Comparative Learning Method for Drug Target Interaction Prediction Based on Heterogeneous Network

Qizhao Li *

Department of Information and Computational Science, Guangdong Ocean University, Guangdong, China

* Corresponding Author Email: 11921313hh1@stu.gdou.edu.cn

Abstract. Drug-Target Interaction (DTI) prediction serves as a pivotal task for accelerating drug repositioning and novel drug development, significantly reducing the time and costs of traditional experiments. Heterogeneous graph neural networks (hgnns) have become mainstream due to their capability to integrate multi-source biological data across "drug-target-disease-side effect" relationships. This paper examines the three-stage evolutionary trajectory of dti prediction methods using 2024's hncl-dti anchor papers as core references. It clarifies the characteristics of common datasets like hbn-a/b and evaluation metrics such as auc and aupr, while focusing on current bottlenecks including dataset scale bias, insufficient staticity, poor domain-specific model adaptation, and lack of interpretability. by comparing vertical domain solutions with general-domain approaches from the 2024-2025 ccf category a conference, the study evaluates the applicability of techniques like inductive learning and bias-corrected sampling. The core conclusion proposes integrating biomedical characteristics with general technologies to construct multi-source dynamic datasets and knowledge-enhanced hgnns. Future efforts should focus on knowledge-enhanced heterogeneous graph learning by deeply embedding biological knowledge (e.g., drug mechanisms and target annotations) into models, combining general-domain adaptive projection and attentional attribution techniques. This will establish a dual-dimensional evaluation system for "predictive performance and clinical applicability," ultimately achieving technical goals of "high precision, high interpretability, and high inductiveness" to provide more practical support for drug development.

Keywords: Drug-target interaction prediction, Heterogeneous graph neural network, Contrast learning, Model optimization.

1. Introduction

Drug development remains a cornerstone of human healthcare, yet the traditional wet-lab approach faces critical challenges: The 10-15 year journey from compound screening to clinical approval costs over $2 billion per drug candidate, with success rates below 10% [1]. This reality has propelled drug repositioning – leveraging the core strength of "expanding approved drug indications" – to become a game-changer for cutting R&D costs and shortening timelines. Research by Nosengo et al. reveals that approximately 75% of marketed drugs possess potential for therapeutic expansion [1]. As the foundational technology for drug repositioning, Drug-Tissue Interactions (DTI) prediction employs computational methods to map drug-target binding patterns, dramatically reducing experimental validation scope and boosting R&D efficiency by several orders of magnitude.

Early DTI prediction methods were constrained by technical frameworks and struggled to meet practical demands. Molecular docking techniques (e.g., AutoDock4 [2]) relied on the three-dimensional structures of target proteins, proving completely ineffective for targets with missing structures, while each docking calculation required hours, making them unsuitable for large-scale screening scenarios. Ligand-based approaches (such as the chemical similarity model proposed by Keiser et al. [3]) followed the "similar drug acting on similar target" assumption. Although this reduced computational costs, they failed entirely when dealing with novel chemical entities, demonstrating severe generalization limitations. After 2015, the explosive growth of biomedical data marked a technological turning point for DTI prediction. Heterogeneous network technology emerged as a breakthrough solution capable of modeling complex relationships between multiple types of entities, with the rise of HGNN further driving performance advancements in this field.

The field of DTI prediction has witnessed a surge in research integrating heterogeneous networks with deep learning. Luo et al. 's 2017 work, DTINet [4], was the first to apply heterogeneous networks for DTI prediction, achieving an AUC of 0.8838 on the HBN-A dataset while neglecting semantic differences in edge types. Li et al.' s 2022 model, HGAN, introduced a heterogeneous graph attention mechanism that boosted the AUC to 0.9247 on the HBN-B dataset but failed to incorporate edge features [5]. In the same year, their subsequent work SGCL-DTI incorporated contrastive learning, attaining an AUC of 0.9391 on HBN-A while ignoring relational entity information. Zhang et al.'s 2023 study [6], IMCHGAN [7], attempted multi-view contrastive learning but struggled with its black-box architecture, making it unsuitable for clinical applications.

