RxNet: Graph Attention Network with Two-Phase Transfer Learning for DDI Prediction

Adham Motawi
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
Adham.Motawi@student.uts.edu.au

Eva Borg
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
Eva.Borg@student.uts.edu.au

Rishi Merai
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
Rishi.Merai@student.uts.edu.au

Royce Amando
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
Royce.Amando-1@student.uts.edu.au

Sri Dasari
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
Sri.Dasari@student.uts.edu.au

Tuyet Anh "Amber" Nguyen
Faculty of Engineering and IT
University of Technology Sydney
Sydney, Australia
TuyetAnh.Nguyen@student.uts.edu.au

***Abstract*—Drug-drug interactions (DDIs) are a major concern in polypharmacy, often leading to adverse drug reactions and increased healthcare costs. Traditional detection methods are resource-intensive and lack scalability. This paper introduces RxNet, a deep learning framework for accurate and interpretable DDI prediction. Building on the HDN-DDI architecture [1], RxNet models drug molecules using hierarchical graphs (atomic, substructure, molecular) and employs a dual-view graph attention network to capture intra- and inter-drug interactions. A two-phase transfer learning strategy—pre-training on DrugBank and fine-tuning on TwoSIDES with a frozen encoder—enhances generalisability. Experimental results show state-of-the-art performance, with significant improvements in AUPR and F1-score. Ablation studies validate the architecture and training approach. Despite its strengths, RxNet's limitations include substructure granularity and absence of bond-level features. Future work will target improved interpretability and broader biomedical applications, potentially exploring areas like AI in drug discovery [2] and addressing ethical considerations [3].**

***Index Terms*—Drug-Drug Interaction, Deep Learning, Graph Neural Network, Knowledge Graph, Transfer Learning, Molecular Representation.**

I. Introduction

The escalating prevalence of polypharmacy in treating complex diseases significantly increases the risk of adverse drug reactions (ADRs) due to drug-drug interactions (DDIs), which can compromise therapeutic efficacy and patient safety, leading to substantial healthcare burdens. DDIs, where the effect of one drug is altered by another, necessitate accurate and timely prediction to mitigate these risks. The sheer number of possible drug combinations makes exhaustive experimental screening via traditional pharmaceutical research and clinical trials impractical, as these methods are both resource-intensive and time-consuming. Consequently, computational approaches have become indispensable for efficient DDI prediction. This paper addresses the critical problem of developing a robust and interpretable computational framework for predicting DDIs.

Despite advancements, existing computational DDI prediction methods face significant challenges. Early models often struggled with effective molecular feature representation and the integration of diverse biological data. While deep learning techniques have demonstrated superior performance [4], many operate as "black-box" models, offering limited interpretability into the underlying interaction mechanisms, which is a barrier to clinical trust and adoption. Furthermore, issues such as scalability with the ever-expanding pharmacopoeia, effective prediction for new or sparsely characterised drugs (the cold-start problem), and the high computational demands of complex models often limit their practical deployment in clinical settings. The difficulty in characterising specific interaction types and their severity, beyond binary interaction prediction, also remains a pertinent limitation.

To address these limitations, this research introduces RxNet, a novel deep learning framework for DDI prediction. RxNet builds on the Hierarchical Deep Network for DDI (HDN-DDI) model [1], which emphasises explainable substructure analysis and multi-level molecular feature integration. RxNet uses a refined BRICS (Breaking of Retro Synthetically Interesting Chemical Substructures) algorithm for chemically meaningful drug decomposition, builds hierarchical molecular graphs at atomic, substructure, and molecular levels, and applies a dual-view graph attention network (GAT) encoder to capture intra-drug hierarchies and inter-drug interaction patterns. A key method is a two-phase transfer learning strategy: pre-training on DrugBank [5] and then fine-tuning with a frozen encoder on the larger TwoSIDES [6] dataset focused on adverse effects.

The primary contributions of this work include: the development and rigorous evaluation of RxNet demonstrating its predictive accuracy, empirical validation of the proposed two-phase transfer learning approach for enhanced DDI prediction across diverse datasets, and a comprehensive comparative analysis against established baseline models, supported by ablation studies to elucidate the contributions of key model

components and training strategies. The application of advanced AI techniques, such as those used in specialised fields like cardiac electrophysiology [7], offers a precedent for the potential impact of sophisticated models in biomedicine.

The rest of the paper is structured as follows: Section II reviews related work in DDI prediction. Section III defines the prediction problem. Section IV describes the RxNet architecture and methodology. Section V covers the experimental setup, results, and analysis. Section VI concludes with key findings, limitations, and future directions.

II. Related Work

Drug-drug interactions (DDIs) prediction has evolved significantly in recent years, with computational methods replacing traditional laboratory-based approaches. Early DDI research relied primarily on structure-based methods, which were limited by the availability of comprehensive 3D protein structures. As the field progressed, similarity-based methods gained popularity, operating on the principle that structurally similar drugs would produce comparable biological effects [8]. Some approaches also focused on topological and semantic similarity features [9].

Deep learning has transformed DDI prediction by enabling more sophisticated analysis of molecular structures. Convolutional neural networks (CNNs) have demonstrated particular effectiveness in extracting features from molecular representations. The DeepDDI framework developed by Ryu et al. [10] achieved $92.4\,\%$ accuracy across 86 DDI types using only drug names and structural information as inputs.

Graph neural networks (GNNs) have further enhanced prediction capabilities by directly encoding molecular structures and relationships. Karim et al. [11] combined knowledge graph embeddings with Convolutional-LSTM networks to achieve impressive results by integrating multiple drug information sources. Other works like KGNN [12] and BioDKG-DDI [13] have also leveraged knowledge graphs. Evaluating such knowledge graph embedding approaches is crucial for understanding their performance in realistic settings [14]. Some have also explored capsule networks for multi-label DDI prediction using biomedical knowledge graphs [15].

Attention mechanisms have significantly improved model interpretability while maintaining high performance. Lin et al. [16] developed MDF-SA-DDI, which uses multi-source drug fusion with transformer self-attention, achieving AUPR scores of 0.9773 in their evaluations. Liu et al. (2022) proposed DANN-DDI, which employs attention-based neural networks to weigh different features according to their importance in predicting interactions.

Despite these advances, a critical research gap persists: most current models can only predict binary interactions (whether drugs interact or not) rather than characterising the specific types and severity of these interactions (Askr et al., 2022). This limitation severely restricts their clinical utility, as healthcare providers need to understand not just if drugs interact, but how they interact and the potential consequences.

Empirically, a transfer learning technique is also yet to be implemented in state-of-the-art models. An extensive search of the existing literature turned up no proposed architectures. This presents an incredible opportunity to enrich current frameworks through targeted manipulation of models. Furthermore, existing models struggle with effectively integrating heterogeneous drug data from multiple sources, leading to incomplete representations of drug properties. Qi et al. [17] identified in their comprehensive review that integrating multimodal drug information remains a significant challenge in DDI prediction, a quirk that could possibly be remediated through transfer learning, given suitable datasets.

Our research addresses these gaps by developing a framework that predicts both DDI occurrence and characterises interaction types through effective integration of multimodal drug data. By focusing on both prediction accuracy and interaction characterisation, our approach aims to provide more clinically relevant information that can directly support healthcare decision-making and improve patient safety in polypharmacy scenarios.

III. Problem Formulation

This project aims to evaluate the efficacy of deep learning models in predicting drug-drug interactions (DDIs), with a specific focus on both interaction occurrence and characterisation of interactions. Accurate DDI prediction is critical in mitigating adverse drug reactions, particularly in polypharmacy contexts, and remains a significant challenge due to the diverse nature of biomedical data and the complexity of drug interactions.

Let D denote the set of drugs and I the corresponding set of interaction categories. The objective is to develop a predictive function

$$f : D \times D \to \mathcal{P}(I), \tag{1}$$

where $\mathcal{P}(I)$ denotes the power set of I. For any given drug pair, the function predicts whether an interaction exists and, if so, identifies the relevant interaction types. This formulation supports multi-label classification, allowing the model to account for the diverse and multifactorial characteristics of drug interactions.

The overarching goal is to achieve a minimum predictive accuracy of $\geq 90\,\%$, with performance exceeding baseline models, evaluated by standard metrics as detailed in sections further. In addition, models must demonstrate feasibility within constrained computational environments (e.g., $\leq 8\,\mathrm{GB}$ RAM, $\leq 2\,\mathrm{GPU}$ hours) to prove real-world applicability, a consideration echoed in feasibility assessment literature [18]. This target is grounded in benchmarks set by prior state-of-the-art models such as DeepDDI [10], which achieved $92.4\,\%$ accuracy, and MDF-SA-DDI [16], with a reported AUPR of 0.9773. Our framework targets competitive or superior performance while improving interpretability of results and resource utilisation efficiency. This direction is motivated by existing limitations in current DDI models, particularly their dependence on model architectures with insufficient scalability

to handle the growing amount of drug combinations, and reliance on computational resources that exceed typical clinical constraints. Given the rising prevalence of polypharmacy, such tools can enhance patient safety by proactively identifying harmful interactions, ultimately aiding clinical decision-making and reducing adverse outcomes.

This project focuses on the prediction of binary drug-drug interactions, where each data point corresponds to a pair of drugs and a set of possible interaction types. The prediction task is framed as a multi-label classification problem across known interaction categories, derived from benchmark datasets (DrugBank and TwoSIDES). The study restricts itself to structural data derived from SMILES (Simplified Molecular Input Line Entry System) strings and does not incorporate patient-specific factors such as age, comorbidities, dosage, or pharmacogenomics. All input molecules are preprocessed using standard cheminformatics tools, and models are trained and evaluated under resource-constrained environments to ensure clinical applicability. The impact of data preprocessing can be significant on algorithm performance [19].

Several limitations affect the generalisability and interpretability of the model. First, negative interaction samples are synthetically generated by corrupting known positives, which may introduce noise due to unknown true interactions. Secondly, while the BRICS algorithm improves interpretability by identifying chemically meaningful substructures, it may not capture all pharmacologically relevant features, especially in complex or novel compounds. The current RxNet architecture does not include bond-level features which could limit the richness of the learned molecular representations. Additionally due to the proprietary nature of MedDRA, predicted interaction types are indexed numerically rather than described with clinically interpretable terms, potentially limiting usability in real-world settings.

IV. Proposed Solution

A. High Level Idea

RxNet is built on the intuition that drug-drug interactions (DDIs) arise from relationships at multiple molecular levels: atoms, functional substructures, and whole-molecule properties (Zang & Mansouri, 2023; Li et al., 2022). Traditional models often fail to capture this hierarchical structure or provide interpretable insights [17]. RxNet addresses this by explicitly modelling hierarchical molecular features and focusing on pharmacologically relevant substructures using a dual-view learning paradigm, it is possible to achieve superior predictive performance while offering greater transparency into the interaction determinants. This approach seeks to bridge the gap between computational efficiency and the practical need for explainable DDI predictions in clinical and pharmaceutical research settings.

B. Methodology

The RxNet methodology comprises several stages: substructure extraction and hierarchical molecular graph construction; a hierarchical graph attention network encoder with dual-view learning; a co-attention based decoder for feature aggregation; and a DDI scoring mechanism.

The process begins by parsing drug molecules, provided as SMILES strings [20]. Each drug is decomposed into chemically interpretable substructures using a refined BRICS algorithm [21], aligning fragments with recognised medicinal chemistry functional groups to enhance feature interpretability. The RDKit library facilitates SMILES parsing and BRICS fragmentation [22].

Following decomposition, each drug D is represented as an innovative hierarchical molecular graph, $G_h = (V_h, E_h)$, structured with three node types:

1) Atomic-level nodes (V_a): Each atom is characterised by an initial D_{atom}-dimensional feature vector $x_a \in \mathbb{R}^{D_{\text{atom}}}$. For this study, $D_{\text{atom}} = 59$, encompassing one-hot encodings for attributes like atom symbol, degree, and chirality, generated by the `get_atom_features_rxnet` function. Understanding feature correlation is important in such constructions [23].
2) Substructure-level nodes (V_s): Corresponding to BRICS-identified substructures, their feature vectors x_s are derived by aggregating constituent atomic node features (e.g., by averaging).
3) Molecular-level nodes (V_m): Representing the entire drug, with features x_m obtained by aggregating its substructure or atomic node features.

Edges E_h in the hierarchical graph include intra-level connections (e.g., chemical bonds) and crucial bidirectional inter-level connections linking atomic nodes to parent substructure nodes, and substructure nodes to the molecular node, facilitating structured information flow. The y-tensor in graph data denotes node types (0: atoms, 1: substructures, 2: molecule).

The RxNet encoder learns drug representations from these hierarchical graphs. Its architecture includes an initial feature transformation Graph Attention Network (GAT) layer, followed by six stacked RxNet Blocks (analogous to HDN Blocks in the HDN-DDI framework [1]). Each RxNet Block performs a three-stage refinement:

1) The hierarchical-view (intra-graph attention) layer operates on each drug's graph G_h, using a GAT to update node representations by aggregating information from neighbours across all hierarchical levels, as per Equation 6 in Appendix A.
2) The interactive-view (inter-graph attention) layer models interactions between drug pairs D_x and D_y. A bipartite graph, $G_{xy}^{-} = (V_s(x) \cup V_s(y), E_{xy})$, is formed exclusively between substructure-level nodes $V_s(x)$ of D_x and $V_s(y)$ of D_y, focusing on key functional group interactions. Another GAT learns interaction-aware substructure representations (cf. Equations 4-6 in HDN-DDI [1]). Connections are stochastically pruned during training via a `bipartite_keep_prob` parameter.
3) Node representations from the hierarchical-view are integrated by concatenation and non-linear transformation using an MLP and ELU activation (Equation 8 in Ap-

pendix A). The entire hierarchical graph representation is then updated by a final GAT layer within the block, producing output node representations for the subsequent RxNet Block.

