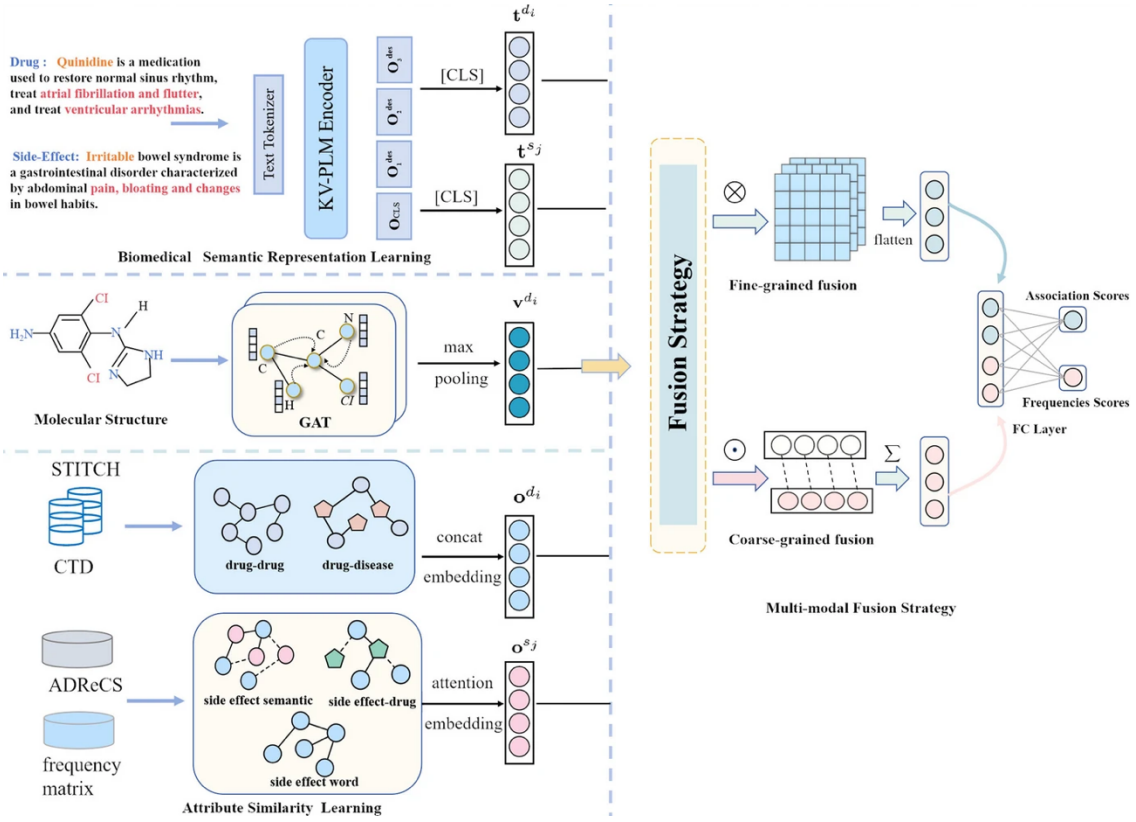


Figure 2.4.7: Architecture of HMMF [53]

The effectiveness of the HMMF framework is validated through extensive experiments, demonstrating superior performance compared to existing methods. The model shows exceptional capability in predicting drug-side effect frequencies,

particularly in cold-start scenarios where new drugs are introduced. This research underscores the potential of combining biomedical texts, molecular structures, and attribute similarities to improve the accuracy and generalizability of drug side effect predictions.

2.4.8 Summary of Key Literature and Limitations

The key literature on drug side effect frequency prediction discussed above are summarized with their significant contribution and limitations in this section. All the papers utilize the dataset curated by Galeano [30] and Zhao [31] with different papers incorporating different types of features. The summarized table is as follows:

Model	Technique Used	Contribution	Limitation
Galeano's Model [30]	Non-Negative Matrix Factorization	Created the benchmark DSF dataset and the first paper to look into DSF prediction task.	Cannot be adopted to predict side effect frequencies of cold start drugs as they solely used DSF matrix.
MGPred [31]	Graph Attention Model	Incorporated multi-view data for both side effects and drugs – similarity information, known frequency distribution, and word embedding	Drugs that are not included in the training set will not have an edge in the constructed heterogenous graph thus will not be able to predict the side effect frequencies of new drugs.
SDPred [37]	Similarity Based Deep Learning Approach	Introduced the ability to predict the frequency of cold start drugs. They also utilized 10 different	Overly dependent on complete similarity information of new drugs.

		drug similarities and 4 different side effect similarities such as chemical structure similarity, target and word similarity, semantic similarity, etc.	
DSGAT [38]	Graph Attention Model	Utilizes molecular graphs to learn representations of the drugs. Ability to predict the frequency of side effects for cold start drugs.	Does not take into consideration all the relevant information of drugs such as drug target protein information.
NRFSE [39]	Neighborhood Regularization	They utilized multi-view data - 3 views for drugs – side effect frequency, chemical structure, and Gene Ontology annotation of drug target. They also utilized frequencies across drugs, and MedDRA terms.	Does not take into consideration all the relevant information of drugs such as drug target protein information and semantic text description.
Park's Model [40]	Deep Learning Based	Incorporated drug target information. Also, utilized all the heterogeneous features - chemical similarity between drugs, molecular graphs, fingerprints, and protein targets simultaneously.	Limited number of features. Just 4 drug features used to generate the embeddings.

		They used Glove word embeddings and MedDRA categories. They used Adaboost ensemble technique to integrate the features of drugs.	
HMMF [53]	Hybrid Multi-Modal Fusion Framework	Introduced concurrent multi-modal learning from molecular structure, semantic information, similarity features for drugs and semantic similarity and side effect semantic text descriptors for side effects. First paper to look into semantic text information as a drug and side effect feature.	Not the most effective representation model used to encode the multi-modal information

Table 2.4.1: Summary of Papers on DSF Prediction

2.5 Papers on Large Language Models

2.5.1 BERT

The paper by Devlin et al. [29] addresses significant limitations in previous language representation models, particularly unidirectionality, which constrained their ability to leverage context from both directions. Traditional models like

OpenAI GPT [28] used unidirectional architectures, limiting their effectiveness in tasks requiring full context, such as question answering and named entity recognition. The distinction between feature-based methods like ELMo [42], which required task-specific architectures, and fine-tuning approaches like GPT also posed challenges. ELMo, for instance, concatenated independently trained left-to-right and right-to-left models but failed to fully exploit bidirectional context.

BERT (Bidirectional Encoder Representations from Transformers) overcomes these limitations by pre-training deep bidirectional representations, allowing it to use context from both directions simultaneously. BERT employs two tasks - masked language model (MLM) and next sentence prediction (NSP) during pre-training, which enables it to capture richer contextual information and understand relationships between sentences. This approach allows BERT to achieve state-of-the-art performance across various NLP tasks with minimal task-specific modifications.

BERT is trained on large-scale corpora, including BooksCorpus [65] and English Wikipedia [66], to learn diverse language patterns. The pre-training involves MLM, where random tokens are masked and predicted using bidirectional context, and NSP, which helps understand sentence relationships. BERT's architecture uses WordPiece tokenizer [43] with a 30,000 token vocabulary, trained over 1,000,000 steps.

The paper's contributions include introducing BERT, a model that pre-trains deep bidirectional representations, employing MLM to capture bidirectional context, and using NSP to enhance sentence-pair understanding. BERT's simple and unified architecture allows for fine-tuning on various NLP tasks, leading to state-of-the-art results on eleven NLP benchmarks, including the GLUE benchmark [44] and SQuAD [45]. This demonstrates BERT's effectiveness in improving predictive performance and overcoming the constraints of previous models.

2.5.2 ChemBERTa-2

This paper by Ahmad et al. [48] introduces ChemBERTa-2, a transformer model based on RoBERTa [54], aimed at enhancing molecular property prediction through advanced pretraining on SMILES strings. Building on the original ChemBERTa [56], ChemBERTa-2 optimizes the pretraining process and significantly increases the dataset size to 77 million compounds from PubChem. This model leverages the principles of large-scale pretraining, similar to those used in natural language processing models like GPT-3 [49], to learn salient representations that can be fine-tuned for various downstream tasks.

ChemBERTa-2 employs two main pretraining strategies: Masked Language Modeling (MLM) and Multi-task Regression (MTR). MLM involves masking 15% of the tokens in each SMILES string and training the model to predict these masked tokens, thereby helping the model understand the context within the molecular representations. MTR, on the other hand, focuses on predicting 200 molecular properties calculated from SMILES strings using RDKit [50]. These strategies are applied to a large corpus of 77 million SMILES strings, making it one of the largest datasets used for molecular pretraining.

The study conducts an extensive hyperparameter search to optimize the model's performance. This involves varying configurations of hidden sizes, attention heads, dropout rates, and other parameters to ensure the model is effectively trained. For evaluation, ChemBERTa-2 is fine-tuned on several regression and classification tasks from the MoleculeNet benchmark suite [51], including datasets like BACE, Clearance, Delaney, Lipophilicity, BBBP, ClinTox, HIV, and Tox21. Performance metrics such as ROC-AUC for classification tasks and RMSE for regression tasks are used to assess the model’s effectiveness.

ChemBERTa-2 achieves competitive results on nearly all MoleculeNet tasks, outperforming existing architectures in several instances. The results highlight the benefits of pretraining on larger datasets, demonstrating that extensive pretraining

can lead to significant improvements in downstream molecular property prediction tasks. The study suggests that the improvements in pretraining directly translate to better performance on downstream tasks, emphasizing the importance of large-scale data and effective pretraining strategies.

In their discussion, the authors note that while ChemBERTa-2 shows promise, future work should involve benchmarking against other graph-based architectures and extending pretraining to even larger datasets. They also emphasize the need to understand the conditions under which datasets might benefit from pretraining, as this can further refine the model's application and performance.

This work demonstrates the importance of large-scale pretraining in developing robust and accurate molecular representations, paving the way for future advancements in molecular property prediction and related fields.

2.5.3 SimCSE

The paper by Gao et al. [41] addresses limitations in previous sentence embedding models by introducing SimCSE, a contrastive learning framework that improves the quality and effectiveness of sentence embeddings. Traditional NLP data augmentation techniques often degrade performance, and many models suffer from representation collapse and anisotropy, limiting their expressiveness. The SimCSE framework uses dropout as minimal data augmentation and employs both unsupervised and supervised contrastive learning to enhance embeddings.

In the unsupervised SimCSE, sentences are encoded twice with different dropout masks, treating the resulting embeddings as positive pairs, and using other sentences in the batch as negatives. This approach maintains semantic meaning while providing variation to improve embedding quality. In the supervised SimCSE, entailment pairs from Natural Language Inference (NLI) [67] datasets are used as positives and contradiction pairs as hard negatives, further enhancing alignment

and uniformity. They validated the quality of the embeddings using Spearman's Rank Correlation.