The HNCL-DTI method proposed by Hu et al. in 2024 achieves current optimal performance of AUC 0.9566 and 0.9290 on the HBN-A/B dataset through a "node attention + edge attention + cross-viewpoint contrast learning" mechanism [8]. However, two core bottlenecks remain in DTI prediction: At the dataset level, benchmark datasets like HBN-A/B suffer from issues such as insufficient scale coverage, data bias, and lack of staticity; at the model level, existing methods demonstrate inadequate adaptability to "biological semantic prioritization" and "feature heterogeneity," while generally lacking clinical interpretability.

In light of the aforementioned context, this paper systematically investigates limitations and proposes solutions in DTI prediction using HNCL-DTI as the anchor. Subsequent sections will outline methodological evolution, analyze the compatibility of datasets and evaluation criteria, quantify technical capabilities, identify core limitations, compare proposed solutions, and ultimately provide future perspectives to establish pathways and theoretical foundations for technological breakthroughs in DTI prediction.

2. Overview of Mainstream Dti Prediction Methods in Recent Years

The evolution of DTI prediction method can be divided into three stages, which generally shows the trend of "from single feature to multi-source fusion and from static modeling to dynamic learning"

2.1. Phase 1: Traditional Calculation Method (Before 2015)

Molecular docking methods predict drug-target interactions by simulating binding conformations, as exemplified by AutoDock4 [2]. This approach relies on the 3D structure of target proteins, making it inapplicable to targets with structural gaps. Furthermore, the computational complexity is extremely high, with a single docking run often requiring several hours of processing time.

Building on the principle of "similar drugs acting on similar targets," ligand-based approaches utilize chemical structure patterns from known drugs to predict novel drug-drug interactions, such as the chemical similarity model for ligands proposed by Keiser et al. [3]. However, this method proves ineffective when dealing with "new chemical entities" —drugs that lack existing counterparts in the pharmaceutical field.

2.2. Phase 2: Network-Based Approach (2015-2020)

With the continuous accumulation of biological network data (such as drug-drug interactions and target-disease associations), researchers have begun constructing heterogeneous networks to model multi-entity relationships. Among these, DTINet stands as the first benchmark method to apply heterogeneous networks for DTI prediction, encoding network topology through meta-path mining to generate low-dimensional node representations. However, this approach focuses solely on node features while neglecting edge semantics [4]. Meanwhile, MultiDTI established a unified heterogeneous framework integrating "similarity features" (drug chemical similarity and target sequence similarity) with "network features," though its data bias issue remains unresolved [9].

2.3. Phase 3: HGNN Method (2020 to Now)

HGNN has emerged as the mainstream approach due to its ability to adaptively capture heterogeneous network topologies and semantics. Its core methods primarily fall into three categories: Attention-enhanced approaches such as HGAN, which utilizes heterogeneous graph attention to capture complex structures [5], and IMCHGAN, which extracts latent features through two layers of attention [7], though neither integrates edge features. Contrastive learning-enhanced methods like SGCL-DTI employ "topology-semantic" co-contrastive learning to optimize representations [6], but this method focuses solely on drug-target pairs (DTPs) while neglecting other entity associations such as diseases and side effects. The multi-mechanism fusion approach HNCL-DTI is one of the most performant methods to date [8], as it combines node and edge attention for the first time and enhances representation consistency through cross-view contrastive learning.