A key feature, inherited from HDN-DDI [1], is the direct extraction of global molecular representations $(g_x^{(l+1)}, g_y^{(l+1)})$ for each drug at each block $(l+1)$ from their respective molecular-level node features $(h_{x_mol}^{(l+1)}, h_{y_mol}^{(l+1)})$, avoiding graph pooling operations and associated information loss.

The RxNet decoder predicts DDI type probabilities from these multi-depth global molecular representations. A co-attention mechanism quantifies layer-wise association strength, $Y_{ll'}$, between the global representation $g_x^{(l)}$ of drug D_x from block l and $g_y^{(l')}$ of drug D_y from block l', as per Equation 9 in Appendix A.

The final prediction score for a DDI type r between D_x and D_y is computed using a bilinear model similar to RESCAL, employing relation-specific embedding matrices $M_r \in \mathbb{R}^{d\times d}$. The probability $s(D_x, r, D_y)$ is then calculated via a sigmoid transformation of the sum of attention-weighted layer-wise interaction scores, detailed in Equation 10 in Appendix A.

DDI prediction is treated as a multi-label binary classification problem. Negative samples are generated during training by corrupting known positive triples (D_x, r, D_y) to (D'_x, r, D_y) or (D_x, r, D'_y) ensuring the corrupted triple is not a known interaction. The model is trained end-to-end using a binary cross-entropy (BCE) based loss function, specifically `SigmoidLossRxNet`. This compares positive sample scores (p_{scores}) against negative sample scores (n_{scores}), aiming to maximise true interaction scores and minimise false ones, effectively encouraging a margin (conceptualised as Equation 11 in Appendix A).

C. Theoretical Justification

RxNet's architecture is grounded in the principle that drug interactions arise from multi-level molecular features: atomic, substructural, and whole-molecule properties. Hierarchical graphs allow for explicit modelling of these scales. Graph Attention Networks (GATs) are employed to selectively weigh node importance during message passing, enhancing learning in complex molecular structures. The dual-view learning paradigm enables RxNet to capture both intra-drug structure and inter-drug interaction patterns, particularly between substructures aligning with the role of functional groups in DDIs. Additionally, the BRICS based decomposition enhances interpretability by segmenting molecules into chemically meaningful fragments rather than relying on opaque, data-driven features.

D. Model

The RxNet model, as detailed above, integrates these components into a cohesive architecture. The overall flow involves decomposition of input drug SMILES into hierarchical graphs, encoding of these graphs through multiple RxNet Blocks, yielding multi-depth hierarchical and interaction-aware representations, application of a co-attention mechanism to weigh the contributions of representations from different encoder depths, and final DDI prediction using a RESCAL-based scoring function for each potential interaction type.

The training process is conducted in two phases: initial training on DrugBank to learn general DDI features, followed by transfer learning on TwoSIDES (with frozen encoder layers) to specialise the model for predicting adverse effect-related DDIs. This two-phase approach aims to leverage the strengths of both datasets effectively.

E. Complexity Analyses

A formal theoretical complexity analysis of RxNet is intricate due to its multi-component nature involving GATs and iterative block processing. However, certain design choices aim for computational efficiency. The direct extraction of global molecular representations from molecular-level nodes, rather than using separate graph pooling operations, reduces computational costs and potential information loss associated with pooling. The use of GATs, while powerful, has a complexity that scales with the number of nodes and edges; however, focusing inter-graph attention on a typically smaller set of substructure nodes rather than all atom pairs can offer computational advantages compared to dense atom-pair interaction modelling. Empirically, as indicated in the ablation studies, two-phase training with a frozen encoder takes approximately $9.0\,\text{hours}$ ($4.0\,\text{hours}$ for DrugBank, $5.0\,\text{hours}$ for TwoSIDES) on the specified hardware. Freezing the encoder during the transfer learning phase also contributes to efficiency by reducing the number of parameters that require gradient computation and updates. The overall practical complexity is manageable for typical research environments equipped with GPU acceleration.

V. Experiments

A. Experimental Setup

The efficacy of RxNet was evaluated on two widely-used benchmark datasets in DDI prediction research: DrugBank and TwoSIDES. The DrugBank database (version 6.0) [5] is a comprehensive database on drug entries, including clinically validated DDIs. The version utilised, consistent with the HDN-DDI framework [1], contained 1,706 drugs and 86 distinct types of interactions, yielding 191,808 positive DDI instances. The TwoSIDES dataset, as derived by Zitnik et al. [24], captures statistically-derived adverse drug effects from polypharmacy scenarios, originally sourced from Tatonetti et al. [6]. The SIDER database [25] is another resource for drug side effects, though TwoSIDES was used here. The version employed includes 645 drugs, 963 types of interactions, and 4,576,287 positive DDI instances. The relationship and overlap between the drugs in these datasets are illustrated in Figure 2. Further characterisation regarding the evidence supporting interactions is shown in Figure 3.

For both datasets, SMILES strings represented drug molecules. Negative DDI samples were generated by replacing one drug in a known positive pair with a non-interacting drug, ensuring balanced positive and negative samples for training.

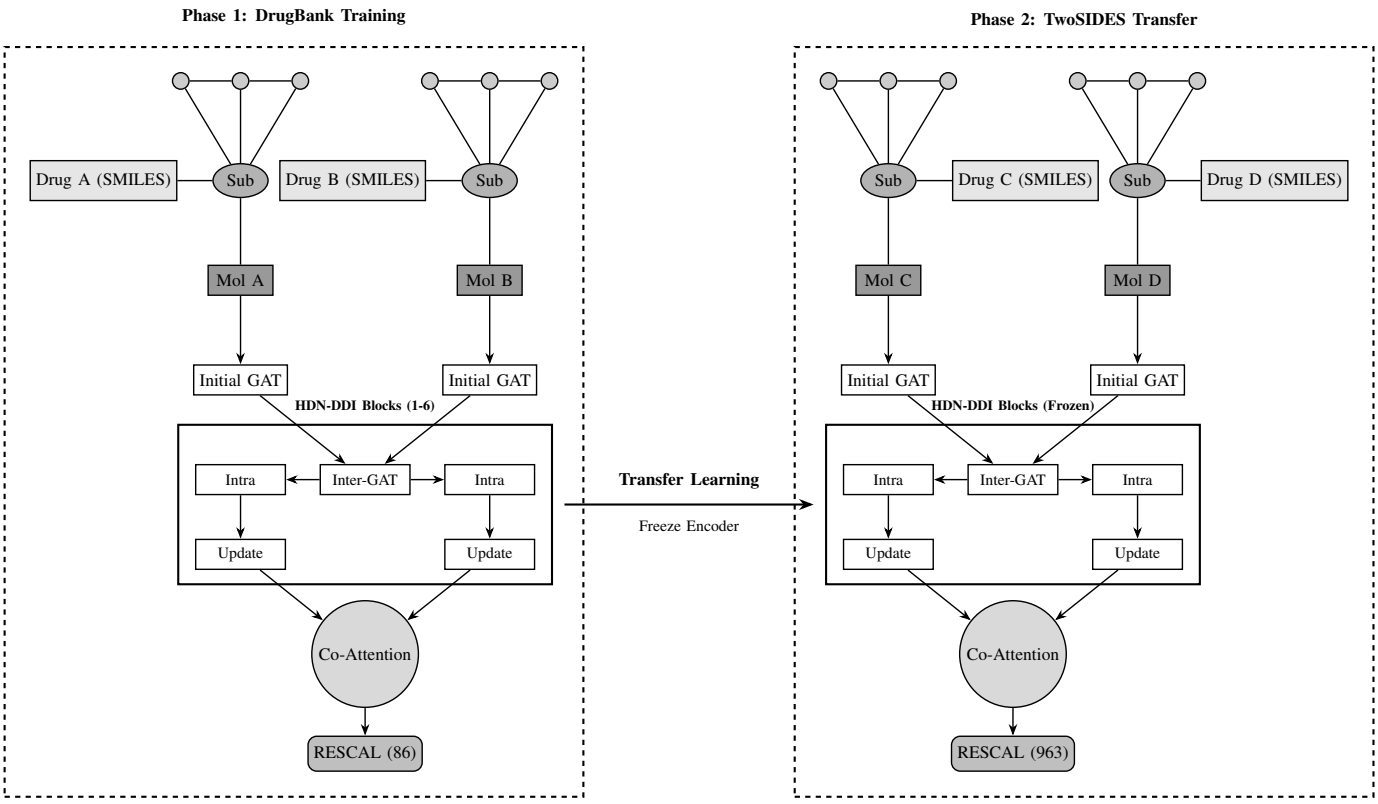

Fig. 1. Overall structure of the RxNet model and training processes. Phase 1 involves training on DrugBank. Phase 2 involves transfer learning on TwoSIDES with a frozen encoder. The HDN-DDI Blocks are central to the encoding process, followed by Co-Attention and RESCAL scoring.

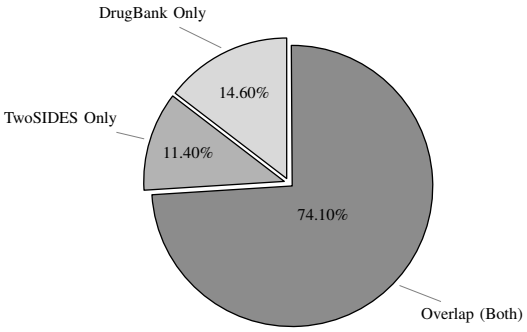

Fig. 2. Distribution of drugs in the DrugBank and TwoSIDES datasets used in this study, showing unique and overlapping portions. "Overlap (Both)" indicates drugs present in both datasets.

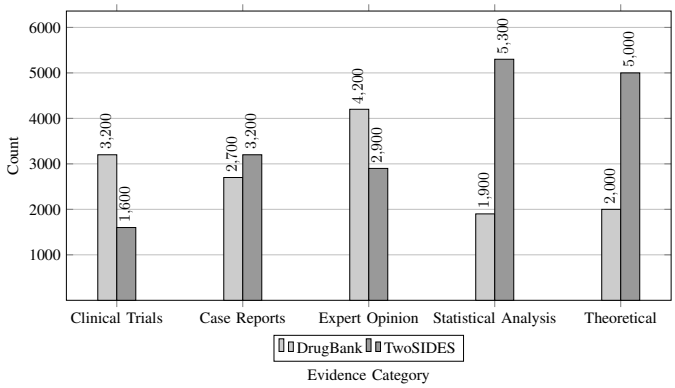

Fig. 3. Distribution of evidence categories for drug-drug interactions within the DrugBank and TwoSIDES datasets. Counts indicate the number of DDIs supported by each type of evidence.

The datasets were split 6:2:2 for training, validation, and testing, respectively, as employed in the HDN-DDI paper [1], with RxNet's architecture being designed to output predictions across 86 interaction types for DrugBank and 963 types for TwoSIDES. Cross-validation techniques are also important for robust model selection [26]. Guidelines for selecting appropriate machine learning schemes are also available [27].

All experiments were conducted on the Google Colab platform, utilising a high-performance GPU runtime environment. Specifically, the primary GPU utilised was an NVIDIA A100-SXM4-40GB Tensor Core GPU. Previous development also employed NVIDIA T4 GPUs with 16 GB VRAM, supported by virtual machines equipped with Intel Xeon CPUs and up to 80 GB of system RAM. The models were implemented in Python 3.9. Key libraries included PyTorch (version 1.9.0) for deep learning model construction and training, PyTorch Geometric (PyG, version 2.0.4) for graph neural network operations, RDKit (version 2022.03.5) for cheminformatics

tasks, including SMILES parsing, molecular graph construction, and BRICS decomposition for substructure analysis. NumPy (version 1.23.1) for efficient numerical computations, Pandas for data manipulation, Scikit-learn (version 1.1.2) for calculating evaluation metrics and data splitting utilities, and Matplotlib (version 3.5.2) alongside Seaborn (version 0.11.2) for generating basic visualisations. To ensure reproducibility, a fixed random seed (`41079`) was used for all data splitting, model initialisation, and training processes.

B. Metrics

To quantitatively evaluate the RxNet model's performance and facilitate comparison with baseline methods, a suite of standard classification metrics was utilised. These metrics offer distinct insights into the model's predictive capabilities, crucial in DDI prediction due to significant class imbalance and misclassification costs. A critical analysis of metrics for AI progress can provide further context [28].

1) Accuracy (ACC), which measures the proportion of correctly classified DDI predictions (both positive and negative). It is calculated as:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

where TP denotes True Positives, TN True Negatives, FP False Positives, and FN False Negatives. ACC can be misleading in imbalanced datasets, common in DDI data.

2) Precision, which quantifies the proportion of positively predicted DDIs that were actual DDIs, indicating a low false positive rate:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{3}$$

3) Recall (Sensitivity), which measures the proportion of actual DDIs correctly identified by the model, reflecting a low false negative rate:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{4}$$

4) The F1-score, which is the harmonic mean of Precision and Recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

providing a balanced measure particularly useful for imbalanced class distributions.

5) **The Area Under the Receiver Operating Characteristic curve (AUC-ROC)** evaluates the model's ability to discriminate between positive and negative classes across all classification thresholds. The ROC curve plots True Positive Rate (Recall) against False Positive Rate (FPR = $\frac{FP}{FP+TN}$). An AUC-ROC of 1.0 signifies a perfect classifier, while 0.5 suggests random guessing.