The authors used various datasets for training and evaluation, including NLI datasets for supervised learning and STS datasets for evaluation. They also explored the effectiveness of different supervised datasets like QQP [46] and Flickr30k [47]. The unsupervised SimCSE was trained using sentences from English Wikipedia.

Key contributions include the use of dropout noise for data augmentation, preventing representation collapse, and the use of a contrastive learning framework that pulls semantically similar sentences closer while pushing apart dissimilar ones. This method improves the uniformity and alignment of embeddings, leading to superior performance in semantic textual similarity tasks. The SimCSE framework demonstrates versatility, robustness, and significant advancements in the quality of sentence embeddings.

2.5.4 Angle-Optimized Text Embeddings

In their paper [55], Xianming Li and Jing Li address the limitations of traditional text embedding models, particularly the problem of vanishing gradients caused by the saturation zones of the cosine function used in optimization objectives. To overcome this, they introduce a novel model named AnglE, which optimizes text embeddings by focusing on angle differences in a complex space. This approach mitigates the adverse effects of the saturation zones in the cosine function, improving the model's ability to learn subtle distinctions between texts.

The researchers conducted experiments on both existing short-text Semantic Textual Similarity (STS) datasets and a newly collected long-text STS dataset from GitHub Issues. They demonstrated that AnglE outperforms state-of-the-art STS models, which often ignore the cosine saturation zone. The AnglE model utilizes multiple encoders to understand various data types, including molecular structures

and biomedical textual information, and integrates these features through coarse and fine-grained fusion strategies.

The study reveals that AnglE achieves superior performance across various tasks, including short-text STS, long-text STS, and domain-specific STS scenarios with limited labeled data. The results show that AnglE's angle optimization technique significantly enhances text embedding quality and the overall performance of semantic similarity tasks. The paper underscores the importance of considering angle optimization in the development of robust and effective text embedding models, especially in applications requiring high-quality semantic textual similarity.

2.5.5 Summary of Key Literature and Limitations

The key literature on LLMs discussed above are summarized with their significant contribution and limitations in this section. The summarized table is as follows:

Model	Pretraining Techniques	Tokenizer	Contribution	Limitation
BERT [29]	Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks on the BooksCorpus [65] and English	WordPiece Tokenizer	Introduced bidirectional transformers [27] by using just the encoder module. Revolutionized state-of-the-art in numerous NLP tasks.	Requires large number of data, overfits on small datasets. Computationally expensive. NSP task investigated to be less effective in later papers such as RoBERTa [54]

	Wikipedia [66]			
ChemBERTa-2 [48]	MLM with focus on SMILES strings pretrained on the PubChem dataset [34]	SMILES Tokenizer	Optimized BERT for generating high quality embeddings of chemical data, especially useful for downstream prediction tasks.	Limited generalization to non-chemical text. Effectiveness relies heavily on high-quality training data. Noisy data can cause the performance to suffer.
SimCSE [41]	Contrastive learning by pretraining on the Natural Language Inference (NLI) [67] dataset.	WordPiece Tokenizer	Improved sentence embeddings by focusing on learning representations that better capture semantic similarity between sentence pairs.	Dependent on the quality of augmentations; not as effective for domain-specific data without fine-tuning.
Angle Optimized Text Embeddings [55]	Contrastive Learning on Semantic Textual Similarity task with Angle Optimization	WordPiece Tokenizer	Enhanced text embeddings by optimizing angular relationships in the vector space, improving	ay require specific training strategies and tuning to achieve optimal results, potentially complex to

	Techniques on the NLI dataset [67]		semantic similarity tasks.	implement.

Table 1.5.1: Summary of Papers on LLM

CHAPTER 3

Proposed Methodology

3.1 Material and Data

For our research, we have used Galeano [30] and Zhao's [31] dataset. The frequencies were extracted from the SIDER [35] dataset. A section of the dataset is illustrated below:

Drug	Side Effect	Frequency
podophyllotoxin	inflammation	5
podophyllotoxin	pruritus	5
gadobutrol	dermatitis	3
gadobutrol	dyspnoea	3
gadobutrol	injection site pain	3
gadobutrol	headache	4
gadobutrol	dizziness	3
gadobutrol	nausea	3
gadobutrol	rash	3
quinapril	chest pain	4
quinapril	diarrhoea	4

Table 2.1.1 DSF Dataset

There are a total of 750 drugs and 994 side effects, with 37,741 frequency values of known drug side effect pairs. The dataset has 5 frequency values – 1 (very rare), 2 (rare), 3 (uncommon), 4 (frequent) and 5 (very frequent). For DSHF, the frequency values over 3 was converted to 1 and the rest as 0 [31].

The SMILES string of the drugs were obtained from DrugBank [36] and PubChem [34] by using the PubChem IDs obtained from SIDER [35]. A section of the SMILES string data is illustrated as follow:

Drug	SMILES Representation
betaine	C[N+](C)(C)CC(=O)[O-]
bupropion	CC(C(=O)C1=CC(=CC=C1)Cl)NC(C)(C)C
estradiol	CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O
mannitol	C(C(C(C(C(CO)O)O)O)O)O
N-acetylcysteine	CC(=O)NC(CS)C(=O)O

Table 3.1.2: Drug SMILES Dataset

The set of side effects were taken from the SIDER Database [35]. A section of the list of side effects is illustrated as follows:

Side Effects
abdominal discomfort
abdominal distension
abdominal pain
abdominal pain lower
abdominal pain upper

Table 3.1.3: Side Effect Dataset

3.2 Methodology

In this section, we will discuss the proposed methodology of our research extensively. The architecture of our model includes the following steps – 1) Feature Acquisition and Dataset Construction, 2a) Embedding Generation of drug SMILES strings and side effect names, 2b) Embeddings Generation of the biomedical semantic text information, 3) Fine-tuning the biomedical semantic text embeddings, 4) Concatenation of the drug embeddings and side effect embeddings and cosine similarity of the embeddings, and 5) Frequency Prediction using Multi-Layer Perceptron. The whole process is illustrated in Figure 3.2.1:

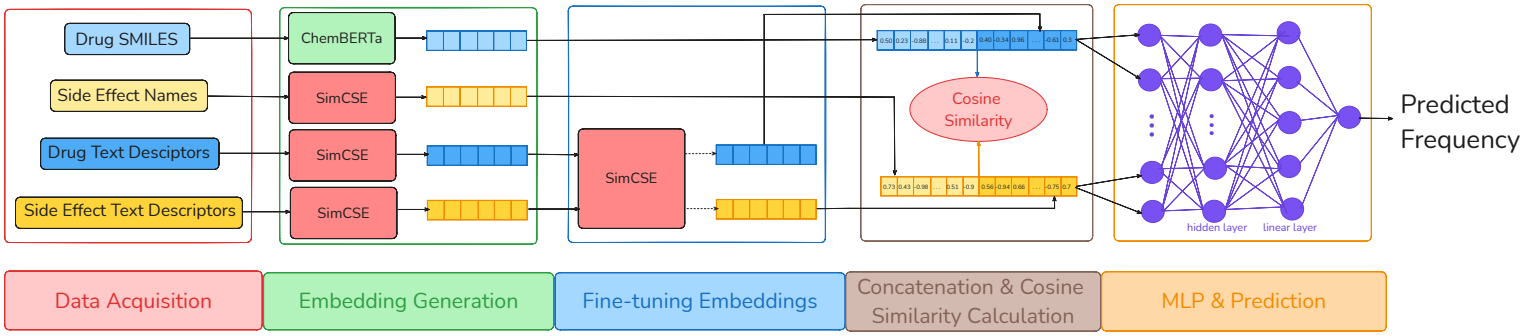


Figure 3.2.1: The proposed LLMPred Framework

3.2.1 Feature Acquisition and Dataset Generation

In the DSF network, we have two key entities – drugs and side effects. We acquire features of these entities such as biomedical semantic information of drugs and side effects from DrugBank and Wikipedia. The tables below illustrate a segment of the collected datasets.

Drug	Drug Description
Alprostadil	Alprostadil is produced endogenously and causes vasodilation by means of a direct effect on vascular and ductus arteriosus (DA) smooth muscle, preventing or reversing the functional closure of the DA that

	occurs shortly after birth. This results in increased pulmonary or systemic blood flow in infants. In infants, it is used for palliative, not definitive, therapy to temporarily maintain the patency of the ductus arteriosus until corrective or palliative surgery can be performed in neonates who have congenital heart defects and who depend upon the patent ductus for survival. In adults, it is used for the treatment of erectile dysfunction due to neurogenic, vasculogenic, psychogenic, or mixed etiology.
Phenytoin	Phenytoin is classified as a hydantoin derivative and despite its narrow therapeutic index, it is one of the most commonly used anticonvulsants. [A33595,A188832,A189219] Since it's introduction about 80 years ago, phenytoin has not only been established as an effective anti-epileptic, but has also been investigated for several other indications such as bipolar disorder, retina protection, and wound healing. [A188826,A188832] Clinicians are advised to initiate therapeutic drug monitoring in patients who require phenytoin since even small deviations from the recommended therapeutic range can lead to suboptimal treatment, or adverse effects.[A189219,A35884] Both parenteral and oral formulations of phenytoin are available on the market.[A189219]
Salbutamol	Salbutamol is a short-acting, selective beta2-adrenergic receptor agonist used in the treatment of asthma and COPD. It is 29 times more selective for beta2 receptors than beta1 receptors giving it higher specificity for pulmonary beta receptors versus beta1-adrenergic receptors located in the heart. Salbutamol is formulated as a racemic mixture of the R- and S-isomers. The R-isomer has 150 times greater affinity for the beta2-receptor than the S-isomer and the S-isomer has been associated with toxicity. This led to the development of levalbuterol, the single R-isomer of salbutamol. However, the high cost of levalbuterol compared to salbutamol has deterred wide-spread use

	of this enantiomerically pure version of the drug. Salbutamol is generally used for acute episodes of bronchospasm caused by bronchial asthma, chronic bronchitis, and other chronic bronchopulmonary disorders such as chronic obstructive pulmonary disorder (COPD). It is also used prophylactically for exercise-induced asthma. [Label, A174379,A174400]

Table 3.2.1: Biomedical Text Information of Drugs

Side Effect	Drug Description
Anemia	Anemia is a deficiency in red blood cells, commonly caused by chemotherapy, NSAIDs, and some antibiotics.
Bronchitis	Bronchitis is inflammation of the bronchial tubes, often caused by infections or irritants.
Depression	Depression is a mental health disorder characterized by persistent sadness and loss of interest, often requiring treatment.

Table 3.2.2: Biomedical Text Information of Side Effects

Using these data, we curate our final dataset to perform our target task of drug side effect frequency on. An example of a row of the final curated dataset is illustrated in Table 8.2.3:

Drug SMILES	Drug Description	Side Effect	Side Effect Description	Frequency
C1=CC=C(C=C1)CCCC(=O)O	Phenylbutyric acid is a fatty acid naturally produced by	Shock	Shock is a critical condition where blood circulation is insufficient to meet the body's	4

	colonic bacteria fermentation.		needs, often requiring emergency treatment.	