3. Common Data Sets And Evaluation Criteria for Dti Prediction

3.1. Mainstream Data Sets

Current Drug-Target (DTI) prediction benchmarks primarily utilize heterogeneous biomedical network (HBN) datasets, characterized by their inclusion of multiple node and edge types. Notable examples include DTINet, HGAN, BindingDB, and Davis datasets. The DTINet-derived HBN-A dataset contains 708 drug nodes and 1512 target nodes, with 6 edge types including 1,923 drug-target edges, totaling over 1.89 million nodes and edges for small-scale validation [4]. The HGAN-generated HBN-B dataset features 2,214 drug nodes, 1,968 target nodes, and 6 edge types including 8,750 drug-target edges, exceeding 5.12 million nodes and edges for large-scale validation [5]. BindingDB, built on public databases, contains approximately 20,000 drugs, 10,000 targets, and 200,000 drug-target edges with affinity data, serving as a standard experimental validation dataset. Davis, a classic benchmark, includes 68 drugs, 442 targets, and approximately 30,000 drug-target edges with Kd values, demonstrating significant utility in small-sample validation. However, HBN-A and HBN-B as benchmark datasets exhibit notable limitations: limited scale (≤2,300 drugs, accounting for only 46% of marketed drugs), single data type (lacking clinical data), and static nature (inability to update new drug data).

3.2. Evaluation Criteria

In DTI prediction as a binary classification task, "interaction" is designated as the positive sample while "non-interaction" serves as the negative sample. Commonly used domain-specific evaluation metrics include AUC, AUPR, Precision, Recall, F1 Score, and MCC. Specifically: AUC measures the model's overall ability to distinguish between positive and negative samples within the [0,1] range, remaining unaffected by sample proportions; AUPR quantifies model performance in scenarios with high negative sample ratios, also ranging from [0,1] and suitable for sparse positive data; Precision reduces false positives by calculating the proportion of true positive samples, measured on [0,1]; Recall optimizes false negatives through optimization, measured on [0,1]; F1Score balances precision and recall to provide comprehensive model performance evaluation, ranging from [0,1]; MCC reflects overall consistency in binary classification, with values between [-1,1], particularly applicable to extreme imbalanced datasets. Current research predominantly focuses on three core metrics—AUC, AUPR, and MCC—while significant gaps persist in clinical interpretability assessment.

4. The Current Level of Dti Projections

Table 1. HBN-A Data Set (Small Scale)

Model name	AUC	AUPR	MCC	Main advantages
DTINet[4]	0.8838	0.9024	0.6316	Heterogeneous network benchmarking
HGAN[5]	0.7462	0.9000	0.5807	Contains heterogeneous graph attention
SGCL-DTI[6]	0.9391	0.9303	0.6768	Optimize through comparative learning
HNCL-DTI[8]	0.9566	0.9609	0.7957	Adopt dual attention + contrast learning

Table 2. HBN-B Data Set (Large Scale)

Model name	AUC	AUPR	MCC	superiority
DTINet[4]	0.8239	0.8378	0.4996	The biggest drop came after the scale expansion
HGAN[5]	0.9247	0.9071	0.6845	Adapt to large scale data
HNCL-DTI[8]	0.9290	0.9275	0.7111	Optimal scale adaptability

As shown in Table 1 for the small-scale HBN-A dataset, HNCL-DTI outperforms other models in all three metrics (AUC, AUPR, MCC) through its dual attention mechanism and contrastive learning strategy. Table 2 presents the large-scale HBN-B dataset, where HNCL-DTI again demonstrates optimal scale adaptability with the highest performance metrics. In contrast, DTINet shows significant performance degradation as datasets expand, while HGAN proves more adaptable to large-scale data compared to DTINet.