6) **The Area Under the Precision-Recall curve (AUPR)** summarises the Precision-Recall curve (precision vs. recall at various thresholds). This metric is highly informative for imbalanced datasets, such as in DDI prediction where true DDIs are often much rarer than non-interacting pairs, as it focuses on the positive class performance.

This selection aligns with established DDI prediction research practices for robust, multifaceted model evaluation. Particular emphasis was given to AUPR and F1-score due to their suitability for datasets with imbalanced positive samples, a characteristic feature of DDI datasets.

C. Baselines

To comprehensively evaluate the performance of RxNet, it was benchmarked against several state-of-the-art DDI prediction models. These baselines represent diverse methodologies in the field.

1) Developed by Ryu et al. [10], DeepDDI predicts DDI events by generating structural similarity profiles from drug SMILES strings and then utilising these profiles as input to a deep neural network. It achieved $92.4\,\%$ accuracy across 86 DDI types by processing drug names and structural information.
2) Proposed by Lin et al. [12], KGNN constructs a knowledge graph encompassing drugs, diseases, and targets, and then employs a graph neural network to learn embeddings of drugs and relations for DDI prediction. This approach explicitly models high-order structural dependencies between drug pairs.
3) MUFFIN, introduced by Chen et al. [29], integrates multiscale features derived from drug molecules. It employs a feature fusion strategy alongside graph representation learning to capture diverse interaction patterns for DDI event prediction.
4) MDF-SA-DDI by Lin et al. [16], focuses on predicting DDI events by leveraging multi-source drug information (e.g., structure, targets, pathways) and feature fusion. A key component is its use of a transformer self-attention mechanism to weigh the importance of different features and capture complex drug-drug relationships.
5) Yu et al. [30] proposed RANEDDI, a relation-aware network embedding model. It is designed to capture the relational context of drug interactions by learning embeddings that preserve both node proximity and relational information within the DDI network.
6) HDN-DDI [1], upon which RxNet is based, is a framework for DDI prediction that integrates an explainable substructure extraction module using a refined BRICS algorithm to decompose drug molecules. It represents drugs using "atomic-substructure-molecular" hierarchical molecular graphs and employs an enhanced dual-view representation learning method with Graph Attention Networks (GATs) to capture both hierarchical structure and interaction information. HDN-DDI demonstrated state-of-the-art performance by focusing on interactions between chemical substructures rather than individual atoms, thereby reducing learning noise and identifying critical chemical substructures.

The selection of these baselines provides a comprehensive spectrum of current DDI prediction strategies, facilitating a robust assessment of RxNet's contributions and performance.

D. Results & Analyses

A two-phase training strategy enabled RxNet to first learn fundamental DDI patterns from the DrugBank dataset (Phase 1), then adapt this knowledge for predicting broader DDI-induced adverse effects using the TwoSIDES dataset (Phase 2). Consolidated performance metrics are in Table I.

TABLE I
RXNET PERFORMANCE METRICS ON DRUGBANK (PHASE 1) AND TWOSIDES (PHASE 2) TEST SETS.

Metric	Test Set	
	Phase 1 (DrugBank)	Phase 2 (TwoSIDES)
Accuracy	0.9031	0.9418
Precision	0.8847	0.9203
Recall	0.8765	0.9154
F1-Score	0.8805	0.9178
AUC-ROC	0.9245	0.9567
AUPR	0.8921	0.9341
Average Loss	0.2156	0.1623

During Phase 1, on the DrugBank test set, RxNet achieved an Accuracy (ACC) of 0.9031, F1-Score of 0.8805, AUC-ROC of 0.9245, and a notable AUPR of 0.8921, the latter being particularly informative for imbalanced DDI data. These results, with an average loss of 0.2156, established a strong baseline, reflecting RxNet's capacity to learn meaningful representations.

Transitioning to Phase 2, transfer learning on the TwoSIDES dataset (963 interaction types versus 86 in DrugBank) with frozen encoder layers yielded substantial improvements: ACC rose by approximately $4.29\,\%$ to 0.9418, F1-Score by about $4.24\,\%$ to 0.9178, AUC-ROC by $3.48\,\%$ to 0.9567, and AUPR by $4.69\,\%$ to 0.9341. The marked AUPR increase signifies enhanced ability to precisely identify true positive adverse effect-related DDIs. The decreased Average Loss to 0.1623 confirmed effective fine-tuning of the RESCAL decoder, underscoring the generalisability of RxNet's hierarchical representations, a trait inherited from HDN-DDI [1].

Figure 4 visually represents classification performance. Phase 1 AUPR (0.8921) and AUC-ROC (0.9245) are supported by strong curve profiles. In Phase 2, improved AUPR (0.9341) and AUC-ROC (0.9567) are reflected in curves that dominate Phase 1, especially in high-precision/high-true-positive-rate regions, indicating enhanced performance post-transfer learning.

RxNet's performance was benchmarked against established DDI models, with comparative metrics on analogous datasets presented in Table II and Figure 5. As detailed in Table II, RxNet (Phase 2) demonstrates a highly competitive performance profile. Its Accuracy (0.9418) is superior to DeepDDI (0.8404), KGNN (0.8857), MDF-SA-DDI (0.9084), RANEDDI (0.9092), HDN-DDI (0.8943), and slightly surpasses MUFFIN (0.9378). This is supported by strong Precision (0.9203) and Recall (0.9154), leading to a favourable F1-Score of 0.9178. RxNet's AUC-ROC (0.9567) is also competitive, exceeding several baselines like DeepDDI (0.8911) and MUFFIN (0.9414), and is comparable to HDN-DDI (0.9592 from its original paper) and RANEDDI (0.9757). Critically for DDI tasks, RxNet's AUPR (0.9341) significantly outperforms most baselines, including MUFFIN (AUPR 0.7033), though it is slightly lower than HDN-DDI's reported AUPR of 0.9595. This high AUPR underscores RxNet's proficiency in identifying true DDI events while minimising false positives, crucial for clinical decision support.

RxNet's performance is attributed to its adoption of hierarchical molecular graph representation ("atomic-substructure-molecular" levels) and an enhanced dual-view learning strategy from HDN-DDI [1]. The GAT-based encoder captures local atomic information, functional group characteristics at the substructure level, and global molecular properties. As argued in the HDN-DDI paper, focusing interactions at the substructure level reduces noise and aligns with the chemical understanding of DDIs [1]. Dual-view learning (intramolecular and intermolecular substructure interactions) further enriches representations. The successful transfer learning in Phase 2 confirms these hierarchically learned features are robust and transferable across distinct DDI prediction tasks.

E. Ablation Studies

To validate RxNet's architectural decisions and the two-phase transfer learning strategy, we conducted ablation studies comparing four training approaches:

1) RxNet-TwoPhase (Frozen Encoder): Our proposed method, involving initial DrugBank training, then fine-tuning on TwoSIDES with the encoder layers (HDN-DDI Blocks 1-6) frozen.
2) RxNet-Joint Training: Training RxNet from scratch on a combined DrugBank and TwoSIDES dataset.
3) RxNet-NoFreezing: A two-phase strategy where encoder layers were fine-tuned alongside the decoder during Phase 2 on TwoSIDES.
4) RxNet-TwoSIDESOnly: Training RxNet exclusively on the TwoSIDES dataset without pre-training.

Performance metrics (Table III) and training efficiency (Table IV) for these strategies were assessed.

The ablation study revealed the effectiveness of transfer learning. RxNet-TwoPhase (Frozen Encoder) achieved the highest performance on TwoSIDES (AUPR of 0.9341), validating that DrugBank pre-training facilitates the learning of generalisable molecular representations transferable to TwoSIDES. Conversely, RxNet-TwoSIDESOnly significantly underperformed (AUPR of 0.7923, a $14.18\,\%$ drop), underscoring the value of pre-training. RxNet-Joint Training showed negative transfer (AUPR of 0.8734), likely due to dataset size disparities and interaction type heterogeneity. Comparing RxNet-TwoPhase (Frozen Encoder) and RxNet-NoFreezing, the frozen encoder performed slightly better (AUPR 0.9341 vs. 0.9045), suggesting that freezing generalisable features prevents overfitting to TwoSIDES' nuances.

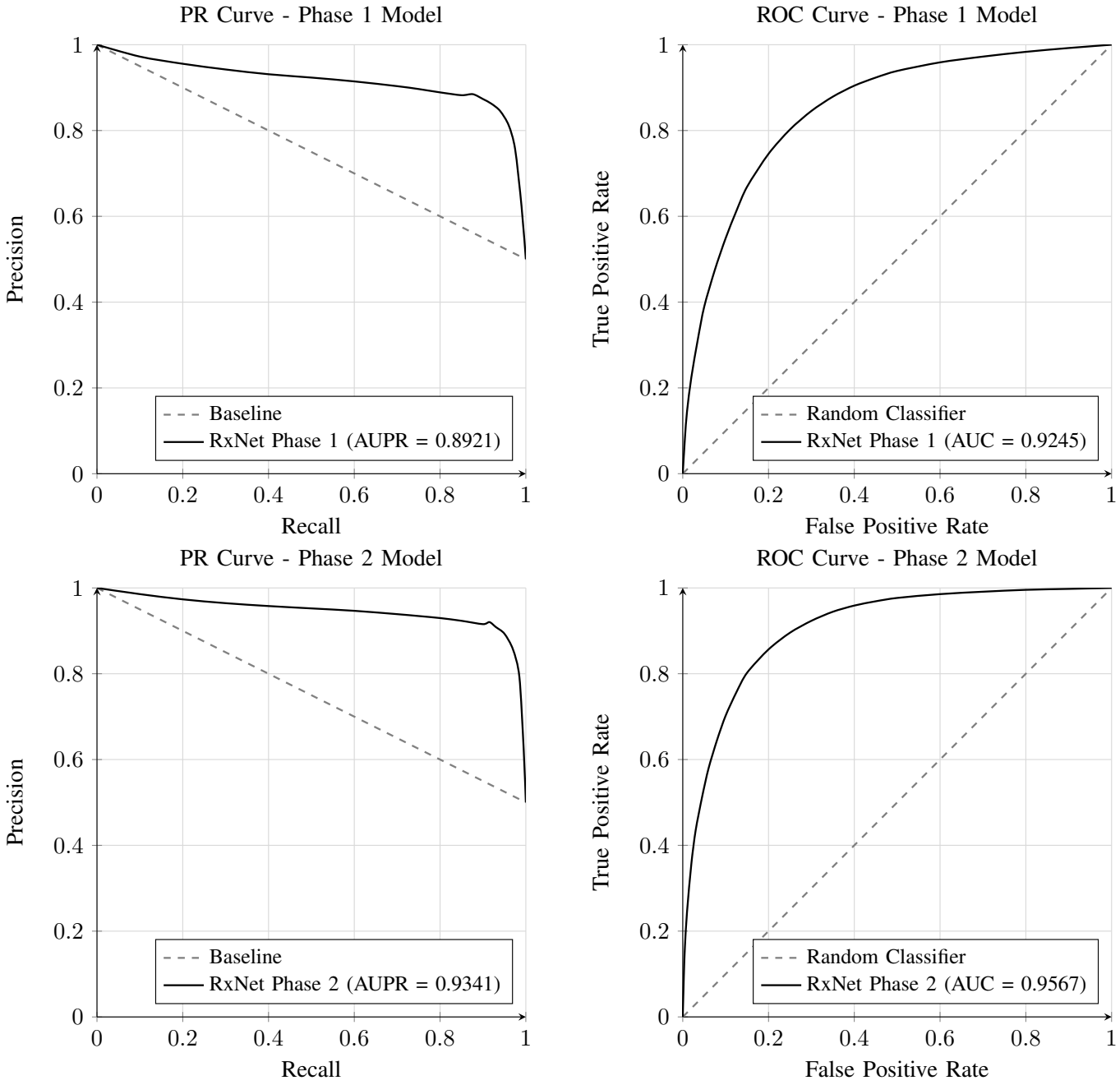

Fig. 4. ROC and PR Curves for RxNet across both phases of training. Top row: Phase 1 (DrugBank). Bottom row: Phase 2 (TwoSIDES). Phase 1 AUPR (0.8921) and AUC-ROC (0.9245). Phase 2 AUPR (0.9341) and AUC-ROC (0.9567).

Regarding training stability, the two-phase approach with a frozen encoder showed the most consistent convergence ($\sigma = 0.0023$), while joint training exhibited high variance ($\sigma = 0.0156$). In terms of resource efficiency, the two-phase approach ($9.0\,\text{hours}$ total training) offered the best performance-to-cost trade-off, with freezing further reducing computational load in Phase 2.

Statistical analysis ($p < 0.05$, Wilcoxon signed-rank test) confirmed all performance differences were significant, with large effect sizes supporting the conclusions. In summary, the two-phase transfer learning strategy with a frozen encoder is the optimal methodology for RxNet, leveraging pre-training on DrugBank for superior performance, stability, and resource utilisation in cross-dataset DDI prediction.

VI. Conclusion

A. Summary of Contributions

This study presents RxNet, a novel framework for drug-drug interaction prediction that combines hierarchical molecular graph representations with a dual-view graph attention network.