Table 3.2.3: Example of a row of the sample dataset

3.2.2 Embedding Generation

In our research, we have used two different BERT-based [29] LLM models to generate the embeddings required for our task – ChemBERTa-2 [48] and SimCSE [41]. We have used ChemBERTa-2 to generate embeddings of the SMILES string representation [20] of drugs and used SimCSE to generate embeddings of the side effect names. We also generated embeddings of the biomedical text descriptions of the drugs and side effects using SimCSE.

3.2.2.1 Embedding SMILES String Using ChemBERTa-2

The SMILES string representation of drugs was done via the BERT-based model specifically pretrained on SMILES strings – ChemBERTa-2 [48]. The general architecture of the model is illustrated in Figure 3.2.2:

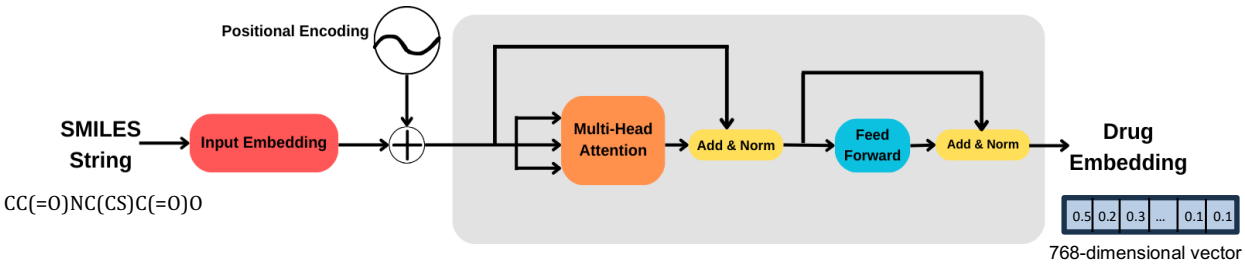


Figure 3.2.2: ChemBERTa-2 Architecture

The SMILES string is first broken down into tokens. Tokens are decomposed units of sentences or strings which can constitute words, sub-words or even characters. Here, the SMILES strings are broken down into tokens using the SMILES tokenizer. SMILES tokenizer is the default tokenizer for ChemBERTa-2 and it

outperforms all other tokenizers for tokenizing SMILES string. For example, the SMILES string CC[N+](C)(C)Cc1ccccc1Br is tokenized into 'CC[N+]C', 'C[N+]C)', '[N+]C)C', 'C)C)', ')C)C', 'C)Cc', ')Ccc', 'Cccc', 'cccc', 'cccc', 'cccc', 'ccc6', 'cc6Br'. These tokens are then mapped to integers called token ids. These token ids are then mapped to 768 length vectors called token embeddings.

Token embeddings are numerical vector representation of tokens in an input sequence. For example, the token 'cc6Br' might be mapped to a vector of length 768 that looks like [0.12, 0.34, -0.88, … , 0.95, 0.27]. This mapping is done based on the lookup table constructed during pretraining. This step is essential to BERT's architecture as it converts discrete tokens into continuous vectors enabling it to capture semantic information.

After the token embeddings are generated, positional embeddings are generated. Positional encoding is a crucial element of the BERT model which enables the model to obtain information regarding the position of a token in a sequence relative to all the other tokens. This is important as BERT processes input tokens parallelly rather than sequentially so without positional encoding, it would not have any way to keep track of the order of tokens [29]. Positional encoding is calculated using sine and cosine functions for even and odd positions of the embedding:

$$PE_{(position,2i)} = \sin\left(\frac{position}{10000^{\frac{2i}{d_{model}}}}\right) \quad (1)$$

$$PE_{(position,2i+1)} = \cos\left(\frac{position}{10000^{\frac{2i}{d_{model}}}}\right) \quad (2)$$

Here, *position* refers to the position of the token in a sequence. *i* is the index of the specific dimension of the 768-length vector and d_{model} refers to the dimension, or the length of the embedding of a model, which, in this case is 768. This can be variable based on the model - BERT-large based models have 1024 dimensions where BERT-base based models have 768 dimensions. These positional embeddings

are then added to the token embeddings. This creates a 768-length embedding vector X to be passed forward to the encoder.

$$X = [x_{1,}x_2, \dots, x_n] \tag{3}$$

This embedding is then passed onto the encoder block which is the same encoder block that is found in transformers [27]. The encoder block is where the refinement of the embedding vector X takes place so that the representation is richer and context-aware.

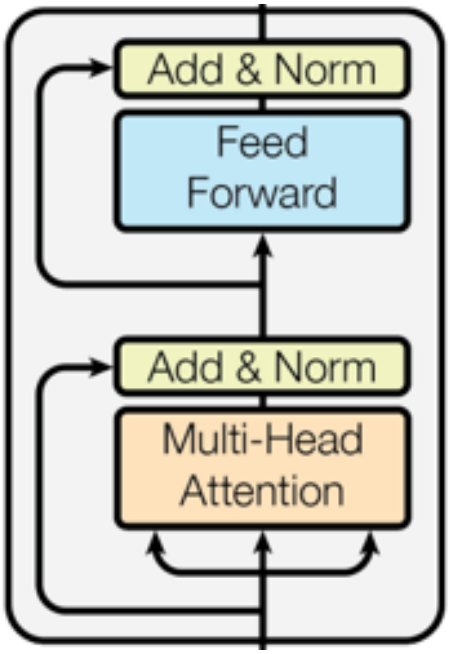


Figure 3.2.3: Encoder Block of BERT [29]

It consists of a multi-head self-attention block through which the embeddings are passed. The embeddings are linearly transformed into 3 different matrices – Query (Q), Key (K), and Value (V). These are the 3 inputs that are going into the Multi-Head Attention module illustrated above. These are calculated as:

$$Q = XW_Q \tag{4}$$

$$K = XW_K \tag{5}$$

$$V = XW_V \tag{6}$$

Here, W_Q, W_K and W_V are learnable weight matrices.

The self-attention scores are computed as follows:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Here, d_k is the dimensionality of the key vectors.

The weight of the values based on the query and key is calculated with:

$$\alpha_{ij} = \frac{e^{\frac{(q_i k_j)}{\sqrt{d_k}}}}{\sum_{j=1}^{n} e^{\frac{(q_i k_j)}{\sqrt{d_k}}}} \tag{8}$$

Here, α_{ij} is the weight of the vector between the *i-th* query and the *j-th* key.

The output from the multi-head attention is then calculated:

$$output_i = \sum_{j=1}^{n} \alpha_{ij}\, v_j \tag{9}$$

The output is then taken, and the residual connection is added to it to mitigate the problem of vanishing gradient. This summed output is normalized to ensure a stable mean and variance of the activations:

$$Norm(x) = \frac{x - mean(x)}{std(x) + \in} \tag{10}$$

Here, ϵ is a constant to prevent division by 0.

These embeddings are then passed onto a fully connected position wise feed forward network that is applied to each position in a sequence independently:

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \quad (11)$$

W_1 and W_2 are weight matrices, and b_1 and b_2 are bias terms and x is the input to the feedforward neural network. This will be added with the residual connection again and normalized.

This process is repeated over the 12 encoder blocks refining the embeddings in a way as to encode the contextual information in it. The output from the final encoder block will be the final embedding of the SMILES string.

3.2.2.2 Embedding Side Effect Names Using SimCSE

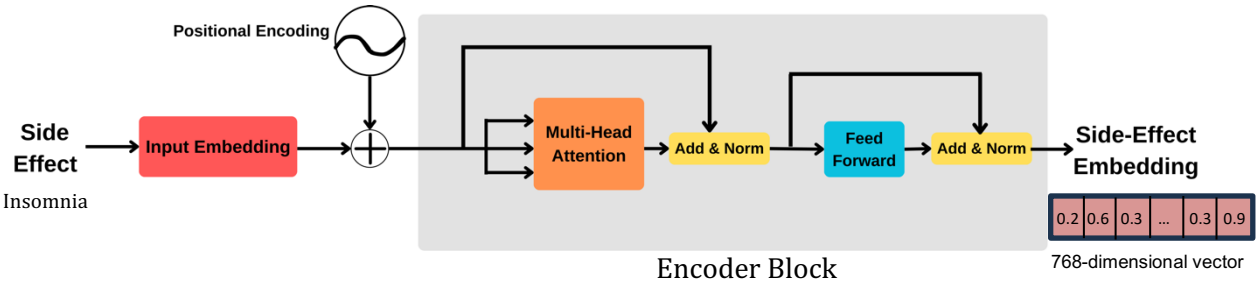


Figure 3.2.4: SimCSE Architecture

SimCSE [41] works on the exact same BERT architecture as ChemBERTa described in the previous section. It generates a 768-length vector for each side effect name. The key difference between ChemBERTa and SimCSE is the concept of contrastive learning employed during its pretraining. It is pretrained on Natural Language Inference (NLI) datasets where there are two types of pairs of sentences - positive pairs where sentence pairs entail each other or are similar, and hard negative pairs where the sentences are contradictory. Figure 3.2.4 and Figure 3.2.5 illustrates examples of how the embeddings of similar and dissimilar pairs of sentences are refined.