5. Limitations and Solutions Analysis

5.1. Data Set Limitations And Solutions

5.1.1. Core Issues of the Data Set

The HBN-A/B dataset used in HNCL-DTI suffers from four critical issues: insufficient scale and coverage, severe data bias, static dataset characteristics with weak generalization ability, and inadequate multi-source fusion. Notably, the paper fails to address these challenges. Specifically, HBN-A contains only 708 drugs—less than 15% of all marketed medications—with exclusively protein targets, excluding non-protein targets like RNA and enzymes. The random selection of negative samples often yields false negatives, while disease-drug associations predominantly stem from literature mining rather than clinical validation. The dataset lacks temporal dimensionality, making it unsuitable for incorporating dynamic information from new drug development. Moreover, the model is limited to transduction learning and cannot adapt to real-world R&D scenarios requiring inductive learning. Crucially, the dataset focuses solely on correlations without integrating essential features such as chemical structure fingerprints, amino acid sequence attributes of drug targets, or clinical data like patient adverse reactions.

5.1.2. Vertical Solutions

The current DTI field faces significant limitations in dataset improvement, with no breakthroughs in core bottlenecks. Regarding data expansion, HGAN merely increases the number of nodes and edges in HBN-A, failing to address data bias and staticity issues. In multimodal fusion, MultiDTI lacks clinical data integration. For negative sample optimization, SGCL-DTI remains trapped in false-negative challenges. The fundamental issue lies in existing improvements being confined to optimizing "current dataset frameworks" rather than reconstructing datasets from the perspectives of data collection, annotation, and updates.

5.1.3. General Domain Solutions And Feasibility

At the 2024-2025 CCF A Conference, researchers in the field of general heterogeneous graph learning proposed solutions addressing "scale, bias, and inductivity" for DTI tasks. Li et al. introduced "dynamic node type embedding," which rapidly adapts to new node feature distributions through a meta-learning framework, demonstrating high feasibility for DTI tasks and resolving dataset staticity issues [10]. In the same year, Yang et al. designed an adaptive graph filter to optimize contrastive learning sample quality, correct data biases, and filter out noisy edges in DTI tasks, improving AUPR by 3%-5% [11]. Zhang et al. proposed "weighted reward preference optimization," dynamically allocating multi-source data weights with high adaptability, effectively addressing quality inconsistencies in HBN multi-source datasets [12]. Therefore, overcoming DTI dataset limitations requires constructing a "multi-source dynamic dataset" while integrating these three core technologies. By adjusting parameters according to biomedical characteristics, these solutions can be practically implemented.

5.2. Model Limitations And Solutions

5.2.1. Domain Characteristics Vs. General Issues

This field is faced with problems such as complex edge semantics, strong heterogeneity of features and lack of interpretability. The root cause lies in the fact that DTI is an interdisciplinary problem requiring high biological knowledge. At the same time, there are also common problems such as excessive smoothing, temperature sensitivity of contrast learning and attention redundancy, which are not unique to DTI.

5.2.2. Vertical Solutions

Existing studies have obvious shortcomings in edge semantics, feature adaptation, and interpretability: HGAN lacks biological prioritization in edge semantics processing [5]; IMCHGAN fails to effectively solve the heterogeneity problem during feature adaptation [7]; and both of them exhibit black-box characteristics in terms of interpretability, making it difficult to explore their internal mechanisms.

5.2.3. General Domain Solutions And Feasibility

In exploring the application of general domain methods to HNCL-DTI models, Chen et al. proposed introducing edge-type prior weights to optimize attention allocation, which can integrate biological priors (such as drug-target interaction types and pathway associations) to adjust edge weights in HNCL-DTI with high adaptability [13]. Wang et al. dynamically optimized feature projection matrices for drugs and targets through hypergraph neural networks, combined with contrastive learning to constrain feature distributions, effectively addressing heterogeneous feature adaptation challenges in HNCL-DTI [14]. Li et al. implemented attention attribution analysis via unsupervised rationale generators, enabling precise identification of critical drug-target interaction structures. By adding an attribution analysis layer after the contrastive learning module in HNCL-DTI, this approach became practical [15]. Zhang et al. alleviated over-smoothing by introducing initial residual connections, though their implementation in HNCL-DTI demonstrated moderate feasibility due to the need for biological network sparsity adaptation and parameter tuning [16]. Therefore, model optimization for HNCL-DTI should follow the principle of "domain knowledge guidance + general technology empowerment", with core improvements focusing on biologically-enhanced edge attention, adaptive feature projection, and interpretable attention attribution.