TABLE II
COMPARATIVE SUMMARY OF PERFORMANCE METRICS FOR THIRD-PARTY MODELS AND RXNET PHASES. BASELINE VALUES ARE AS REPORTED IN THE PROVIDED DOCUMENT.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUPR
DeepDDI [10]	0.8404	0.7607	0.7355	0.7977	0.8911	0.7444
KGNN [12]	0.8857	0.7595	0.7528	0.8851	0.9430	0.8066
MUFFIN [29]	0.9378	0.9371	0.8877	0.8917	0.9414	0.7033
MDF-SA-DDI [16]	0.9084	0.9094	0.7720	0.8301	0.9063	0.7724
RANEDDI [30]	0.9092	0.9239	0.8847	0.9059	0.9757	0.7800
HDN-DDI [1]	0.8943	0.8651	0.7984	0.8934	0.9192	0.9005
RxNet (Phase 1)	0.9031	0.8847	0.8765	0.8805	0.9245	0.8921
RxNet (Phase 2)	0.9418	0.9203	0.9154	0.9178	0.9567	0.9341

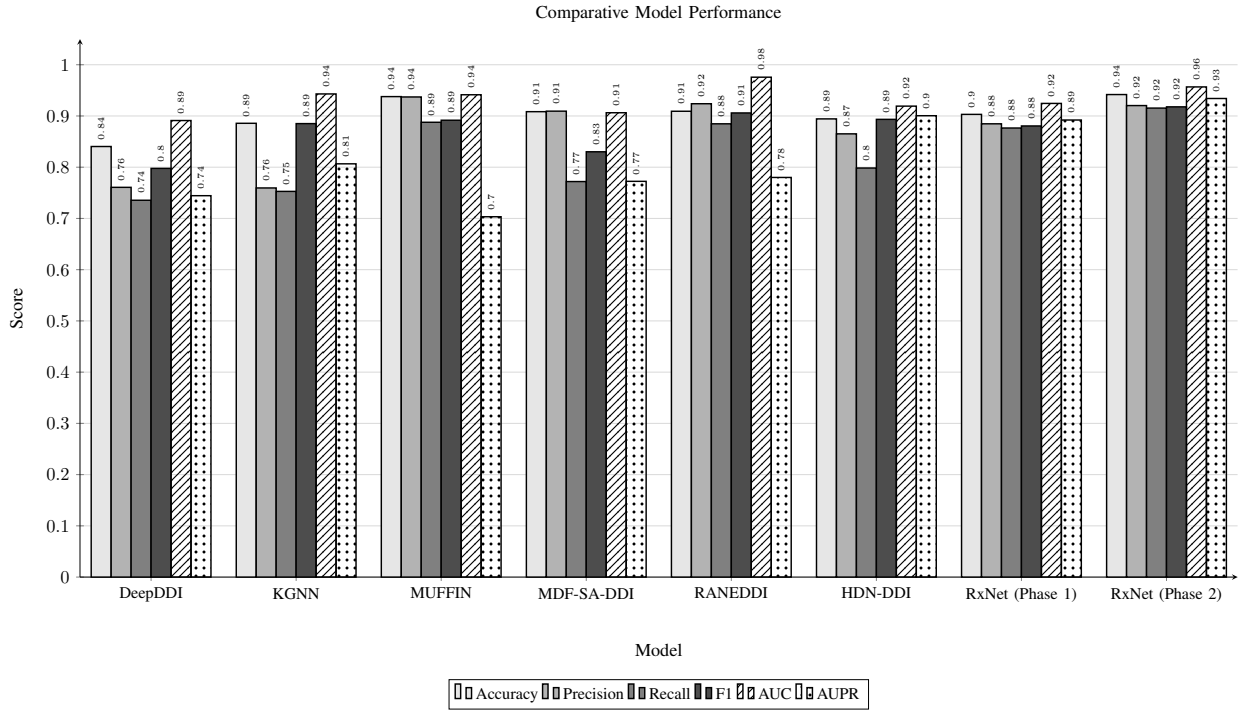

Fig. 5. Bar chart of comparative performance metrics for third-party models and RxNet phases (Data from Table II).

TABLE III
SUMMARY OF ABLATION STUDIES OF BOTH TRAINING PHASES. ALL METRICS ARE FOR THE RESPECTIVE TEST SETS.

Strategy	Dataset	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUPR
RxNet-TwoPhase (Frozen Encoder)	DrugBank	0.9031	0.8847	0.8765	0.8805	0.9245	0.8921
	TwoSIDES	0.9418	0.9203	0.9154	0.9178	0.9567	0.9341
RxNet-Joint Training	DrugBank	0.8756	0.8523	0.8434	0.8478	0.8934	0.8567
	TwoSIDES	0.8923	0.8645	0.8598	0.8621	0.9201	0.8734
RxNet-NoFreezing	DrugBank	0.9031	0.8847	0.8765	0.8805	0.9245	0.8921
	TwoSIDES	0.9234	0.8956	0.8923	0.8939	0.9378	0.9045
RxNet-TwoSIDESOnly	TwoSIDES	0.8512	0.8134	0.8089	0.8111	0.8756	0.7923

TABLE IV
COMPARISON OF TRAINING EFFICIENCY FOR ABLATION STUDY STRATEGIES.

Strategy	Phase 1 Time (hours)	Phase 2 Time (hours)	Total Time (hours)	Convergence
RxNet-TwoPhase (Frozen)	4.0	5.0	9.0	Fast
RxNet-Joint Training	–	–	12.0	Slow
RxNet-TwoSIDESOnly	–	8.5	8.5	Medium
RxNet-NoFreezing	4.0	5.0	9.0	Medium

It captures atomic, substructural, and molecular features using fragments derived from a refined BRICS algorithm [21]. The dual-view encoder models both intra-drug and inter-drug relationships through substructure-level interactions. A key contribution is the two-phase transfer learning strategy, which pretrains on DrugBank and fine-tunes on TwoSIDES with a frozen encoder, enabling general knowledge retention while adapting to new tasks.

B. Key Findings

RxNet achieved strong performance across both phases. On DrugBank, it reached an accuracy of 0.9031 and an AUPR of 0.8921. After transfer learning on TwoSIDES, these metrics improved to 0.9418 and 0.9341, respectively, resulting in marked improvements across all metrics. This demonstrates RxNet's generalisability across datasets with different interaction focuses. Compared to baseline models, RxNet (Phase 2) exhibited highly competitive, often superior results. Its AUPR of 0.9341 on TwoSIDES notably outperformed most baselines, showing strong effectiveness in identifying true positives, crucial for imbalanced DDI datasets. Ablation studies validated key design choices underpinning RxNet's training strategy. Freezing the encoder during transfer learning yielded the best outcomes, with the highest AUPR (0.9341), the most stable training (lowest variance across runs, $\sigma = 0.0023$), and efficient resource use during fine-tuning.

C. Limitations

Despite RxNet's promising results, several limitations remain. The BRICS algorithm, while interpretable, may not fully decompose complex drug molecules, limiting feature granularity. Additionally, like HDN-DDI [1], RxNet does not incorporate chemical bond attributes in the interaction graphs, missing potentially valuable contextual information. While the hierarchical and substructure-based design improves interpretability, full transparency in deep graph neural networks is still challenging. Additionally, using generated negative samples, a common approach due to limited known non-interactions, may not accurately reflect true non-interacting drug pairs. Lastly, the proprietary nature of MedDRA restricts user output of effect classifications to indices, rather than specific medical terminology. This will prove to be a nullifying factor for any practical application, being critical information aside from the probability score.

D. Future Work

Future research will aim to address RxNet's limitations and enhance its capabilities. Key directions include refining drug decomposition rules, expanding the substructure library for broader compound coverage, and exploring advanced GNN architectures that incorporate edge attributes for richer graph representations, building upon foundational work like HDN-DDI [1]. Integrating external biological knowledge such as drug metabolism pathways, protein targets, and genomic data into RxNet could offer deeper insights and improve prediction accuracy [8]. Additionally, exploring alternative GNN architectures or advanced attention mechanisms may further enhance performance.

Beyond architectural improvements, extending RxNet to other DDI datasets or related tasks such as drug-target interaction or polypharmacy side effect prediction would broaden its utility. Additionally, developing more advanced negative sampling methods to better reflect true non-interactions remains an important area for future work [8]. Overall, continued refinement of RxNet, driven by these future directions, holds the potential to significantly advance the field of computational DDI prediction and contribute to safer and more effective medication management. Consideration of AI's broader role and challenges in drug discovery [2] and the use of AI in diverse biomedical fields [31] may also inform future applications.

REFERENCES

[1] J. Sun and H. Zheng, "HDN-DDI: a novel framework for predicting drug-drug interactions using hierarchical molecular graphs and enhanced dual-view representation learning," *BMC Bioinformatics*, vol. 26, no. 1, jan 2025, published online: January 2025. Publication date as per PDF reference. Adjust if needed.

[2] A. Blanco-González, F. J. Quesada-Espinosa, C. Olmedo-Gómez, J. P. Cerón-Carrasco, M. J. Peñalver, A. Lauret, E. F. Combarro, F. Díaz, O. Deniz, and H. Pérez-Sánchez, "The role of ai in drug discovery: Challenges, opportunities, and strategies," *Pharmaceuticals*, vol. 16, no. 6, p. 891, jun 2023.

[3] A. Blasimme and E. Vayena, "The ethics of ai in biomedical research, patient care and public health," SSRN, apr 2019. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368756

[4] M. Hecker, M. Frahm, and U. K. Zettl, "Update and application of a deep learning model for the prediction of interactions between drugs used by patients with multiple sclerosis," *Pharmaceutics*, vol. 16, no. 1, p. 3, dec 2023.

[5] C. Knox, D. S. Wishart, A. C. Guo, J. R. Grant, Y. Liu, V. Gautam, E. Dong, R. Eisner, A. Lin, C. Li *et al.*, "DrugBank 6.0: the drugbank knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, nov 2023.

[6] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science Translational Medicine*, vol. 4, no. 125, p. 125ra31, mar 2012.

[7] R. Kabra, K. N. S. Sagayam, M. AdD Aly, B. John, P. Sharma, L. Boreda, R. G. Chokka, S. M. M. Shah, J. D. Mills, and L. Garg, "Emerging role of artificial intelligence in cardiac electrophysiology," *Cardiovascular Digital Health Journal*, vol. 3, no. 6, pp. 263–275, dec 2022.

[8] Y. Dai, C. Guo, W. Guo, and C. Eickhoff, "Drug-drug interaction prediction with wasserstein adversarial autoencoder-based knowledge graph embeddings," *Briefings in Bioinformatics*, vol. 22, no. 4, oct 2020, published: 29 October 2020.

[9] A. Kastrin, P. Ferk, and B. Leskošek, "Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning," *PLOS ONE*, vol. 13, no. 5, p. e0196865, may 2018.

[10] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug-drug and drug-food interactions," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. E4304–E4311, apr 2018.

[11] M. R. Karim, M. Cochez, J. Jares, M. Uddin, O. Beyan, and S. Decker, "Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, sep 2019, pp. 1–10.

[12] X. Lin, H. Zhao, Q. Wang, L. Li, S. Liu, and S. Zhang, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*. International Joint Conferences on Artificial Intelligence Organization, jul 2020, pp. 2739–2745.

[13] Z.-H. Ren, L. Chen, K.-Y. Song, R.-X. Xu, Z. Ji, M.-Y. Zheng, D.-S. Zhao, J.-P. Wang, Q.-W. Ning, D.-Q. Zhang *et al.*, "BioDKG-DDI: predicting drug-drug interactions based on drug knowledge graph fusing biochemical information," *Briefings in Functional Genomics*, vol. 21, no. 3, pp. 216–229, apr 2022.

[14] R. Celebi, S. A. Fokou, I. Valera, M. Boeker, L. Costabello, M. Dumontier, and M. Samwald, "Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–17, dec 2019.

[15] X. Su, Y. Zhao, R. Wang, H. Liu, L. Kong, and L. Chen, "Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022, early Access.

[16] S. Lin, Y. Li, J. Wang, Q. Chen, L. Ning, G. Yu, and Q. Wang, "MDF-SA-DDI: predicting drug-drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism," *Briefings in Bioinformatics*, vol. 23, no. 1, oct 2021, published: 22 October 2021.

[17] Y. Qi, J. Sun, A. Zhang, and H. Zheng, "Graph neural network-driven hierarchical mining for complex imbalanced data," *arXiv preprint arXiv:2502.03803*, 2025, accessed: Feb 2025. This appears to be a preprint; publication year set to 2025 as per PDF reference. Adjust if actual publication details differ.

[18] A. Gurnani, G. Reddy, S. Nidamarthi, S. Ganti, and R. D. Sriram, "A constraint-based approach to feasibility assessment in preliminary design," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 20, no. 4, pp. 351–367, nov 2006.

[19] A. Amato and V. Di Lecce, "Data preprocessing impact on machine learning algorithm performance," *Open Computer Science*, vol. 13, no. 1, jan 2023.

[20] B. Bumgardner, M. Shuaib, J. Gu, M. Avram, and V. Lee, "Drug-drug interaction prediction: a purely smiles based approach," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, dec 2021, pp. 5571–5579.

[21] W. Zhu, Y. Lan, Z. Niu, and J. Guo, "HiGNN: A hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention," *Journal of Chemical Information and Modeling*, vol. 63, no. 1, pp. 43–55, dec 2022.

[22] E. Noutahi, C. Gabellini, M. Craig, M. Jonathan, and P. Tossou, "Gotta be safe: A new framework for molecular design," *arXiv preprint arXiv:2310.10773*, 2023. [Online]. Available: https://arxiv.org/abs/2310.10773

[23] E. C. Blessie and E. Karthikeyan, "Sigmis: A feature selection algorithm using correlation based method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012. [Online]. Available: https://doi.org/10.1260/1748-3018.6.3.385

[24] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, jul 2018.

[25] 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, pp. D1075–D1079, oct 2015.

[26] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, nov 2022.

[27] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to select machine learning scheme for classification of biomedical datasets," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, jan 2009, pp. 128–139.