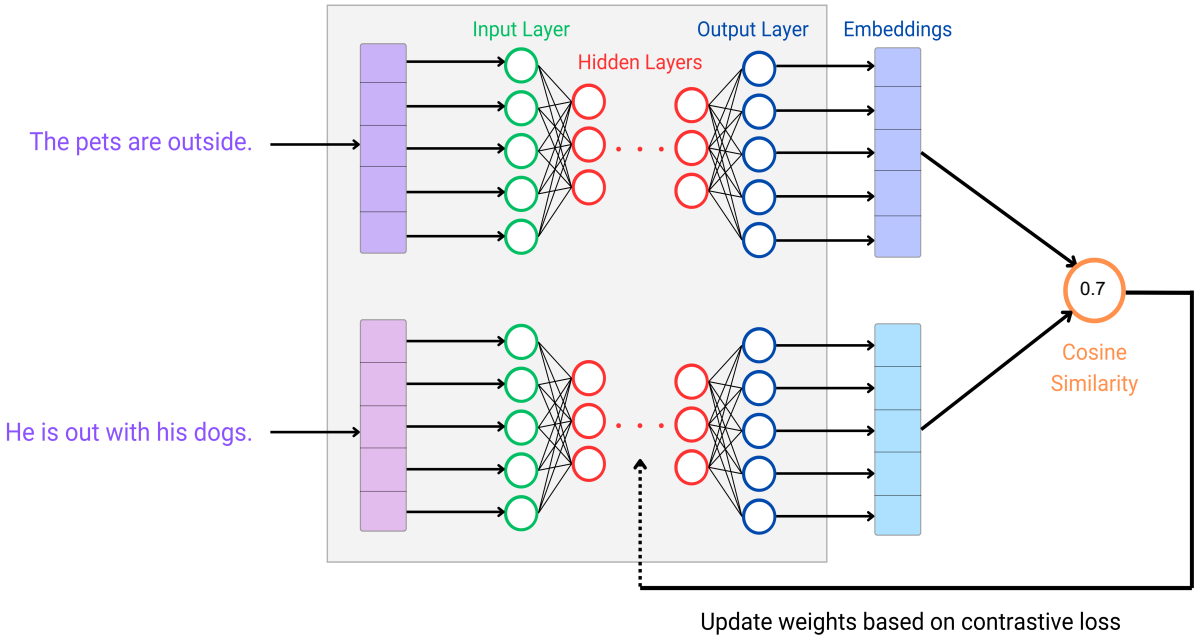


Figure 3.2.5: Pretraining of SimCSE on similar pairs of sentences

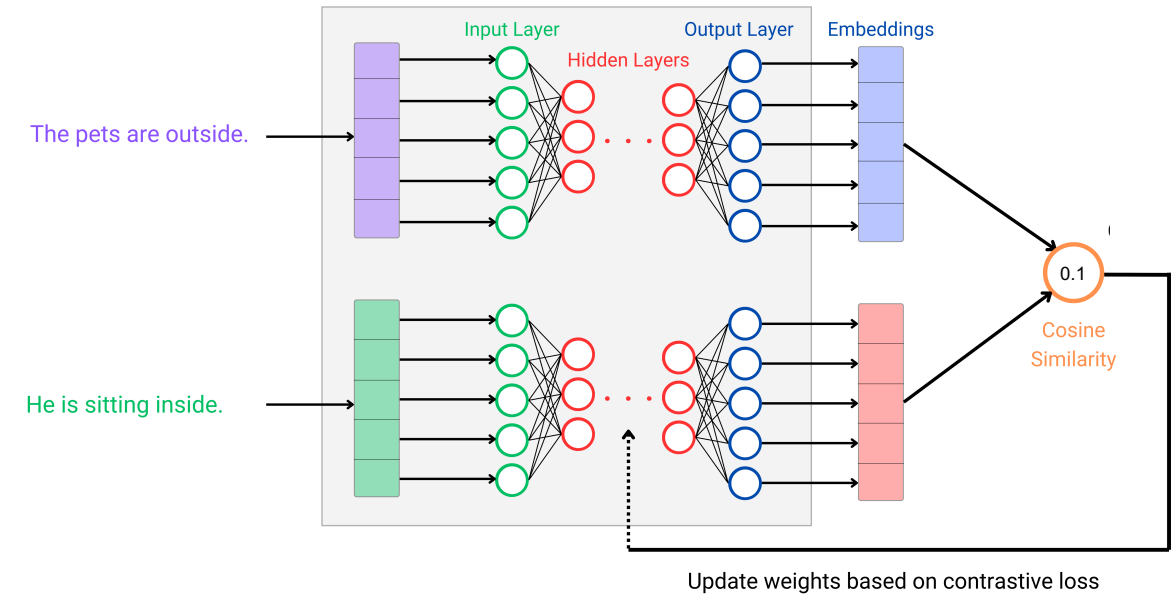


Figure 3.2.6: Pretraining of SimCSE on dissimilar pair of sentences

The pair of sentences are being input into the neural network with default weights and biases and embeddings are generated for the two separate sentences. Cosine similarity of the two sentences is calculated:

$$cosine\ similarity(h_1, h_2) = \frac{\sum_{i=1}^{d} h_{1i} \cdot h_{2i}}{\sqrt{\sum_{i=1}^{d} {h_{1i}}^2} \cdot \sqrt{\sum_{i=1}^{d} {h_{2i}}^2}} \quad (12)$$

Here, h_1 and h_2 are the embedding of the two sentences, and d denotes the dimensionality of the embeddings, which, in this case, is 768.

Based on the similarity of the sentences and cosine similarity, if the sentences are similar as per Figure 3.2.4, the cosine similarity value will be higher so the weights and biases of the neural network will be updated based on contrastive loss function in a way as to maximize the cosine similarity, and if the sentence pair is dissimilar, as seen in Figure 3.2.5, the cosine similarity value would be lower and the weights and biases would be updated based on contrastive loss function in a way as to minimize the cosine similarity. The contrastive loss function is:

$$\mathcal{L}_{constrastive} = -\log\left(\frac{e^{\frac{cosine\_sim(h_1,h_2)}{\tau}}}{\sum_{i=1}^{2N} e^{\frac{cosine\_sim(h_1,h_i)}{\tau}}}\right) \quad (13)$$

Here, h refers to the embeddings, τ is a temperature hyperparameter and $2N$ is the batch size.

3.2.2.3 Embedding Biomedical Text Descriptors Using SimCSE

Both the biomedical text descriptors of the drugs and the side effects are embedded using SimCSE [41], explained in Section 3.2.2.2. The biomedical semantic text descriptor for the drugs is converted into a 768-length vectors. The biomedical semantic text descriptor for the side effects is also embedded into a 768-length vector.

3.2.3 Fine-Tuning the Embeddings of the Biomedical Text Descriptors

The two different embeddings for the biomedical semantic text descriptor of drugs and side effects are then fine-tuned using combinations of different loss functions in order to improve the representation of the text descriptors better and thus improve the performance.

The architecture of the fine-tuning flow is illustrated in Figure 3.2.6:

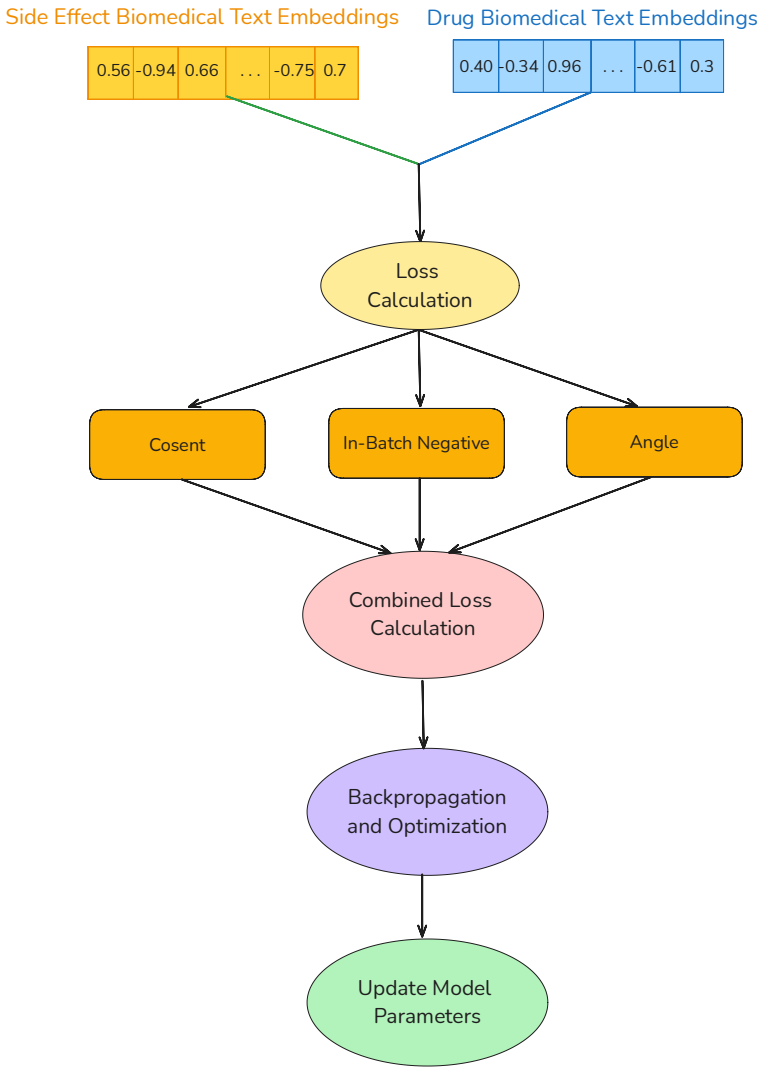


Figure 3.2.7: Fine-Tuning Framework for SimCSE

This fine-tuning architecture uses a combination of loss functions discussed by Li and Li in the paper AnglE [55]. It discusses combining three loss functions – CoSENT [52], In-Batch Negative [57], and AnglE [55].

3.2.3.1 CoSENT Loss

According to CoSENT [52], the primary goal is to maximize the cosine similarity for positive sample pairs and minimize it for negative sample pairs. This principle underpins the design of the cosine function used in generalization and end-to-end optimization of similarity between representations. By ensuring that the similarity score for positive pairs is higher than that for negative pairs, CoSENT addresses the need for more nuanced similarity measures beyond the binary classification approach typical in Natural Language Inference (NLI). The use of a temperature hyperparameter, τ, helps to control the degree of generalization, thereby reducing overfitting and enhancing the model's performance on unseen data. The equation for $\mathcal{L}_{cosent}$ is:

$$\mathcal{L}_{cos} = \log\left[1 + \sum_{f(X_{drug},\ X_{side}) > f(X_{drug2},\ X_{side2})} e^{\frac{\cos(X_{drug2},\ X_{side2}) - \cos(X_{drug},\ X_{side})}{\tau}}\right] \quad (14)$$

Where:

- X_{drug} and X_{side} - embeddings of the biomedical semantic text descriptors of drugs and side effects.
- $s(X_{drug},\ X_{side})$ - labeled frequency scores.
- cos (.) – cosine similarity function
- τ – temperature hyperparameter

3.2.3.2 In-Batch Negative Loss

Contrastive models typically generate positive samples through data augmentation techniques. Within a batch, semantically similar sentences that aren't explicitly labeled as positive samples might end up as in-batch negatives. To address

this, supervised positive samples are utilized, ensuring that truly similar sentences are correctly identified as positives.

The in-batch negative loss function [57] then leverages these in-batch negatives to calculate the loss. This loss function encourages the model to learn embeddings such that positive samples are closer together and negative samples are farther apart in the embedding space. This approach helps in effectively distinguishing between similar and dissimilar sentences, improving the overall quality of the learned sentence embeddings. The equation for the In-Batch Negative Loss is:

$$\mathcal{L}_{ibn} = -\sum_{b}\sum_{i}^{m} log\left[\frac{e^{\frac{\cos\left(X_{b_i}, X_{b_i}^{+}\right)}{\tau}}}{\sum_{j}^{N} e^{\frac{\cos\left(X_{b_i}, X_{b_j}^{+}\right)}{\tau}}}\right] \quad (15)$$

Where:

- b – the batch number
- X_{b_i} and $X_{b_i}^{+}$ - embeddings of the positive samples of a drug and side effect biomedical text descriptor.
- m – number of positive pairs in a batch.
- cos (.) – cosine similarity function
- τ – temperature hyperparameter

3.2.3.3 Angle Loss

Traditional cosine similarity functions used in In-Batch Negative loss [57] can encounter saturation zones, which can hinder the optimization process by causing gradients to vanish, leading to suboptimal model performance. By focusing on optimizing the angle differences in complex space, the AnglE loss [55] addresses these saturation issues. Instead of relying solely on cosine similarity, AnglE introduces a complex space where the angles between vectors are optimized. The complex space representations of text embeddings X_{drug} and X_{side} are split into real

and imaginary parts X_{real} and X_{img}. The representations are defined as $z = a + bi$ and $w = c + di$ with $a = X_{drug,real}$, $b = X_{drug,img}$, $c = X_{side,real}$ and $d = X_{side,img}$.