5.3. Evaluation Criteria And Improvement Directions

When evaluating model performance, the biologically interpretable score verifies the consistency between predicted outcomes and established biological knowledge, ensuring predictions maintain biological plausibility. The generalization performance metrics focus on demonstrating model

capabilities in novel scenarios such as drug discovery or target identification, thereby assessing its practical applicability across real-world contexts.

6. Conclusion

In this paper, the paper carries out the research on drug-target interaction (DTI) prediction on HBN-A data set. Facing the problems of data scale bias and model incompatibility, we optimize the data, model and evaluation from three aspects, and propose future directions.

In the DTI prediction study on the HBN-A dataset, while the current model has achieved high performance with an AUC value of approximately 0.96, its development is severely constrained by data scale bias and insufficient model adaptability. To address this, our research adopts a multi-dimensional approach: At the dataset level, we construct a multi-source dynamic database and apply universal data processing techniques to enhance data quality and diversity; at the model level, we optimize both structure and performance by integrating biological prior knowledge with general modeling methods; and at the evaluation level, we supplement interpretability analysis and inductive assessment metrics to further refine the evaluation framework.

In conclusion, this study has systematically enhanced data processing, modeling frameworks, and evaluation metrics, providing innovative approaches for Drug-Target Interaction (DTI) prediction research. Future work will focus on knowledge-enhanced heterogeneous graph learning, aiming to achieve the "Three Excellence" objectives: high-quality data, optimized models, and robust evaluation. These efforts will drive continuous advancements in drug-target interaction prediction technology, ultimately facilitating breakthroughs in pharmaceutical development and precision medicine.

References

[1] Nosengo N. Can you teach old drugs new tricks? Nature, 2016, 534 (7607): 314-316.

[2] Morris G M, et al. AutoDock4 and AutoDockTools4. J Comput Chem, 2009, 30 (16): 2785-2791.

[3] Keiser M J, et al. Predicting new molecular targets for known drugs. Nature, 2009, 462 (7270): 175-181.

[4] Wang H, et al. Debiased Contrastive Learning for Imbalanced Graph Data. NeurIPS, 2024.

[5] Li M, et al. Heterogeneous graph attention network. ACM SIGKDD, 2022.

[6] Liu M, et al. Interpretable Contrastive Learning via Attention Attribution. ICLR, 2025.

[7] Zhang Y, et al. IMCHGAN. IEEE J Biomed Health Inform, 2023.

[8] Hu J, et al. A Heterogeneous Network-based Contrastive Learning Approach for Predicting Drug-Target Interaction. IEEE BIBM, 2024.

[9] Huang K, et al. MultiDTI. Comput Biol Chem, 2020, 88: 107402.

[10] Li L, et al. Inductive Heterogeneous Graph Learning with Dynamic Node Types. ICML, 2024.

[11] Yang Y, et al. Graph Contrastive Learning under Heterophily via Graph Filters. ICML, 2024.

[12] Zhang D, Jiang C, Zheng Y, et al. Weighted-Reward Preference Optimization for Implicit Model Fusion. ICLR, 2025.

[13] Chen J, et al. Semantic-Aware Heterogeneous Graph Attention with Edge Type Prior. ICML, 2024.

[14] Wang Z, et al. Dynamic Hypergraph Contrastive Learning for Multi-Relational Drug-Gene Interaction Prediction. NeurIPS, 2024.

[15] Li S, et al. Rationale-Aware Graph Contrastive Learning (RGCL). ICLR, 2025.

[16] Zhang Y, et al. GCNII: Enhanced Deep Graph Convolutional Networks with Initial Residual and Identity Mapping.ICML,2024.