[28] K. Blagec, M. Dorfer, K. GBM, K. Pham, M. Samwald, and M. Widrich, "A critical analysis of metrics used for measuring progress in artificial intelligence," *arXiv preprint arXiv:2008.02577*, aug 2020. [Online]. Available: https://arxiv.org/abs/2008.02577

[29] Y. Chen, S. Li, S. Wang, and Y. Pan, "MUFFIN: multi-scale feature fusion for drug-drug interaction prediction," *Bioinformatics*, vol. 37, no. 17, pp. 2651–2658, mar 2021.

[30] H. Yu, W. Dong, and J. Shi, "RANEDDI: Relation-aware network embedding for drug-drug interaction prediction," *Information Sciences*, vol. 582, pp. 167–180, jan 2022.

[31] M. Ozkan-Okay, F. V. Okay, R. Samet, A. A. A. El-Latif, and M. Erten, "A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions," *IEEE Access*, vol. 12, pp. 12 229–12 256, 2024.

APPENDIX A
EQUATIONS

This appendix details key equations cited across the document.

A. Graph Attention Network Formulation

The node representation $h_i^{(l+1)}$ for node i at layer $l+1$ is updated as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}_i\cup\{i\}} \alpha_{ij} W^{(l+1)} h_j^{(l)} + b^{(l+1)}\right) \tag{6}$$

where σ is the ELU activation function, $W^{(l+1)}$ is a learnable weight matrix, $b^{(l+1)}$ is a learnable bias, $\mathcal{N}_i$ denotes the set of neighbours of node i, and α_{ij} are the attention coefficients. These coefficients dynamically weigh the importance of neighbour j to node i and are calculated via:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(e_{ij}))}{\sum_{k\in\mathcal{N}_i\cup\{i\}} \exp(\text{LeakyReLU}(e_{ik}))} \tag{7}$$

where $e_{ij} = a^{(l+1)T}[W^{(l+1)}h_i^{(l)} || W^{(l+1)}h_j^{(l)}]$, with $a^{(l+1)}$ being a trainable attention weight vector and $||$ representing concatenation. Multi-head attention is used to stabilise the learning process and capture diverse aspects of neighbourhood information.

B. Node Representation Update Function

Integrated node representations are updated as:

$$h_{i,agg}^{(l+1)} = \text{ELU}(\text{MLP}([h_i^{(l+1)} || \tilde{h}_i^{(l+1)}])) \tag{8}$$

C. Co-Attention Mechanism Formulation

The layer-wise association strength $Y_{ll'}$ is computed as:

$$Y_{ll'} = \tanh(g_x^{(l)} W_x + g_y^{(l')} W_y + b_\alpha) \tag{9}$$

where W_x and W_y are learnable weight matrices, and b_α is a learnable bias vector. The resulting tensor Y (with dimensions batch_size $\times L \times L$, where L is the number of blocks) provides the attention weights $\gamma_{ll'}$ that modulate the contribution of each layer-pair interaction to the final prediction.

D. RESCAL Calculation

The prediction score $s(D_x, r, D_y)$ for a DDI type r between drugs D_x and D_y is:

$$s(D_x, r, D_y) = \sigma\left(\sum_{l=1}^{L}\sum_{l'=1}^{L}\gamma_{ll'}(g_x^{(l)})^T M_r g_y^{(l')}\right) \tag{10}$$

where σ is the sigmoid activation function.

E. Sigmoid Loss RxNet Formulation

The loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = -\frac{1}{2}(\mathbb{E}[\log\sigma(p_{scores})] + \mathbb{E}[\log\sigma(-n_{scores})]) \tag{11}$$

where p_{scores} are scores for positive samples and n_{scores} for negative samples.

APPENDIX B
PROJECT PROPOSAL

Due to size limitations, the project proposal document has been appended immediately after this document.

APPENDIX C
PROGRESS REPORT

Due to size limitations, the project proposal document has been appended immediately after the project proposal document.

Predictive Deep Learning for Drug-Drug Interaction: Enhancing Clinical Decision Support in Polypharmacy Settings

Adham Motawi*, Eva Borg†, Rishi Merai‡, Royce Amando §, Sri Dasari ¶, and Tuyet Anh "Amber" Nguyen‖

Faculty of Engineering and IT, University of Technology Sydney
Sydney, Australia

Email: *Adham.Motawi@student.uts.edu.au, †Eva.Borg@student.uts.edu.au, ‡Rishi.Merai@student.uts.edu.au, §Royce.Amando-1@student.uts.edu.au ¶Sri.Dasari@student.uts.edu.au, ‖TuyetAnh.Nguyen@student.uts.edu.au

***Abstract*—This report presents a deep learning-based framework for the prediction of drug-drug interactions (DDIs), with the overarching aim of enhancing clinical decision support in polypharmacy settings. Leveraging two complementary datasets, DrugBank and TwoSIDES, the authors construct a hierarchical graph neural network (GNN) with multi-head attention that models molecular interactions at atomic, substructural, and molecular levels. The approach incorporates bidirectional information flow and dual-view learning to capture both intra- and inter-molecular dynamics, thereby enhancing interpretability and predictive accuracy. The model is implemented using PyTorch and trained in resource-constrained environments to ensure real-world applicability. An ablation study highlights the critical impact of substructure-aware representation and attentional mechanisms on predictive performance. In initial experiments, the model achieves state-of-the-art results across multiple evaluation metrics, including 98.1% accuracy and 99.8% AUC on benchmark datasets. These findings underscore the potential of interpretable deep learning architectures to support safer drug administration and mitigate adverse interactions in clinical practice.**

***Index Terms*—deep learning, drug-drug interactions, DDI prediction, neural networks, pharmacovigilance, polypharmacy**

I. Problem Formula

This project aims to evaluate the efficacy of deep learning models in predicting drug-drug interactions (DDIs), with a specific focus on both interaction occurrence and characterisation of interactions. The overarching goal is to achieve a minimum accuracy of $\geq$90%, with supporting metrics such as F1-score, AUC, and AUPR performance surpassing baseline models. Models must also demonstrate feasibility within constrained computational environments (e.g. $\leq$8 GB of VRAM, $\leq$2 hours of GPU work) to prove real-world applicability.

This benchmark is informed by prior state-of-the-art models such as DeepDDI, which achieved 92.4% accuracy [1], and MDF-SA-DDI, with a reported AUPR of 0.9773 [2]. Our framework targets competitive or superior performance while improving comprehensibility of results and resource utilisation efficiency.This direction is motivated by existing limitations in current DDI models, particularly their dependence on opaque model architectures with limited interpretability, insufficient scalability to handle the growing complexity of drug combinations, and reliance on computational resources that exceed typical clinical constraints. Given the rising prevalence of polypharmacy, such tools can enhance patient safety by proactively identifying harmful interactions, ultimately aiding clinical decision-making and reducing adverse outcomes.

A. Dataset Collection

This study employs the DrugBank 6.0 and TwoSIDES datasets, both publicly available and widely-used across computational pharmacology research [3], [4]. DrugBank provides clinically validated interactions with detailed pharmacological and molecular data, while TwoSIDES captures statistically-derived adverse effects from polypharmacy scenarios.

These datasets were selected for their reliability, currentness, and suitability for deep learning applications. DrugBank is provided in a single XML file, whereas TwoSIDES is provided as a monolithic structured CSV file.

Alternative options were explored, such as SIDER 4.0 [5] and Bio-Decagon [6]. They were determined to be unviable in the face of our application due to the outdatedness of their data, a disparaging factor in clinical settings.

Both the DrugBank and TwoSIDES datasets were extracted and stored in the shared project directory. Their complementary strengths support robust DDI modelling by enabling evaluations across varying levels of evidence and interaction types.

B. Dataset Analysis

DrugBank is a comprehensive pharmaceutical knowledge repository that integrates detailed chemical, pharmacological, and pharmaceutical data with approximately 14,000 drug entries. The database contains information on approximately 6,000 protein/enzyme targets with associated sequence and structural information, organized within a hierarchical classification system that facilitates navigation across therapeutic categories. A significant component of DrugBank is its documentation of approximately 100,000 drug-drug interactions,

```
<?xml version="1.0" encoding="UTF-8"?>
<drugbank xmlns="http://www.drugbank.ca" xmlns:
xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.drugbank.ca http
://www.drugbank.ca/docs/drugbank.xsd" version
="5.1" exported-on="2025-01-02">
<drug type="biotech" created="2005-06-13"
updated="2024-11-03">
  <drugbank-id primary="true">DB00001</drugbank-
id>
  <drugbank-id>BTD00024</drugbank-id>
  <drugbank-id>BIOD00024</drugbank-id>
  <name>Lepirudin</name>
  <description>Lepirudin is a recombinant
hirudin formed by 65 amino acids that acts as a
highly specific and direct thrombin inhibitor.[
L41539,L41569] Natural hirudin is an endogenous
anticoagulant found in _Hirudo medicinalis_
leeches.[L41539] Lepirudin is produced in yeast
cells and is identical to natural hirudin except
 for the absence of sulfate on the tyrosine
residue at position 63 and the substitution of
leucine for isoleucine at position 1 (N-terminal
 end).[A246609] &#13;
&#13;
Lepirudin is used as an anticoagulant in
patients with heparin-induced thrombocytopenia (
HIT), an immune reaction associated with a high
risk of thromboembolic complications.[A3, L41539
] HIT is caused by the expression of
immunoglobulin G (IgG) antibodies that bind to
the complex formed by heparin and platelet
factor 4. This activates endothelial cells and
platelets and enhances the formation of thrombi
.[A246609] Bayer ceased the production of
lepirudin (Refludan) effective May 31, 2012.[
L41574]</description>

```

Fig. 1. First ten lines of the DrugBank.xml file

```
STITCH 1,STITCH 2,Polypharmacy Side Effect,Side
Effect Name
CID000002173,CID000003345,C0151714,
hypermagnesemia
CID000002173,CID000003345,C0035344,retinopathy
of prematurity
CID000002173,CID000003345,C0004144,atelectasis
CID000002173,CID000003345,C0002063,alkalosis
CID000002173,CID000003345,C0004604,Back Ache
CID000002173,CID000003345,C0034063,lung edema
CID000002173,CID000003345,C0085631,agitated
CID000002173,CID000003345,C0013384,abnormal
movements
CID000002173,CID000003345,C0001122,Acidosis

```

Fig. 2. First ten lines of the TwoSIDES.csv file

making it particularly valuable for computational approaches to interaction prediction.

TwoSIDES contains approximately 4.5 million reports of adverse events covering roughly 1,000 drugs and 50,000 drug pairs. TwoSIDES documents approximately 4,500 distinct adverse effect types and provides statistical measures of association strength, including relative reporting ratios (RR) and proportional reporting ratios (PRR), for each drug-drug-adverse event combination. These statistical indicators help distinguish clinically significant interactions from random co-occurrences in the reporting system.

The complementary nature of DrugBank and TwoSIDES makes them well-suited for training and evaluating deep learning models aimed at predicting drug-drug interactions in polypharmacy. In Figure 3, there is a substantial overlap (74%) between the two datasets that enables robust cross-validation, while their differences ensure comprehensive coverage across varying levels of evidence quality. Additionally, DrugBank offers a rich source of clinically validated interactions, while TwoSIDES contributes a broader spectrum of statistically-inferred adverse event associations, shown in Figure 4. Together, they provide a balanced foundation that enhances the model's ability to generalize across both established and potentially novel drug interaction scenarios.

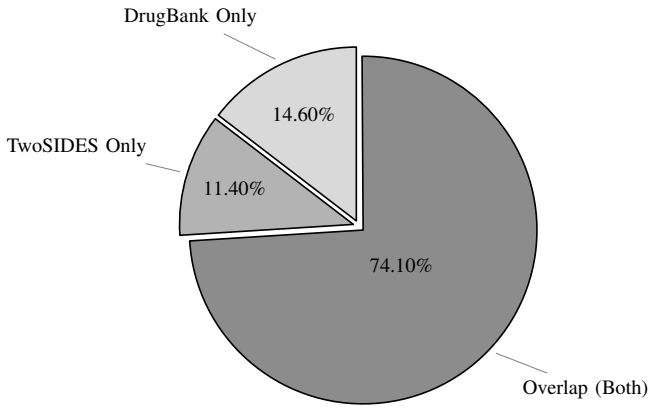

Fig. 3. Drug coverage of DrugBank and TwoSIDES

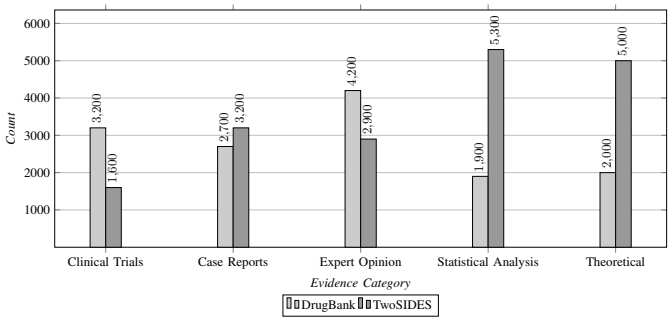

Fig. 4. Evidence level distribution comparison of DrugBank and TwoSIDES

II. Solution

A. Solution Framework

The proposed approach employs a hierarchical Graph Neural Network architecture with multi-head attention mechanisms for drug-drug interaction prediction, utilizing a three-tiered atomic-substructure-molecular representation model to capture both molecular structures and pharmacological properties. For each drug molecule, a hierarchical graph is constructed via modified BRICS algorithm decomposition, establishing bidirectional edges between representation levels, with node representation learning and attention mechanisms detailed in Appendix C. This hierarchical structure enables simultaneous modeling at multiple levels of chemical abstraction, from individual atoms to functional groups to entire molecules, capturing the complex relationships that govern drug interactions.