The angle difference of $\Delta\theta_{zw}$ is:

$$\Delta\theta_{zw} = abs\left[\frac{(ac + bd) + (bc - ad)i}{\sqrt{(c^2 + d^2)(a^2 + b^2)}} \text{ x } \frac{\sqrt{(a^2 + b^2)}}{\sqrt{(c^2 + d^2)}}\right] \quad (16)$$

With the angle difference $\Delta\theta_{zw}$ calculated, the angle loss $\mathcal{L}_{angle}$ is formulated as:

$$\mathcal{L}_{angle} = \log\left[1 + \sum_{f(X_{drug}, X_{side}) > f(X_{drug2}, X_{side2})} e^{\frac{\Delta\theta_{drug\_side} - \Delta\theta_{drug2\_side2}}{\tau}}\right] \quad (17)$$

Where:

- $f(X_{drug}, X_{side})$ – the labeled frequency values
- $\Delta\theta_{drug\_side}$ and $\Delta\theta_{drug2\_side2}$ – normalized angle differences between high and low frequency pairs.
- τ – temperature hyperparameter

3.2.3.4 Combined Loss

After the calculation of all the loss values, the combined loss is calculated as:

$$\mathcal{L}_{combined} = w_{cosent} \times \mathcal{L}_{cosent} + w_{ibn} \times \mathcal{L}_{ibn} + w_{angle} \times \mathcal{L}_{angle} \quad (18)$$

Here, w_{cosent}, w_{ibn} and w_{angle} are weights for each of the loss functions.

3.2.4 Embedding Concatenation and Cosine Similarity

After the generation of the 4 embeddings vectors of drug SMILES representation, side effect names, biomedical text descriptor of drugs, and biomedical text descriptor of side effects, each of length 768, the 2 drug embeddings and the 2 side

effect embeddings are concatenated to create a 1536 length vector representing the drugs and a 1536 length vector representing the side effects.

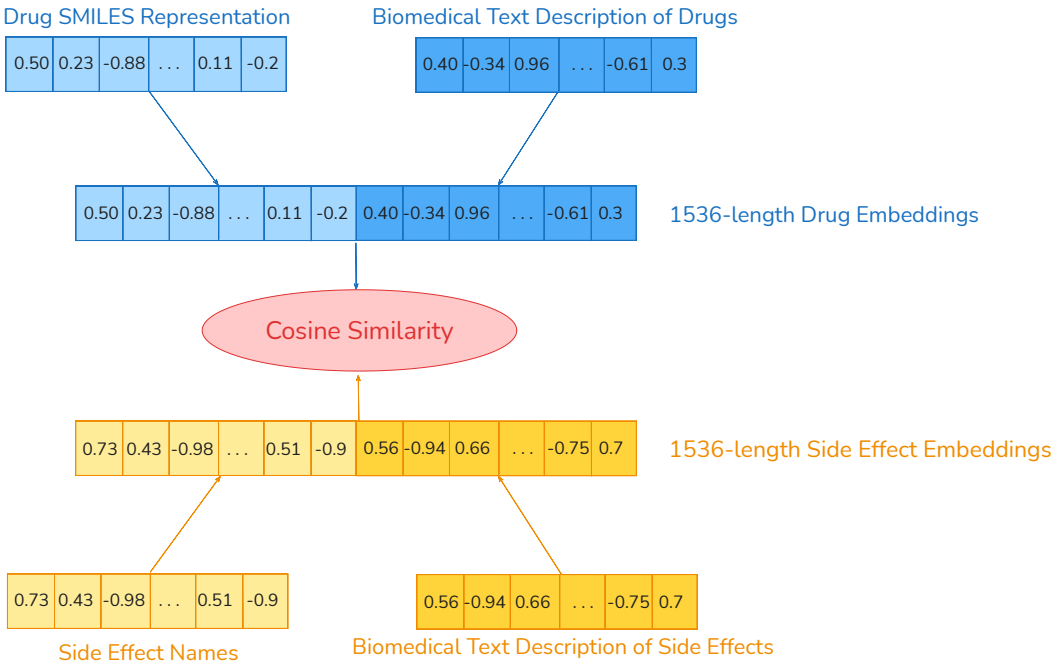


Figure 3.2.8: Concatenated Embeddings of Drugs and Side Effects

This is followed by the calculation of cosine similarity of the two concatenated embeddings of drugs and side effects.

3.2.5 Frequency Prediction using Multi-Layer Perceptron

The concatenated embeddings are then passed on to a multi-layer perceptron (MLP) for classification. The MLP has 4 layers – an input layer where the embeddings are being fed to the model as input, a fully connected hidden layer with the activation function ReLU to introduce nonlinearity, a linear layer for the 5 classes to output the raw scores (logits), and a final softmax function to convert the logits into probabilities that enables the output to be interpreted as the likelihood of each class thus predicting the frequency class.

CHAPTER 4

Computational Experiments

4.1 System Configuration

We have conducted the experiments on the cloud GPU rental platform, RunPod [58] using the specifications listed in Table 4.1.1:

Spec	Detail
Processor	Intel ® Core™ i9-14900k, 3.2GHz, 24 Core(s)
RAM	188GB 5600MHz DDR5
OS	Ubuntu Linux
GPU	NVIDIA ® RTX 6000 ADA 48GB

Table 4.1.1: System Configuration Details

4.2 Dataset

We conducted the experiments on Galeano [30] and Zhao's [31] dataset with additional biomedical drug and side effect semantic text information for DSF prediction. We converted the same dataset with the frequency labels set to 0 if the labels are between 1 to 3, and 1 otherwise for DSFH prediction [31]. It is to be noted that all the state-of-the-art models rely on this dataset.

4.3 Experiments

We ran all the computational experiments using python3.10.1, pytorch2.3.1, transformers4.42.4, numpy1.26.4 and sklearn1.3.2 and other packages.

We started off by aggregating the data into a unified table – drug SMILES, drug semantic text descriptor, side effect name, side effect semantic text descriptor, and the frequency label. We then proceeded to embed each of the 4 features. We used the transformers library to import ChemBERTa-2 and SimCSE. After obtaining the embeddings, we fine-tuned a SimCSE model using the semantic text descriptors' embeddings, and as our loss functions, we used the combination of loss functions Cosent loss, In-batch Negative loss, and AnglE loss. We then concatenated the drug embeddings and side effect embeddings and calculated the cosine similarity between the concatenated embeddings using the sklearn library to compute the SCC. We then passed the embeddings on to an MLP constructed using the pytorch library to obtain the frequency value predictions.

With the hyperparameters listed in Table 10.4.1, we performed 10-fold cross validation. In each of the 10 iterations over all the folds, we iteratively used 9 of the folds for training and 1 of the folds for testing. This ensures that the results that we obtain are trustworthy and can work well on unseen and independent dataset. Also, this ensures that our model is robust, generalizes well and prevents overfitting.

Due to the size of the LLMs being very large with over a 100 million parameters trained for every single model along with the 10-fold cross validation, the experimental process was very resource and time intensive. Therefore, we experimented using the default values of the hyperparameters suggested by the previous literature.

4.4 Hyperparameter Tuning

Hyperparameter tuning is a crucial aspect for getting a machine learning model to perform well. It is the selection of ideal values for hyperparameters. We chose the values of these hyperparameters based on the usage guidelines of the standard fine-tuning processes in different papers The hyperparameters that we used in this research are as follows:

4.4.1 CoSENT Tau

CoSENT tau adjusts the smoothness of the Softmax function by scaling the differences between the cosine similarities of the pairs. A lower tau makes the model more confident, which could lead to overfitting, while a higher tau produces softer probability distributions across the pairs. We used a value of 20.

4.4.2 In-Batch Negative Tau

In-Batch Negative Tau adjusts the differences in similarity measures, helping to differentiate between negative and non-negative pairs within the batch. We used a value of 20 as the In-Batch Negative Tau.

4.4.3 Angle Tau

Angle Tau scales the different between the similarities in angle between pairs of sentences. We used a value of 1 for it.

4.4.4 Weights of the Loss Functions

These weights w_{cosent}, w_{ibn} and w_{angle} refer to the weights of the individual loss functions in the combined loss that controls how much a specific loss will impact the

overall combined loss. We used values of 1 for all of them in this research giving each of them equal impact.

The values of the hyperparameters are listed in Table 4.4.1:

Hyperparameters	Values
CoSENT tau	20
In-Batch Negative Tau	20
Angle Tau	20
w_{cosent}	1
w_{ibn}	1
w_{angle}	1

Table 4.4.1: Hyperparameter Values

4.5 Training Parameters

Training parameters are parameters that influence the training process and ultimately affect the performance of the model. Unlike model parameters, which are learned by the model during training (such as weights in a neural network), training parameters are set before training begins and are not learned from the data but rather set by the users. We discuss the training parameters that we used and the values that we chose for each. We chose the values of these parameters based on the usage guidelines of the standard fine-tuning processes in different papers.

4.5.1 Batch Size

Batch size is a crucial training parameter that refers to the number of training examples (or data points) that the model processes before updating its internal parameters (such as weights) during a single iteration of training. We used a value of 32 for the batch size.

4.5.2 Learning Rate

Learning rate is a training parameter that controls the size of the steps that the model takes while updating its internal parameters (such as weights) during the training process. We used a value of $2e^{-5}$ as the learning rate.

4.5.3 Epochs

Epoch refers to one complete cycle through the entire training dataset. During an epoch, the learning algorithm processes every example in the training dataset exactly once and updates the model's parameters based on the computed gradients. We trained our training set over 5 epochs.

4.5.4 Weight Decay

Weight decay is a regularization technique used in training machine learning models, particularly neural networks, to prevent overfitting. It works by adding a penalty to the loss function, which discourages the model's weights from growing too large. This penalty is proportional to the size of the weights and is controlled by a parameter known as the weight decay coefficient. We used a value of 0.01.

4.5.5 Optimizer

Optimizer is an algorithm that adjusts a model's internal parameters such as weights and biases in order to minimize the loss function. The loss function measures how well the model's predictions match the actual target values, and the optimizer's job is to find the set of parameters that results in the lowest possible loss. We used the Adam optimizer that computes adaptive learning rates for each parameter and has bias correction terms.

The values of the hyperparameters are listed in Table 4.5.1:

Training Parameters	Values
Batch Size	32
Learning Rate	$2e^{-5}$
Epoch	5
Weight Decay	0.01
Optimizer	Adam

Table 4.5.1: Training Parameters

4.6 Evaluation Metrics

The key metrics that we have used to evaluate the performance of our model for DSF and DSHF prediction will be discussed in this section.

4.6.1 Evaluation Metrics for DSF Prediction

For the DSF prediction problem, we employed the metrics Spearman's Rank Correlation Coefficient (SCC), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

4.6.1.1 Spearman's Rank Correlation Coefficient

The Spearman rank correlation coefficient, known as Spearman's rho (ρ), is a non-parametric statistic that measures the strength and direction of the association between two ranked variables. Unlike Pearson's correlation, which looks at linear relationships, Spearman's rho determines how well the relationship between two variables can be described using a monotonic function. This is particularly useful for data that doesn't follow a normal distribution or when dealing with ordinal data. To compute Spearman's rho, the data values are first converted to ranks, and then the differences between the ranks of each pair of variables are calculated.

The equation of SCC is formulated as:

$$SCC = 1 - \frac{6\sum {d_i}^2}{n(n^2 - 1)} \tag{19}$$

Where:

- d_i is the difference between ranks of each pair of observations.
- n is the number of observations.

4.6.1.2 Mean Absolute Error (MAE)

Mean Absolute Error calculates the average magnitude of the errors in a set of predictions, without considering their direction. It measures the average absolute difference between the predicted values and the actual values. MAE is useful in quantifying the prediction error of models in regression analysis. MAE is formulated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

Where:

- n – number of observations.
- y_i – actual value.
- $\hat{y}_i$ – predicted value.

4.6.1.3 Root Mean Square Error

Unlike the absolute average difference calculated in MAE, Root Mean Square Error (RMSE) measures the average magnitude of the errors between predicted values and actual values. RMSE is particularly useful because it provides a unified measure of predictive accuracy by combining both the variance and bias of the model's errors. RMSE is formulated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (21)$$

Where:

- n – number of observations.
- y_i – actual value.
- $\hat{y}_i$ – predicted value.

4.6.2 Evaluation Metrics for DSFH Prediction

For the binary classification problem DSFH prediction, we employed the metrics Accuracy, Precision, Recall, F1-score, Area Under Receiver Operator Curve (AUROC), and Area Under Precision Recall Curve (AUPRC). Before discussing these metrics, it is imperative to understand the terms that are used to calculate these metrics. These can be found from the confusion matrix. This is illustrated in Figure 4.6.1:

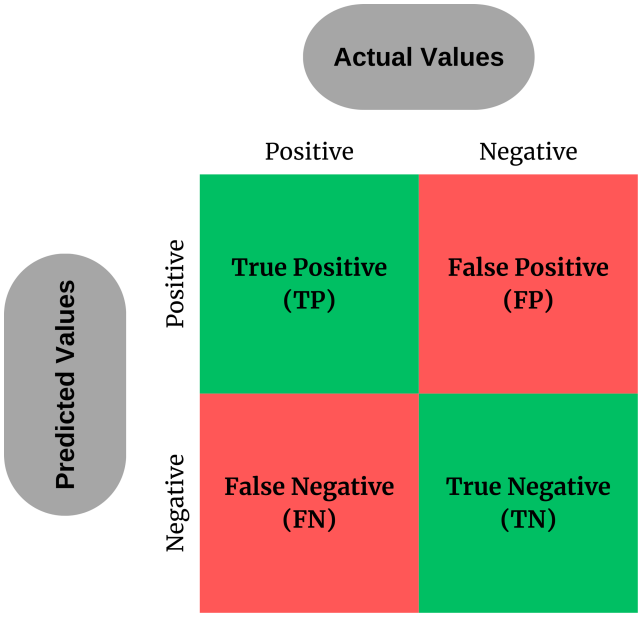


Figure 4.6.1: Confusion Matrix

True Positive: The total number of high frequency values predicted correctly.

True Negative: The total number of low frequency values predicted correctly.

False Positive: The total number of low frequency values incorrectly predicted as high frequency values.

False Negative: The total number of high frequency values incorrectly predicted as low frequency values.

4.6.2.1 Accuracy

Accuracy is defined as the ratio of the number of correct predictions made by the model to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{22}$$

4.6.2.2 Precision

Precision refers to the ratio of the true positive predictions (correctly identified positive instances) to all the instances that are predicted as positive by the model. In this case, this refers to the proportion of all the high frequencies that are correctly predicted to the total high frequency predictions.

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

4.6.2.3 Recall

Recall refers to the ratio of the true positive predictions (correctly identified positive instances) to all the instances that are actually true. In this case, this refers to the proportion of all the high frequencies that are correctly predicted to the total number of samples that are actually labelled as high frequency.

$$Recall = \frac{TP}{TP + FN} \tag{24}$$

4.6.2.4 F1-Score

The F1-Score shows the balance between the precision and recall. In this instance, it ensures that both the high frequency and low frequency side effects are comprehensively identified without a significant number of false positives or negatives.

$$F1 - Score = \frac{2 \ * Precision * Recall}{Precision + Recall} \tag{25}$$

4.6.2.5 Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC illustrates the tradeoff between recall and false positive rate. It measures how well a model can distinguish between two classes. It denotes class separability.

$$AUROC = \int_0^1 Recall\big(FPR^{-1}(x)\big)dx \tag{26}$$

Where:

$$False\ Positive\ Rate\ (TPR) = \frac{FP}{FP + TN} \tag{27}$$

4.6.2.6 Area Under the Precision Recall Curve (AUPRC)

AUPRC illustrates the tradeoff between precision and recall at different thresholds. It is used to comprehend the performance of a model's ability with regards to maintaining high precision while also receiving high recall.

$$AUPRC = \frac{1}{N}\sum_{d=1}^{N}\frac{TP}{FP + TP} \tag{28}$$

4.7 Result

The results obtained by our proposed LLMPred model for DSF and DSHF prediction is summarized in Table 4.7.1 and Table 4.7.2:

Metric	Value
SCC	0.7070
MAE	0.4071
RMSE	0.4346

Table 4.7.1: DSF Prediction Results

Metric	Value
Accuracy	0.8226
Precision	0.8752
Recall	0.8113
F1-Score	0.8421
AUROC	0.8248
AUPRC	0.8337

Table 4.7.2: DSHF Prediction

CHAPTER 5

Comparison and Analysis

5.1 DSF Prediction

The most recent research in the domain of DSF prediction have utilized different approaches such as matrix decomposition, graph attention networks, neighborhood regularization, multi modal fusion strategy, etc.

Model	SCC	MAE	RMSE
Galeano's Model	-	1.2980	0.9530
MGPred	0.065	0.4905	0.6521
SDPred	0.258	0.4212	0.5794
NRFSE	-	0.4330	0.5930
DSGAT	0.431	1.1750	1.4690
Park's model	0.438	1.0570	1.4071
HMMF	-	0.4216	0.5810
LLMPred	**0.707**	**0.4071**	**0.4346**

Table 5.1.1: Comparison of DSF Prediction

In the table above, we compare our model with the current state-of-the-art paper. Our model achieved superior metrics in all the metrics reported in previous state-of-the-art models. We also included earlier state-of-the-art models for

comparison. Figure 5.1.1, 5.1.2 and 5.1.3 visually illustrates the comparison of the performance of our model with all the baseline models:

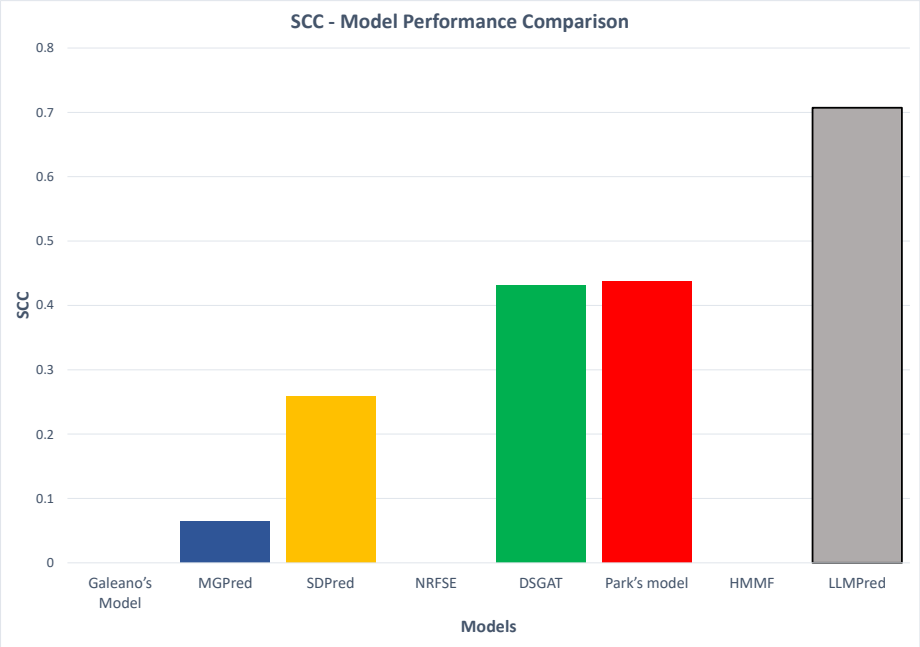


Figure 5.1.1: Comparing SCC

LLMPred achieved state-of-the-art results obtaining SCC scores of 0.707, a 61.46% increase over the second-best Park's model at an SCC score of 0.438. For the metric SCC, the higher the score, the better the performance of the model.

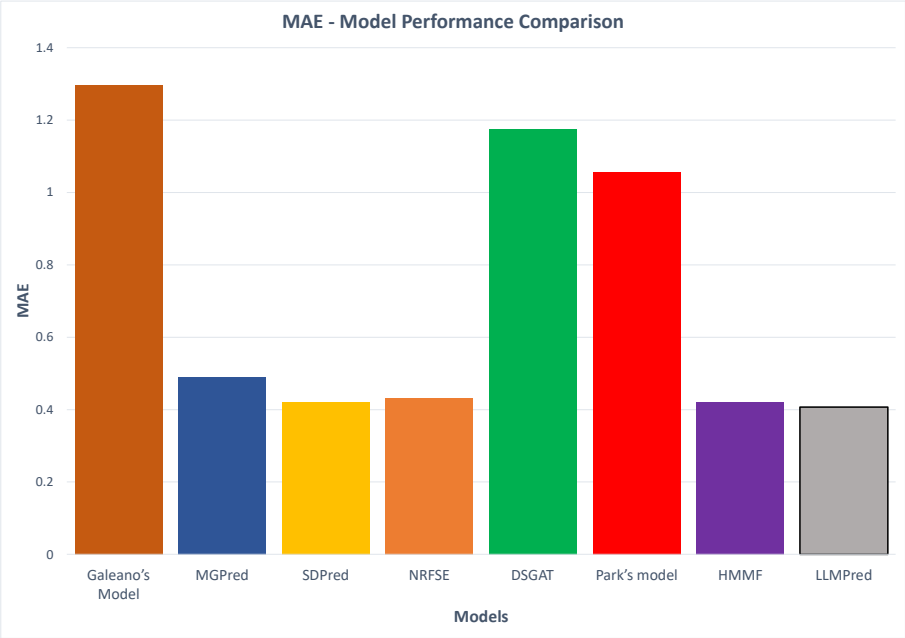


Figure 5.1.2: Comparing MAE

MGPred, SDPred, NRFSE and HMMF performs very similar to LLMPred but our model slightly outperforms the second best performing model by around 3.35%. In this case, the lower the MAE, the better the model.