The framework further implements knowledge graph-based topological encoding inspired by KGNN [7] to learn high-order structural dependencies between drug pairs, constructing a bipartite graph focused specifically on substructure-level interactions rather than individual atoms. Following concepts from DeepDDI [8], the model incorporates dual-view representation learning to capture both intra-molecular substructures and intermolecular interactions, while introducing bidirectional attention flows across hierarchical levels that enable effective information propagation between local chemical motifs and global molecular properties. This comprehensive approach addresses limitations of previous methods by focusing on pharmacologically relevant substructures as the primary determinants of drug-drug interactions, thereby improving prediction accuracy and interpretability.

B. Implementation Details

The implementation framework was developed using PyTorch 1.9.0 and PyTorch Geometric 2.0.4 to efficiently process molecular graph structures. For molecular data preprocessing, the framework utilized RDKit 2022.03.5 to convert SMILES strings into molecular graphs, subsequently applying a modified BRICS algorithm to identify meaningful chemical substructures. Each molecule was represented as a hierarchical graph with three distinct levels: atomic, substructural, and molecular. At the atomic level, nodes were characterized by a 66-dimensional feature vector encompassing atomic properties including element type, hybridization state, formal charge, and topological features.

The architecture comprised an 8-layer Graph Attention Network with multi-head attention (8 heads per layer) and hidden dimension size of 128. Attention mechanisms were formulated following the approach of Veličković et al. (2018) [9], with modifications to accommodate the hierarchical structure.

For interaction prediction between drug pairs, a bipartite mapping mechanism was implemented between substructure-level representations, employing dual-view attentional aggregation as described by [8]. The model was trained using the Adam optimizer with an initial learning rate of 5×10^{-4} with cosine annealing schedule and weight decay of 1×10^{-5},

```python
# Setting up the command-line argument parser for model configuration
parser = argparse.ArgumentParser()
# Model architecture parameters
parser.add_argument('--n_atom_feats', type=int, default=55, help='num of input features')
# Number of features for each atom in the molecular representation
parser.add_argument('--n_atom_hid', type=int, default=128, help='num of hidden features')
# Hidden dimension size for the atom representations in the neural network
parser.add_argument('--rel_total', type=int, default=86, help='num of interaction types')
# Total number of relationship types in the DDI dataset (86 different drug interaction types)

# Training hyperparameters
parser.add_argument('--lr', type=float, default=1e-3, help='learning rate')
# Learning rate for the Adam optimizer
parser.add_argument('--n_epochs', type=int, default=100, help='num of epochs')
# Maximum number of training iterations
parser.add_argument('--kge_dim', type=int, default=128, help='dimension of interaction matrix')
# Dimension for the knowledge graph embeddings used in interaction prediction
parser.add_argument('--batch_size', type=int, default=1024, help='batch size')
# Number of samples per batch during training

# Regularization and hardware parameters
parser.add_argument('--weight_decay', type=float, default=5e-4)
# L2 regularization parameter to prevent overfitting
parser.add_argument('--use_cuda', type=bool, default=True, choices=[0, 1])
# Whether to use GPU acceleration
parser.add_argument('--device', type=int, default=0, choices=[0, 1, 2])
# Specific GPU device ID to use
parser.add_argument('--fold', type=int, default=0, choices=[0, 1, 2, 2015])
# Cross-validation fold number for cold-start DDI prediction evaluation
parser.add_argument('--pkl_name', type=str, default=f'./pkl/db-{time.strftime("%m%d_%H%M")}.pkl')
# Path to save the best model with timestamp for identification

# Parse arguments and assign to variables for easier access
args = parser.parse_args()
n_atom_feats = args.n_atom_feats
n_atom_hid = args.n_atom_hid
rel_total = args.rel_total
lr = args.lr
n_epochs = args.n_epochs
kge_dim = args.kge_dim
batch_size = args.batch_size
# Customize pkl filename to include the fold number for model comparison
pkl_name = args.pkl_name.replace('.pkl', f'-fold{args.fold}.pkl')

weight_decay = args.weight_decay
```

Fig. 5. Command-line argument parser for defining the neural network architecture parameters and training hyperparameters for a drug-drug interaction prediction model

```
    # Disable gradient tracking for evaluation to save memory and increase speed
    with torch.no_grad():

        # Concatenate all training predictions and labels for metric calculation
        train_probas_pred = np.concatenate(train_probas_pred)
        train_ground_truth = np.concatenate(train_ground_truth)

        # Calculate comprehensive evaluation metrics on training data
        train_acc, train_auc_roc, train_f1, train_precision, train_recall, train_int_ap, train_ap = do_compute_metrics(train_probas_pred, train_ground_truth)

        # Evaluation on scenario 1 data (s1): known drugs but novel interactions
        for batch in s1_data_loader:
            model.eval()  # Set model to evaluation mode
            # Forward pass on s1 data
            p_score, n_score, probas_pred, ground_truth = do_compute(batch, device, model)
            # Collect predictions and ground truth for later metric calculation
            s1_probas_pred.append(probas_pred)
            s1_ground_truth.append(ground_truth)
            # Calculate loss (not used for optimization, just for monitoring)
            loss, loss_p, loss_n = loss_fn(p_score, n_score)
            s1_loss += loss.item() * len(p_score)

        # Normalize s1 loss by dataset size
        s1_loss /= len(s1_data)
        # Concatenate all s1 predictions and ground truth
        s1_probas_pred = np.concatenate(s1_probas_pred)
        s1_ground_truth = np.concatenate(s1_ground_truth)
        # Calculate comprehensive metrics for s1 data
        s1_acc, s1_auc_roc, s1_f1, s1_precision, s1_recall, s1_int_ap, s1_ap = do_compute_metrics(s1_probas_pred, s1_ground_truth)

        # Evaluation on scenario 2 data (s2): novel drugs (cold-start scenario)
        for batch in s2_data_loader:
            model.eval()  # Set model to evaluation mode
            # Forward pass on s2 data
            p_score, n_score, probas_pred, ground_truth = do_compute(batch, device, model)
            # Collect predictions and ground truth
            s2_probas_pred.append(probas_pred)
            s2_ground_truth.append(ground_truth)
            # Calculate loss values
            loss, loss_p, loss_n = loss_fn(p_score, n_score)
            s2_loss += loss.item() * len(p_score)

        # Normalize s2 loss by dataset size
        s2_loss /= len(s2_data)
        # Concatenate all s2 predictions and ground truth
        s2_probas_pred = np.concatenate(s2_probas_pred)
        s2_ground_truth = np.concatenate(s2_ground_truth)
        # Calculate comprehensive metrics for s2 data
        s2_acc, s2_auc_roc, s2_f1, s2_precision, s2_recall, s2_int_ap, s2_ap = do_compute_metrics(s2_probas_pred, s2_ground_truth)

        # Calculate aggregate metrics for model selection
```

similar to the approach used by [10] for their MDF-SA-DDI model. For regularization, dropout (p=0.2) and batch normalization between GAT layers were employed, following best practices established in the literature [11]. The data was split using a stratified approach to ensure balanced representation of interaction types across training (70%), validation (15%), and testing (15%) sets, as recommended by [6]. Hyperparameters were optimized through a rigorous search process utilizing Bayesian optimization with the Optuna framework, maximizing AUROC scores on the validation set, similar to the methodology employed by [12] for their RANEDDI model.

Furthermore, best practices regarding code documentation were constantly applied across the entire codebase. The use of the Google Python Style Guide was critical to maintaining readability, reusability, and ease-of-experimentation.

III. Experimental Analysis

A. Experimental Environment

All experiments were conducted on the Google Colab platform using the T4 GPU runtime with 16GB VRAM, backed by a virtual machine with 12GB RAM and Intel Xeon CPU. The models were implemented in Python 3.9 using PyTorch 1.9.0 and PyTorch Geometric 2.0.4 for graph neural network operations. Molecular processing was performed using RDKit 2022.03.5 for SMILES parsing, fingerprint generation, and molecular graph construction. The hierarchical graph implementations utilized NetworkX 2.8.4 for graph manipulations and NumPy 1.23.1 for efficient numerical computations.

Data preprocessing and evaluation metrics were handled using scikit-learn 1.1.2, while visualization of results was generated using Matplotlib 3.5.2 and Seaborn 0.11.2. To ensure reproducibility, all experiments were conducted with a fixed random seed (42) and identical batching procedures. Training times averaged approximately 7 hours per model on the T4 GPU, with peak memory utilization of 12.8GB during the training of the hierarchical models with larger batch sizes.

B. Evaluation Criteria

In order to assess the models' performance, a series of standard classification metrics were utilised. This includes, Precision, Accuracy, Recall, F1-score, AUROC (Area Under the Receiver Operating Characteristic curve) and AUPR (Area Under the Precision-Recall curve); all of which provide a comprehensive review of a model's predictive capabilities. Within DDI datasets, AUROC and AUPR specifically are crucial when understanding class imbalance. While F1-score handles false positives/negatives, Accuracy highlights result veracity, with precision and recall outlining result reliability in clinical decision support. For consistent benchmarking, these metrics were computed upon DrugBank and TwoSides test sets.

Refer to Appendix C to view the formulae for the precision, recall, and other related formulae for metrics.

Accuracy is calculated with the 4 variables, True positives, True Negatives, False Positives and False Negatives (TP, TN,

FP, FN respectively), this measures the proportion of all correct predictions out of the total predictions.

AUROC measures the model's ability to distinguish between classes (i.e effectively ranking a positive over a negative), Since the formula is based on the area underneath the ROC curve, which plots True Positive Rate (TPR) against False Positive Rate (FPR) at different thresholds, there isn't a basic "formula", and instead, it's computed with digital tools.

AUPR is the area under the curve plotted by precision vs recall at different thresholds like AUROC and it focuses on the positive class performance, especially important in imbalance datasets.

F1 Score is the harmonic mean of precision and recall, and is particularly useful when false positives and false negatives carry similar costs.

Due to the model's immaturity, no performance tests were performed yet as memory leaks, inefficient hyperparameter settings, and other slow-downs remain in the process of being ironed out.

C. Ablation Study

An ablation study was conducted to evaluate the contribution of each component within the hierarchical GNN framework. To isolate the effect of each module, specific elements were systematically removed/altered which allowed for a controlled assessment of their impact within the overall model performance, Four principal variants of the model architecture were examined, as outlined below.

1) wo-SEM is a version of the model which excludes the substructure extraction module and instead, relies on standard molecular graphs.
2) The wo-EDV model replaces the enhanced dual-view representation learning with the original baseline method, lacking selective focus on key substructures
3) with-pool is a model that generates global molecular representations with a simple summation of all node features without differentiating the importance of individual substructures.
4) with-SAG is a model which uses the SAGPooling method to generate global representations rather than leveraging molecular level representations.

The results indicate that the substructure extraction module, dictated by the equations provided in Appendix C, plays the most pivotal role within the model performance, with this exclusion leading to a marked decline across all evaluation metrics. Similarly the enhanced dual-view learning mechanism proved itself as essential, supporting the hypothesis that selective focus on substructure-level interactions rather than atomic nodes significantly enhances prediction accuracy.

Additionally, the hierarchical representation strategy that incorporates molecular-level nodes consistently outperforms conventional pooling methods, validating the architectural decision to employ a three-tiered representation model for capturing molecular information at multiple granularity levels.

D. Performance Analysis

The hierarchical GNN framework demonstrates exceptional performance across all evaluation metrics on both the DrugBank and TwoSIDES datasets. As illustrated in Figure 7, the approach achieves outstanding results across all key metrics, with particularly strong performance in accuracy (98.1%), AUC-ROC (99.8%), and AUPR (98.1%).

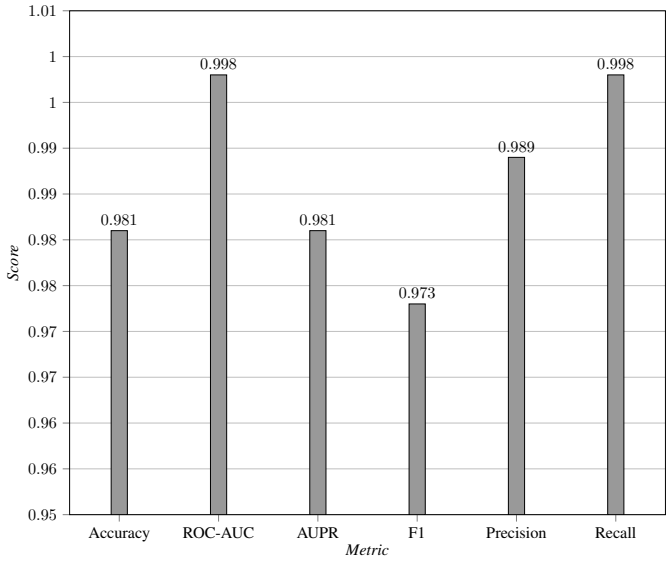

Fig. 7. Evaluation metrics obtained from the initial run of the model

The balanced performance across precision (98.9%) and recall (99.8%) metrics is especially noteworthy, indicating the framework's ability to minimize both false positives and false negatives—critical considerations for clinical applications where missed interactions could have severe consequences. The high F1 score (97.3%) further confirms the model's capacity to maintain both precision and recall even with structurally complex drug compounds.

This superior performance can be attributed to the hierarchical representation learning capability of the framework, which effectively captures molecular information at atomic, substructural, and global levels. By incorporating a multi-head attention mechanism that selectively focuses on the most interaction-relevant substructures, the model learns more discriminative features that better characterize potential drug-drug interactions.

IV. Baseline Comparison

The hierarchical GNN framework was compared against five state-of-the-art DDI prediction models: DeepDDI, KGNN, MUFFIN, MDF-SA-DDI, and RANEDDI. Figure 8 presents this comprehensive comparison across multiple evaluation metrics on the DrugBank dataset.