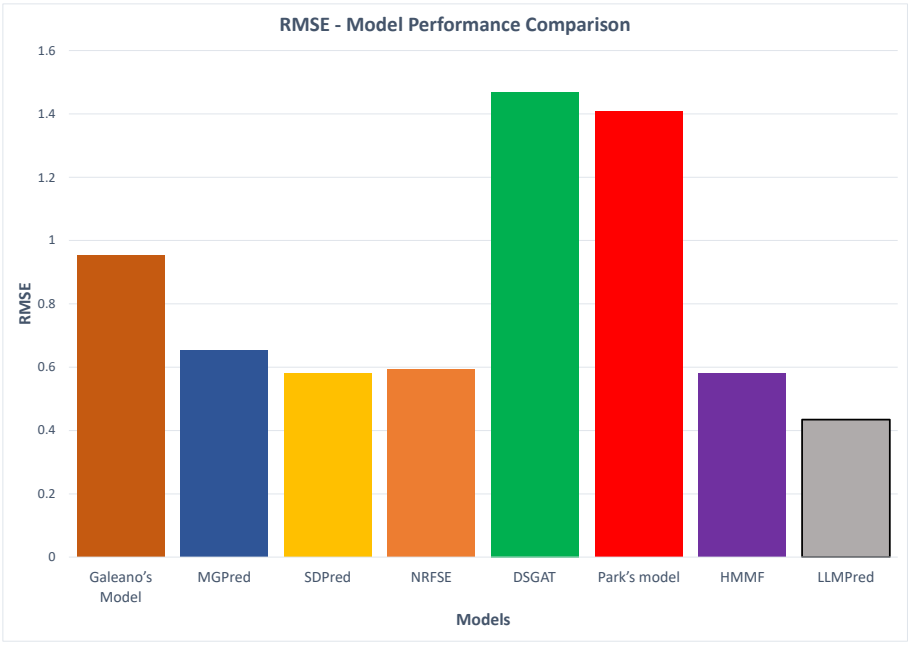


Figure 5.1.3: Comparing RMSE

Our model LLMPred showed superior performance compared to the other baseline models with an improvement from the best model by around 25%. The lower the RMSE, the better the performance.

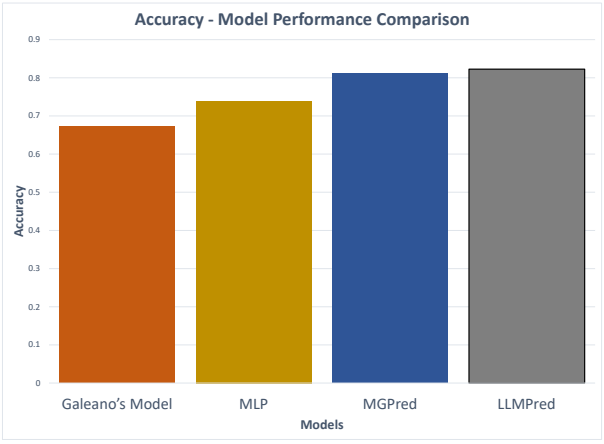
5.2 DSHF Prediction

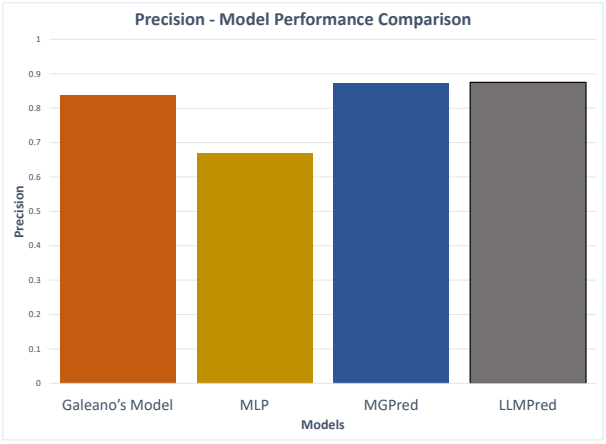
DSFH prediction task has not been explored in many of the papers that looked into frequency values. The most recent research that looked into high frequency prediction is MGPred [31].

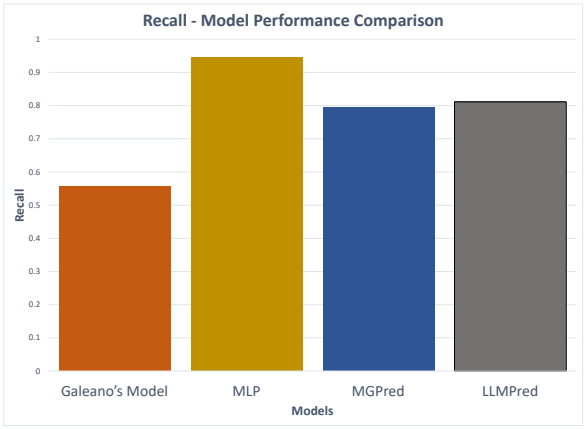
Model	Accuracy	Precision	Recall	F1	AUROC	AUPRC
Galeano's Model	0.6731	0.8391	0.5551	0.6682	0.7433	0.8279
MLP	0.7391	0.6696	**0.9441**	0.7835	0.8944	0.8886
MGPred	0.8107	0.8732	0.7942	0.8318	**0.9012**	**0.9197**
LLMPred	**0.8226**	**0.8752**	0.8113	**0.8421**	0.8248	0.8337

Table 4.1.2: Comparison of DSFH Prediction

In Table 5.1.2, we compare our model with the current state-of-the-art paper on DSHF prediction MGPred. Our model achieved metrics in all the metrics reported in previous state-of-the-art models. We also included earlier state-of-the-art models for comparison. Our model achieved the best accuracy along with a very balanced predictive power as demonstrated by the f1-score. The following figures visually illustrate the comparison of the metrics for the models. Figure 5.1.4 compares the accuracy, precision, recall and f-1 score of the models and Figure 5.1.5 compares the AUROC and AUPRC.







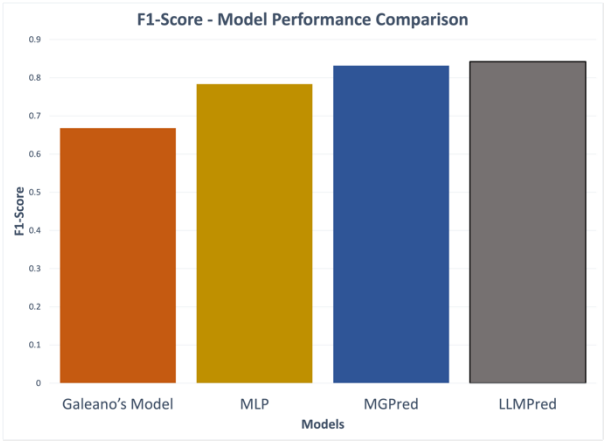
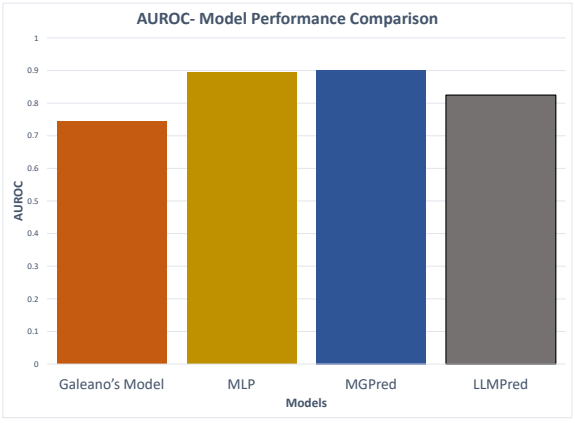


Figure 5.1.4: Accuracy, Precision, Recall and F1-Score



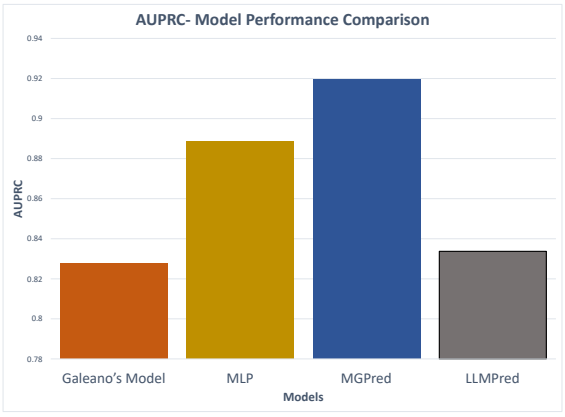


Figure 5.1.5: AUROC and AUPRC

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

In this research, we strived to develop a novel architecture based on similarity measures to generate embeddings using LLMs. We formulated the DSF and DSHF task from the DSF dataset curated in previous research efforts along with new data acquired from multiple sources. The purpose of this methodology is to predict the frequency of specific side effects for certain drugs.

The first step involves generating embeddings for the chemical structures of the drugs and the names of the side effects. This allows the model to capture the inherent similarities between different chemical compounds and their associated side effects.

In parallel, embeddings are generated from biomedical text information related to the drugs and side effects. This step leverages textual data to further enhance the embeddings with context-specific information derived from scientific literature and other textual sources.

After generating these embeddings, the research concatenates them, combining both the structural and textual information. The final step involves predicting the frequency of drug side effects using these combined embeddings using cosine similarity. This comprehensive approach aims to enhance the accuracy and

reliability of side effect predictions by integrating multiple data sources and types of information.

6.2 Limitation and Future Work

Despite obtaining significantly good results in the task of DSF, there are still some limitations of our research. We focused primarily on DSF prediction rather than drug side effect association prediction.

Also, we did not consider external environments such as genetic factors associated with a side effect occurring after a drug is taken. We also used a couple of attributes each for the drugs and side effects.

Furthermore, our research is concerned with side effects of single drugs rather than how a side effect can result from polypharmacy, or the effect of multiple drugs taken at once.

Based on our work, future research could focus on

- Incorporate more features to our dataset and extend it.
- Extend the model architecture for polypharmacy side effect and side effect frequency prediction.
- Carry out drug side effect association prediction.
- Harness the power of very large language models such as Llama 3 with 405 billion parameters.

REFERENCES

[1] The Editors of Encyclopedia Britannica, "Molecule," Encyclopedia Britannica, 30 Jul. 2024. [Online]. Available: https://www.britannica.com/science/molecule. [Accessed: 30-Jul-2024].

[2] Australian Government Department of Health and Aged Care, "What are drugs?" 2024. [Online]. Available: https://www.health.gov.au/topics/drugs/about-drugs/what-are-drugs. [Accessed: 30-Jul-2024].

[3] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nature reviews Drug discovery*, vol. 3, no. 8, pp. 673–683, 2004.