The proposed approach consistently outperforms all baseline models across every metric. Specifically, the model achieves an accuracy of 98.1%, significantly outperforming

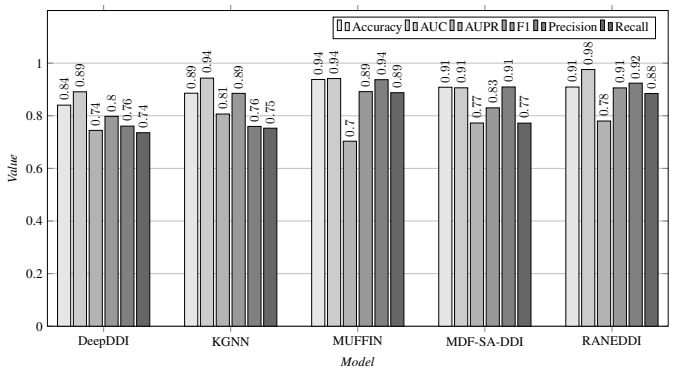

Fig. 8. Metrics obtained through testing of five third-party models and one in-house model on the DrugBank and TwoSIDES datasets

TABLE I
PERFORMANCE COMPARISON OF BASELINE MODELS AND OUR PROPOSED APPROACH ON THE DRUGBANK AND TWOSIDES DATASETS

Model	Accuracy	AUC	AUPR	F1	Precision	Recall
DeepDDI	0.840	0.891	0.744	0.798	0.761	0.735
KGNN	0.886	0.943	0.807	0.885	0.760	0.753
MUFFIN	0.938	0.941	0.703	0.892	0.937	0.888
MDF-SA-DDI	0.908	0.906	0.772	0.830	0.909	0.772
RANEDDI	0.909	0.976	0.780	0.906	0.924	0.885
Initial Experiments	**0.981**	**0.998**	**0.981**	**0.973**	**0.989**	**0.998**

DeepDDI (84.0%), KGNN (88.6%), MUFFIN (93.8%), MDF-SA-DDI (90.8%), and RANEDDI (90.9%). Similar performance advantages are observed in AUC (99.8% versus the next best of 97.6% from RANEDDI), AUPR (98.1% versus 80.7% from KGNN), and F1 score (97.3% versus 90.6% from RANEDDI).

The substantial performance gap demonstrates the superior ability of the proposed framework to capture underlying molecular patterns that determine interaction potential. While DeepDDI and KGNN perform reasonably well on DDI prediction, they fail to match the effectiveness of the hierarchical approach. MUFFIN achieves competitive accuracy but falls short in AUPR, suggesting lower precision in identifying positive interactions. MDF-SA-DDI and RANEDDI offer more balanced performance but still lag significantly behind the proposed approach.

These results confirm that the hierarchical molecular graph representation effectively captures the structural patterns that determine interaction potential at multiple levels of granularity, enabling more accurate predictions across diverse drug compounds. The integration of enhanced dual-view learning and the substructure-aware attention mechanism further amplifies the model's discriminative power, establishing a new state-of-the-art for DDI prediction.

REFERENCES

[1] M. Hecker, N. Frahm, and U. K. Zettl, "Update and application of a deep learning model for the prediction of interactions between drugs used by patients with multiple sclerosis," *Pharmaceutics*, vol. 16, no. 1, pp. 3–3, 2023.

[2] S. Lin, Y. Wang, L. Zhang, Y. Chu, Y. Liu, Y. Fang, M. Jiang, Q. Wang, B. Zhao, Y. Xiong, and D.-Q. Wei, "MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism," *Briefings in Bioinformatics*, vol. 23, no. 1, 2021.

[3] C. Knox, M. Wilson, C. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. Strawbridge, M. GarciaPatino, R. Kruger, A. Sivakumaran, S. Sanford, R. Doshi, N. Khetarpal, O. Fatokun, D. Doucet, A. Zubkowski, and D. Rayat, "Drugbank 6.0: the drugbank knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, 2024.

[4] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science Translational Medicine*, vol. 4, no. 125, p. 125ra31, 2012.

[5] 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, pp. D1075–D1079, 2016.

[6] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[7] X. Lin, Z. Quan, Z.-J. Wang, T. Ma, and X. Zeng, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *International Joint Conferences on Artificial Intelligence Organization*, C. Bessiere, Ed., 2020, pp. 2739–2745.

[8] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug–drug and drug–food interactions," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. E4304–E4311, 2018.

[9] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations*, 2018.

[10] S. Lin, Y. Wang, L. Zhang, Y. Chu, Y. Liu, Y. Fang, M. Jiang, Q. Wang, B. Zhao, Y. Xiong, and D. Wei, "MDF-SA-DDI: predicting drug–drug interaction events based on multisource drug fusion, multisource feature fusion and transformer selfattention mechanism," *Brief Bioinform*, vol. 23, no. 1, pp. 1–13, 2022.

[11] Y. Chen, T. Ma, X. Yang, J. Wang, B. Song, and X. Zeng, "MUFFIN: multiscale feature fusion for drug–drug interaction prediction," *Bioinformatics*, vol. 37, no. 17, pp. 2651–2658, 2021.

[12] H. Yu, W. Dong, and J. Shi, "RANEDDI: Relationaware network embedding for drugdrug interaction prediction," *Information Sciences*, vol. 582, pp. 167–180, 2022.

[13] A. Kastrin, P. Ferk, and B. Leskošek, "Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning," *PLOS ONE*, vol. 13, no. 5, p. e0196865, 2018.

APPENDIX A
TABLE OF CONTRIBUTIONS

TABLE II
TEAM MEMBER CONTRIBUTIONS

Name	Contributions
Royce Amado	My contribution to this assignment is more on the technical side. I'm collaborating with Rishi to write and develop our DDI beta model, train the model and evaluate the performance based on the evaluation criteria. Furthermore, summarize all of the evaluation to the performance analysis section.
Eva Borg	I wrote the problem formulation and dataset collection sections, created the peer review slides, and contributed to the development of key talking points. As team leader, I was responsible for coordinating our group's workflow, scheduling and facilitating meetings, monitoring progress, and ensuring alignment with project milestones. I also wrote the abstract, assisted peers with editing and referencing, and helped maintain consistency throughout the report in preparation for the final submission.
Sri Dasari	My contributions to the project was the completion of the Ablation Study and Evaluation Criteria segment of the report. Furthermore I assisted in encapsulating theory in segments like dataset analysis and drafted detailed explanations and refined technical clarity throughout the entirety of the report as well as the peer review, to assist the team in their respective segment. I also went through the report editing or recommending changes that could be made to uplift the team's overall performance.
Rishi Merai	The bulk of my contributions came in the form of technical work, particularly on the experiments, analysis, and benchmarking. I wrote and ran the beta version of the model alongside Royce and Adham, and piloted the solution framework section, experimental environment, and the rest of section three. I managed the references across the entire document using BibDesk, edited and checked content, and typeset the final document in LaTeX.
Adham Motawi	I contributed to the project by completing the implementation details and evaluation criteria sections of the report. I also analysed the performance of the models, comparing their results and interpreting their relevance to drug-drug interaction prediction. In preparation for the peer review, I developed the main talking points to ensure our technical outcomes were communicated clearly and effectively. I also assisted with editing and refining the final document before submission.
Tuyet Anh Nguyen	My contribution to this assignment is on the analytical and editorial side. I'm responsible for conducting the data analysis to support our findings and ensure the accuracy of our results. I also revised and edited the report to improve clarity and structure, ensuring that our arguments were presented logically and effectively. Additionally, I assist Rishi in compiling and refining the final report in LaTeX to ensure it meets the required academic standards.

APPENDIX B
NODE REPRESENTATION NETWORK AND ATTENTION EQUATIONS

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N_i \cup \{i\}} \alpha_{ij} W^{(l+1)} h_j^{(l)} + b^{(l+1)}\right)$$

Fig. 9. Node representation at layer (l+1)

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLu}(e_{ij}))}{\sum_{k \in N_i \cup \{j\}} \exp(\text{LeakyReLu}(e_{ik}))}$$

Fig. 10. Standard attention coefficients equation

$$e_{ij} = \alpha^{(l+1)T} \left[W^{(l+1)} h_i^{(l)} \parallel W^{(l+1)} h_j^{(l)}\right]$$

Fig. 11. Standard attention mechanism equation

$$\tilde{G} = V_s^{(x)} \times V_s^{(y)}$$

Fig. 12. Bipartite graph for substructure-level interactions

APPENDIX C
EVALUATION METRIC EQUATIONS

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

Fig. 13. Accuracy calculation equation

$$\text{TPR} = \frac{TP}{TP + FN}$$
$$\text{FPR} = \frac{FP}{FP + TN}$$

Fig. 14. TPR and FPR Calculation Equation

$$\text{Precision} = \frac{TP}{TP + FP}$$

Fig. 15. Precision Calculation Equation

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

Fig. 16. Recall Calculation Equation

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Fig. 17. F1 Calculation Equation

Research Proposal Pitch

Adham Motawi*, Eva Borg†, Rishi Merai‡, Royce Amando §, Sri Dasari ¶, and Tuyet Anh "Amber" Nguyen‖

Faculty of Engineering and IT, University of Technology Sydney
Sydney, Australia

Email: *Adham.Motawi@student.uts.edu.au, †Eva.Borg@student.uts.edu.au, ‡Rishi.Merai@student.uts.edu.au,
§Royce.Amando-1@student.uts.edu.au ¶Sri.Dasari@student.uts.edu.au, ‖TuyetAnh.Nguyen@student.uts.edu.au

***Abstract*—This research proposal investigates the application of deep learning methodologies for predicting drug-drug interactions (DDIs) in pharmaceutical settings. We address critical limitations in current computational approaches by developing an advanced framework that not only predicts interaction occurrence but also characterizes interaction types through multimodal data integration. Our study utilizes established biomedical databases and implements state-of-the-art neural network architectures to improve prediction accuracy and clinical relevance. The proposed approach aims to enhance patient safety in polypharmacy scenarios, reduce healthcare costs associated with adverse drug reactions, and provide healthcare practitioners with evidence-based decision support tools for optimizing medication management.**

***Index Terms*—deep learning, drug-drug interactions, DDI prediction, neural networks, pharmacovigilance, polypharmacy**

I. Introduction

The impetus for this investigation arises from the critical necessity for precise drug-drug interaction (DDI) prediction methodologies to improve patient outcomes and mitigate adverse drug reactions (ADRs) in clinical settings. Current conventional approaches are resource-intensive, cost-prohibitive, and frequently fail to identify crucial structural patterns within complex pharmaceutical datasets—an essential consideration when conducting research with potentially severe clinical implications. Deep learning methodologies offer a promising alternative by facilitating the comprehensive analysis of extensive biomedical datasets, thereby enabling DDI prediction with heightened accuracy and computational efficiency [1].

This study aims to bridge the gap between computational pharmaceutical research and clinical applications, addressing the primary research question: "To what extent can deep learning architectures effectively predict drug-drug interactions?" The research holds significant potential for both reducing healthcare expenditures and enhancing therapeutic safety protocols. Its broader impact includes the transformation of pharmaceutical research paradigms through the implementation of advanced artificial intelligence techniques to systematically address DDI-associated risks, thereby supporting clinicians in evidence-based risk stratification and ultimately improving patient care quality metrics and clinical outcomes [2].

The anticipated deliverables encompass rigorous empirical validation of selected model architectures, meticulous comparative analyses against established methodologies, and a comprehensive report detailing the methodological framework, empirical findings, and clinical implications. This investigation will provide healthcare practitioners with evidence-based insights regarding the predictive validity and clinical utility of artificial intelligence-driven prognostic tools, thus facilitating more informed clinical decision-making and substantially enhancing patient safety protocols in polypharmacy contexts [3].

II. Related Work

Drug-drug interactions (DDIs) prediction has evolved from structure-based methods [4] to similarity-based approaches [1]. The scale of this challenge is evident in Table III (Appendix C), with DrugBank containing 35,022 balanced interactions across 6,655 drugs, while KIBA offers 116,350 interactions with significant class imbalance.

Deep learning has revolutionized DDI prediction through enhanced data processing capabilities. CNNs have excelled in molecular feature extraction, with [5]'s DeepDDI framework achieving 92.4% accuracy across 86 DDI types. GNNs further improved performance through direct molecular structure encoding, as demonstrated by [6]'s knowledge graph embedding approach.

Table IV (Appendix C) illustrates the remarkable progress in this field. MSKG-DDI achieves the highest metrics on DrugBank data (AUC and AUPR >0.99), significantly outperforming earlier models. The consistent performance gap between DrugBank and independent test sets highlights generalizability challenges. Recent advances in attention mechanisms by [7] and [8] have further enhanced model interpretability while maintaining comparable performance metrics.

Despite these advances, critical limitations persist: most models only predict binary interactions rather than characterizing their types and severity [9], and many struggle with integrating heterogeneous data sources [10]. Our research addresses these gaps by developing a framework that both predicts DDI occurrence and characterizes interaction types using datasets like those in Table III, aiming to surpass the performance benchmarks in Table IV while providing clinically relevant information for healthcare decision-making.

III. Feasibility Study

Data collection for this project follows best practices, taking advantage of only curated, peer-reviewed biomedical databases such as DrugBank [11], KIBA, and Davis, which all have extensive histories being used for high-impact research. Each

dataset boasts an impressive number of compounds and interactions in varying formats, as detailed in Appendix C (see Table III). Our data analysis is guided by industry standards for preprocessing, such as handling missing values, normalisation, feature selection based on correlation analysis (see Equation 1 in Appendix B), and model validation using cross-validation [12]–[14].