[4] N. Vargesson, "Thalidomide-induced teratogenesis: History and mechanisms," *Birth Defects Research Part C: Embryo Today: Reviews*, vol. 105, no. 2, pp. 140–156, 2015.

[5] Government of Canada, "Adverse reactions and side effects," 2024. [Online]. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/adverse-reaction-reporting/adverse-reactions-side-effects.html. [Accessed: 30-Jul-2024].

[6] Drugwatch, "Misplaced trust: FDA approval concerns," 2024. [Online]. Available: https://shorturl.at/cgX77 [Accessed: 30-Jul-2024]

[7] Government of Canada, "Clinical trials," 2024. [Online]. Available: https://www.canada.ca/en/health-canada/services/clinical-trials.html. [Accessed: 30-Jul-2024].

[8] D. B. Fogel, "Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review," *Contemporary Clinical Trials Communications*, vol. 11, pp. 156-164, Aug. 2018.

[9] F. Curtin and P. Schulz, "Assessing the benefit: risk ratio of a drug--randomized and naturalistic evidence," *Dialogues in Clinical Neuroscience*, 2011.

[10] T. Gandhi, D. Seder, and D. Bates, "Methodology matters. Identifying drug

safety issues: From research to practice," *International Journal for Quality in Health Care*, vol. 12, no. 1, pp. 69–76, 2000.

[11] Canadian Pharmacogenomic Network for Drug Safety, "Adverse Side Effects," [Online]. Available: https://cpnds.ubc.ca/about/. [Accessed: July 30, 2024].

[12] H. Le Louët and P. J. Pitts, "Twenty-First Century Global ADR Management: A Need for Clarification, Redesign, and Coordinated Action," *Therapeutic Innovation & Regulatory Science,* vol. 57, no. 1, pp. 100–103, 2023.

[13] H. Patel, D. Bell, M. Molokhia, J. Srishanmuganathan, M. Patel, J. Car, and A. Majeed, "Trends in hospital admissions for adverse drug reactions in England: Analysis of national hospital episode statistics 1998–2005," *BMC Clinical Pharmacology*, vol. 7, no. 1, 2007.

[14] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients," *JAMA*, vol. 279, no. 15, p. 1200, 1998.

[15] S.-F. Zhou and W.-Z. Zhong, "Drug design and discovery: Principles and applications," *Molecules*, vol. 22, no. 2, p. 279, 2017.

[16] O. J. Wouters, M. McKee, and J. Luyten, "Estimated research and development investment needed to bring a new medicine to market, 2009-2018," JAMA, vol. 323, no. 9, pp. 844-853, Mar. 2020.

[17] M. L. Billingsley, "Druggable targets and targeted drugs: Enhancing the development of new therapeutics," *Pharmacology*, vol. 82, no. 4, pp. 239–244, 2008.

[18] R. K. Harrison, “Phase ii and phase iii failures: 2013–2015,” *Nature Reviews Drug Discovery*, vol. 15, no. 12, pp. 817–818, 2016.

[19] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 2–12, 2015.

[20] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information Computer Science*, vol. 28, no. 1, pp. 31–36, 1988.

[21] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-

referencing embedded strings (SELFIES): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.

[22] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *Journal of Cheminformatics*, vol. 7, no. 1, 2015.

[23] N. M. O'Boyle, "Towards a universal SMILES representation - A standard method to generate canonical SMILES based on the InChI," *J. Cheminform.*, vol. 4, no. 22, 2012.

[24] Daylight Chemical Information Systems, Inc., "SMARTS Theory," [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. [Accessed: 30-Jul-2024].

[25] A. Capecchi, D. Probst, and J. L. Reymond, "One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome," *J. Cheminform.*, vol. 12, no. 43, 2020.

[26] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modelling*, vol. 50, no. 5, pp. 742-754, 2010.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[28] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, et al., "GPT (Generative Pre-trained Transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, 2024.

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] D. Galeano, S. Li, M. Gerstein, and A. Paccanaro, "Predicting the frequencies of drug side effects," *Nature Communications*, vol. 11, no. 1, p. 4575, 2020.

[31] H. Zhao, K. Zheng, Y. Li, and J. Wang, "A novel graph attention model for predicting frequencies of drug-side effects from multi-view data," *Brief. Bioinform*., vol. 22, no. 6, pp. bbab239, 2021

[32] D. Kalla, N. Smith, F. Samaah, and S. Kuraku, "Study and analysis of chat GPT and its impact on different fields of study," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 3, 2023

[33] H. R. Saeidnia, "Welcome to the Gemini era: Google DeepMind and the information industry," *Library Hi Tech News*, vol. ahead-of-print, 2023.

[34] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, et al., "PubChem in 2021: New data content and improved web interfaces," *Nucleic Acids Res.*, vol. 49, no. D1, 2020.

[35] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Research*, vol. 44, no. D1, 2015, doi: 10.1093/nar/gkv1075

[36] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Reearch.*, vol. 46, no. D1, 2017.

[37] H. Zhao, S. Wang, K. Zheng, Q. Zhao, F. Zhu, and J. Wang, "A similarity-based deep learning approach for determining the frequencies of drug side effects," *Brief. Bioinform*., vol. 23, no. 1, pp. bbab449, 2022.

[38] X. Xu, L. Yue, B. Li, Y. Liu, Y. Wang, W. Zhang, and L. Wang, "DSGAT: predicting frequencies of drug side effects by graph attention networks," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab586, 2022.

[39] L. Wang, C. Sun, X. Xu, J. Li, and W. Zhang, "A neighborhood-regularization method leveraging multiview data for predicting the frequency of drug-side effects," *Bioinformatics*, vol. 39, no. 9, pp. btad532, 2023.

[40] S. Park, S. Lee, M. Pak, and S. Kim, "Dual Representation Learning for Predicting Drug-side Effect Frequency using Protein Target Information," *IEEE Journal of Biomedical Health Informatics*, 2024.

[41] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence

Embeddings," *arXiv preprint* arXiv:2104.08821, 2021.

[42] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M. and Okruszek, L., 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, p.114135.

[43] "WordPiece Subword-Based Tokenization Algorithm," Towards Data Science. [Online]. Available: https://towardsdatascience.com/wordpiece-subword-based-tokenization-algorithm-1fbd14394ed7. [Accessed: 1-Aug-2024].

[44] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[45] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[46] "First Quora Dataset Release: Question Pairs," quoradata.quora.com, 2017. [Online]. Available: https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.

[47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[48] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa-2: Towards chemical foundation models," *arXiv preprint arXiv:2209.01712*, 2022.

[49] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[50] A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach, "An open-source chemical structure curation pipeline using RDKit," *Journal of Cheminformatics*, vol. 12, pp. 1-16, 2020.

[51] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: a benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513-530, 2018.

[52] X. Huang, H. Peng, D. Zou, Z. Liu, J. Li, K. Liu, J. Wu, J. Su, and S. Y. Philip, "CoSENT: Consistent Sentence Embedding via Similarity Ranking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[53] W. Liu, J. Zhang, G. Qiao, J. Bian, B. Dong, and Y. Li, "HMMF: a hybrid multi-modal fusion framework for predicting drug side effect frequencies," *BMC Bioinformatics*, vol. 25, no. 1, p. 196, 2024.

[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[55] X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2023.

[56] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[57] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, “Consert: A contrastive framework for self-supervised sentence representation transfer,” *arXiv preprint arXiv:2105.11741*, 2021.

[58] RunPod, "RunPod - Cloud GPU Rental," available at: https://www.runpod.io, [Accessed: Aug. 1, 2024]

[59] SWIMM, "Embeddings in Machine Learning: Types, Models, and Best Practices," [Online]. Available: https://swimm.io/learn/large-language-models/embeddings-in-machine-learning-types-models-and-best-practices [Accessed: August 19, 2024].

[60] T. G. Kristensen, J. Nielsen, and C. N. Pedersen, "Methods for similarity-based virtual screening," *Computational and Structural Biotechnology Journal*, vol. 5, no. 6, p. e201302009, 2013.

[61] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn,

"STITCH 5: Augmenting protein–chemical interaction networks with tissue and affinity data," *Nucleic Acids Research*, vol. 44, no. D1, 2015.

[62] Y. Tanaka, H. Y. Chen, P. Belloni, U. Gisladottir, J. Kefeli, J. Patterson, A. Srinivasan, M. Zeitz, G. Sirdeshmukh, J. Berkowitz, and K. LaRow Brown, "OnSIDES (ON-label SIDE effectS resource) Database: Extracting Adverse Drug Events from Drug Labels using Natural Language Processing Models," *medRxiv*, pp. 2024-03, 2024.

[63] D. Demner-Fushman, S. E. Shooshan, L. Rodriguez, A. R. Aronson, F. Lang, W. Rogers, K. Roberts, and J. Tonning, "A dataset of 200 structured product labels annotated for adverse drug reactions," *Scientific Data*, vol. 5, no. 1, pp. 1-8, 2018.

[64] L. Yu, M. Cheng, W. Qiu, X. Xiao, and W. Lin, "IDSE-HE: Hybrid embedding graph neural network for drug side effects prediction," *Journal of Biomedical Informatics*, vol. 131, p. 104098, 2022.

[65] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," *arXiv preprint arXiv:1506.06724*, 2015

[66] Wikipedia contributors, "Wikipedia, The Free Encyclopedia," *Wikipedia, The Free Encyclopedia*. [Online]. Available: https://en.wikipedia.org. [Accessed: August 18, 2024]

[67] S. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large, annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 632-642.

[68] A. Borrel, "Development of Computational Methods to Predict Protein Pocket Druggability and Profile Ligands using Structural Data," Ph.D. dissertation, 2016.

[69] X. Qing, X. Y. Lee, J. De Raeymaeker, J. R. H. Tame, K. Y. J. Zhang, M. De Maeyer, and A. R. D. Voet, "Pharmacophore modeling: advances, limitations, and current utility in drug discovery," *Journal of Receptor, Ligand and Channel*

Research, pp. 81-92, 2014.

[70] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742-754, 2010.

[71] G. Deol, "Predicting Environmental Carcinogens with Logistic Regression, KNN, Gradient Boosting and Neural Networks," *Medium*, 2020. [Online]. Available: https://medium.com/@gurkamaldeol/predicting-environmental-carcinogens-with-logistic-regression-knn-gradient-boosting-and-7973f88eb8b3. [Accessed: Aug. 21, 2024].

VITA AUCTORIS

NAME:	Siyam Sajnan Chowdhury
PLACE OF BIRTH:	Chattogram, Bangladesh
YEAR OF BIRTH:	1998
EDUCATION:	Presidency International School, Chattogram, Bangladesh, 2017.
	East Delta University, BSc. in Computer Science & Engineering, Chattogram, Bangladesh, 2022.
	University of London, Bachelor of Laws, London, United Kingdom, 2022.
	East Delta University, MSc. in Data Analytics and Design Thinking for Business, Chattogram, Bangladesh, 2024.
	University of Windsor, M.Sc. in Computer Science (Artificial Intelligence Stream), Windsor, ON, Canada, 2024.