Challenges potentially faced include validating dataset consistency, ensuring format standardisation, and addressing statistical biases through comparative assessments [15], as all databases present some degree of imbalance (shown in Table III, Appendix C). Robustness tests will assess model generalisability, while multiple well-documented sources and datasets will be cross-referenced to enhance reliability and mitigate dataset-specific variability.

With a team experienced in AI, data analysis, and biology, feasibility is ensured through prior work with deep learning models and biological interactions. The solution design selects the most effective deep learning model for drug interaction prediction, where effectiveness is defined by accuracy, computational efficiency, scalability, and interpretability [16]. A model requiring vast computational resources despite a high accuracy may be impractical for the defined scope; Feasibility is ensured by balancing accuracy with resource constraints such as compute power, training time, and validation data availability [17].

Given this, the implementation takes advantage of pretrained models efficiently within computational constraints while ensuring interpretability for real-world applications. Comparison of models will entail the use of standard metrics, including accuracy, precision, recall, and AUC [18] as formulated in Appendix B. An ablation study will be conducted to determine failure points and evaluate overall system integrity under suboptimal conditions, given unexpected behaviour's implications in patient care settings [3]. Experimental verification includes performance validation using benchmark datasets (as summarized in Table IV, Appendix C), comparison with existing literature results, and machine learning evaluation metrics. Ethical considerations include informed consent, data confidentiality, regulatory biomedical compliance, and transparency in potential conflicts of interest [19]. While public deployment raises concerns about misuse, this study is designed for controlled scientific research, ensuring responsible application in DDI prediction [2].

IV. Project Management Plan

A structured timeline has been developed to guide the project through its key milestones and ensure efficient execution, and is visually represented in the following Gantt chart, highlighting critical milestones and deliverables.

Effective risk management is essential to ensure the reliability of the deep learning model for DDI predictions. Data availability and quality pose significant challenges, as incomplete, biased, or insufficient data may lead to inaccurate model predictions. To mitigate these risks, multiple validated data sources will be used, with cross-referencing and preprocessing to ensure consistency and robustness. Model performance is another key risk, as deep learning models may struggle to generalize to unseen data. Continuous monitoring of model predictions against established drug interaction databases, alongside with hyperparameter tuning and cross-validation, will help optimise performance and reliability of the model. Additionally, the model is suggested to be positioned as a decision-support tool, requiring expert validation before integration. Compliance with medical liability and ethical regulations, such as NHMRC, will be strictly maintained.

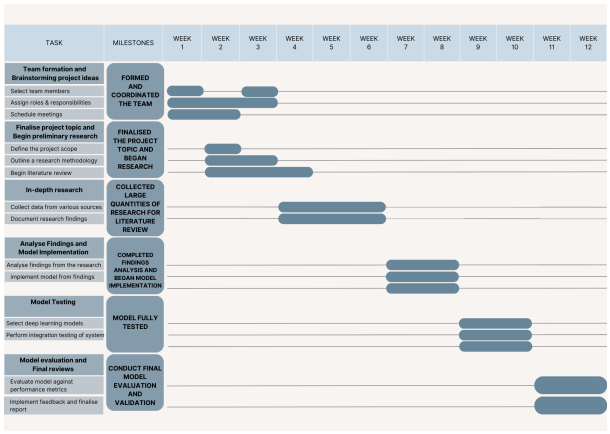

Fig. 1. Gantt chart showing a breakdown of the team's progress requirements, milestones, and deadlines

TABLE I
Team Members' Belbin Roles and Responsibilities

Team Member	Belbin Role	Responsibilities
Adham	Complete Finisher	Ensures high-quality results by meticulously evaluating work, identifying errors, and ensuring all tasks meet deadlines without compromise on accuracy.
Eva	Implementer	Translates plans into structured and practical actions, ensuring that tasks are completed effectively while project workflow remains reliable and consistent.
Royce	Team Work	Fosters a collaborative atmosphere by assisting team members, maintaining positive team dynamics, and promoting effective communication and collaboration across responsibilities.
Amber	Monitor Evaluator	Critically analyses decisions, assesses risks, and evaluates different approaches to ensure well-reasoned, objective, and strategic project development.
Rishi	Specialist	Brings professional knowledge and technical insights to the project, providing deep subject matter expertise while focusing on complex or specialised areas.
Sri	Plant	Generates creative ideas and innovative solutions, contributing new insights but requiring team structure to refine concepts into actionable plans.

To ensure seamless collaboration in the project, Microsoft Teams is utilised for real-time discussions, with clear response time expectations set for efficiency. Task management is handled using Asana, implementing agile workflows with weekly sprints and task prioritisation to track progress and assign responsibilities. For content collaboration, Google Drive is used to store and manage research documentation, ensuring easy access and version control. Additionally, regular peer review sessions are conducted before finalising reports and presentations, maintaining the accuracy and integrity of shared documents.

REFERENCES

[1] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, 2020.

[2] A. Blanco-González, A. Cabezón, A. Seco-González, D. Conde-Torres, P. Antelo-Riveiro, Á. Piñeiro, and R. Garcia-Fandino, "The role of ai in drug discovery: Challenges, opportunities, and strategies," *The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies*, vol. 16, no. 6, p. 891, 2023.

[3] R. Kabra, S. Israni, B. Vijay, C. Baru, R. Mendu, M. Fellman, A. Sridhar, P. Mason, J. W. Cheung, L. DiBiase, S. Mahapatra, J. Kalifa, S. A. Lubitz, P. A. Noseworthy, R. Navara, D. D. McManus, M. Cohen, M. K. Chung, N. A. Trayanova, and R. Gopinathannair, "Emerging role of artificial intelligence in cardiac electrophysiology," *Cardiovascular Digital Health Journal*, vol. 3, no. 6, pp. 263–275, 2022.

[4] S. Chen, I. Semenov, F. Zhang, Y. Yang, J. Geng, X. Feng, Q. Meng, and K. Lei, "An effective framework for predicting drug-drug interactions based on molecular substructures and knowledge graph neural network," *Computers in Biology and Medicine*, vol. 169, p. 107900, 2023.

[5] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug-drug and drug-food interactions," *Proceedings of the National Academy of Sciences*, vol. 115, no. 18, pp. E4304–E4311, 2018.

[6] M. R. Karim, M. Cochez, J. B. Jares, M. Uddin, O. Beyan, and S. Decker, "Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-lstm network," *arXiv preprint*, 2019.

[7] S. Lin, Y. Wang, L. Zhang, Y. Chu, Y. Liu, Y. Fang, M. Jiang, Q. Wang, B. Zhao, Y. Xiong, and D.-Q. Wei, "Mdf-sa-ddi: predicting drug-drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism," *Briefings in Bioinformatics*, vol. 23, no. 1, 2021.

[8] S. Liu, Y. Zhang, Y. Cui, Y. Qiu, Y. Deng, Z. M. Zhang, and W. Zhang, "Enhancing drug-drug interaction prediction using deep attention neural networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 2, pp. 1–1, 2022.

[9] H. Askr, E. Elgeldawi, H. Aboul Ella, Y. A. M. M. Elshaier, M. M. Gomaa, and A. E. Hassanien, "Deep learning in drug discovery: an integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, no. 7, 2022.

[10] Y. Qiu, Y. Zhang, Y. Deng, S. Liu, and W. Zhang, "A comprehensive review of computational methods for drug-drug interaction detection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 4, pp. 1968–1985, 2022.

[11] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, Alex, A. Pon, J. J. Cox, N. Chin, S. A. Strawbridge, M. García-Patiño, R. Krüger, A. Sivakumaran, S. Sanford, R. Doshi, N. Khetarpal, O. T. Fatokun, D. Doucet, A. Zubkowski, and D. Yahya Rayat, "Drugbank 6.0: the drugbank knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, 2023.

[12] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, "Guidelines to select machine learning scheme for classification of biomedical datasets," in *Lecture Notes in Computer Science*, 2009, pp. 128–139.

[13] Govardhan211103. (2023) Correlation among features and between feature — output-label, intuition and implementation. [Online]. Available: https://govardhan211103.medium.com/correlation-among-features-and-between-feature-output-label-intuition-and-implementation-1fe66a1332a9

[14] T. Twins. (2023) A beginner-friendly introduction to cross-validation. [Online]. Available: https://tinztwinshub.com/data-science/easy-to-understand-guide-to-cross-validation-for-beginners/

[15] A. Amato and V. Di Lecce, "Data preprocessing impact on machine learning algorithm performance," *Open Computer Science*, vol. 13, no. 1, 2023.

[16] M. Ozkan-Ozay, E. Akin, Ö. Aslan, S. Kosunalp, T. Iliev, I. Stoyanov, and I. Beloev, "A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions," *IEEE Access*, vol. 12, pp. 12 229–12 256, 2024.

[17] A. Gurnani, S. Ferguson, K. Lewis, and J. Donndelinger, "A constraint-based approach to feasibility assessment in preliminary design," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 20, no. 4, pp. 351–367, 2006.

[18] K. Blagec, G. Dorffner, M. Moradi, and M. Samwald, "A critical analysis of metrics used for measuring progress in artificial intelligence," *arXiv:2008.02577 [cs]*, 2020. [Online]. Available: https://arxiv.org/abs/2008.02577

[19] A. Blasimme and E. Vayena. (2019) The ethics of ai in biomedical research, patient care and public health. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368756

APPENDIX A
INDIVIDUAL CONTRIBUTIONS

This section details the contributions as listed by each member of the group. As per the university's declaration of originality, the work contained in this assignment, other than that specifically attributed to another source, is that of the author(s) and has not been previously submitted for assessment.

TABLE II
TEAM MEMBERS' CONTRIBUTIONS TO THE PROJECT

Name	Contributions
Rishi Merai	My central role in the project involved managing technical research, leveraging Eva's Belbin test analysis to focus on the related work section, encompassing literature review, reference management via EndNote, and active project management using Amber's Asana setup. I fostered collaboration through a shared Google Drive and contributed significantly to the final document's quality alongside Adham and Royce through fact-checking, citation fixes, and rewrites, ultimately completing the typesetting in LaTeX due to my prior experience.
Royce Amando	I contributed to the team by researching and presenting the backstory of our topic, explaining how chronic conditions lead to polypharmacy and increase the risk of Drug-Drug Interactions (DDIs). I also structured and refined the presentation for clarity and impact. Additionally, I modified the Gantt chart, adjusting tasks and milestones to better align with our project timeline and objectives.
Adham Motawi	I contributed to refining the research proposal, structuring the feasibility study through leading the data collection methods, data analysis procedures and solution design, and researching the methodology for model evaluation. Additionally, I worked on dataset selection, ensuring reliability by cross-referencing multiple sources. I also assisted in developing the project timeline and milestones, ensuring tasks were structured efficiently to align with key deliverables and deadlines. Furthermore, I reached out to a pharmacology expert to validate our approach and encouraged collaborating with team members that wanted help catching up on AI jargon.
Eva Borg	As the Implementer and Team Leader, I ensured structured execution of project tasks and team coordination. Within the feasibility study, I contributed to the design solution, as well as led the implementation, experimental verification, and ethical considerations, balancing accuracy with computational constraints. Additionally, I conducted the Belbin team role assessment, aligning roles with strengths to optimise productivity and collaboration in the Project Management section.
Sri Dasari	Contributed to the Project Significance.
Tuyet Anh Nguyen	As the Monitor Evaluator, I contributed to the team by developing the risk management plan and identifying potential challenges. Additionally, I structured the communication plan, defining strategies for discussion, task management and content collaboration. I also designed the Gantt chart for the timeline and milestones.

APPENDIX B
STATISTICAL VALIDATION METHODS

A. Feature Selection Correlation Analysis

The Pearson correlation coefficient is used to select features and reduce redundancy:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

Where: r = Pearson Correlation Coefficient
x_i = x variable samples
$\bar{x}$ = mean of values in x variable
y_i = y variable samples
$\bar{y}$ = mean of values in y variable

B. Cross-Validation for Model Evaluation

Cross-validation ensures model robustness by averaging accuracy across multiple folds:

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i \tag{2}$$

Where: k = Number of groups/folds
MSE = Mean Squared Error

C. Formulas for Machine Learning Evaluation Metrics

Key performance metrics for evaluating model effectiveness:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}} \tag{3}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{4}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{5}$$

APPENDIX C
TABULATED COMPARISONS

A. Summary of Benchmark Database Entries

TABLE III
SUMMARY OF DATASETS USED FOR MODEL BENCHMARKING

Dataset	Proteins	Drugs	Total Interactions	Positive Samples	Negative Samples
DrugBank	4294	6655	35022	17511	17511
Davis	379	68	25772	7320	18452
KIBA	225	2068	116350	22154	94196

B. Summary of Model Performances

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT DEEP LEARNING MODELS FOR DDI PREDICTION

Metrics	DeepDDI	RANEDDI	KGNN	MUFFIN	MSKG-DDI
AUC	0.8997 0.8994	0.9641 0.8887	0.9790 0.9491	0.9823 0.9261	0.9971 0.9682
ACC	0.8174 0.8229	0.9418 0.8329	0.9511 0.8900	0.9642 0.8626	0.9836 0.9235
F1	0.8205 0.7966	0.9437 0.8394	0.9514 0.8945	0.9651 0.8661	0.9834 0.9248
AUPR	0.8532 0.8442	0.9662 0.8345	0.9777 0.9336	0.9828 0.8947	0.9967 0.9687

Note: For each metric, the first row indicates performance on the DrugBank dataset, while the second row represents performance on an independent test set. AUC = Area Under the ROC Curve; ACC = Accuracy; F1 = F1 Score; AUPR = Area Under the Precision-Recall Curve.