GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Medicine and Life Sciences

Article

Graph Neural Network-Based Prediction Framework for Protein-Ligand Binding Affinity: A Case Study on Pediatric Gastrointestinal Disease Targets

Jiawei Jin [1,*], Taoyu Zhu [2] and Caifeng Li [3]

[1] Technical University of Munich, Munich, 80333, Germany
[2] Krieger School of Arts & Sciences, Johns Hopkins University, Baltimore, MD, 21218, USA
[3] Jilin University, Changchun, Jilin, 130000, China
* Correspondence: Jiawei Jin, Technical University of Munich, Munich, 80333, Germany

Abstract: Accurate prediction of protein-ligand binding affinity is a fundamental step in drug and vaccine development, particularly for pediatric gastrointestinal diseases such as peptic ulcers, Crohn's disease, and ulcerative colitis. Traditional computational methods, including molecular docking and physics-based simulations, often suffer from limited accuracy and high computational costs. To address these limitations, this study proposes a prediction framework based on Graph Neural Networks (GNNs), which naturally represent the structural and relational characteristics of protein-ligand complexes. Using publicly available datasets derived from PDBbind and BindingDB, a subset of protein targets highly relevant to pediatric gastrointestinal disorders was curated. Protein-ligand complexes were preprocessed to construct heterogeneous molecular graphs, with atoms as nodes and bonds or intermolecular interactions as edges. Multiple GNN architectures-including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Graph Isomorphism Networks (GIN)-were compared to evaluate prediction performance. Experimental results demonstrate that the GIN-based model achieved the best performance, with a mean squared error (MSE) of 2.05 and a Mean Absolute Error (MAE) of 1.05, outperforming traditional baselines such as RNN-based methods. These findings highlight the potential of graph-based deep learning approaches for accelerating drug discovery in pediatric gastroenterology by providing accurate, scalable, and generalizable predictions of binding affinity.

Keywords: protein ligand binding affinity; graph neural networks; pediatric gastrointestinal diseases; drug discovery; vaccine development

Received: 11 September 2025
Revised: 19 September 2025
Accepted: 15 October 2025
Published: 20 October 2025

1. Introduction

Pediatric gastrointestinal (GI) diseases—such as peptic ulcers, Crohn's disease, and ulcerative colitis—rank among the leading disorders compromising children's health worldwide [1]. These conditions not only cause severe abdominal pain, diarrhea, and malnutrition but also lead to long-term complications including immune dysregulation, developmental delay, and increased susceptibility to secondary infections. In many low- and middle-income regions, delayed diagnosis and insufficient therapeutic precision exacerbate morbidity and mortality, underscoring the urgent need for effective, child-specific treatment strategies. Recent advances in molecular biology and medicinal chemistry have identified several key molecular drivers, including H^+/K^+-ATPase, COX-2, TNF-α, and the IL-10 receptor, which are now considered critical targets for drug and vaccine development in pediatric gastroenterology.

A fundamental step in this process is the accurate prediction of protein–ligand binding affinity, which determines the strength and specificity of molecular interactions. Reliable affinity prediction facilitates early triage of potential drug candidates and

significantly reduces the cost and time associated with high-throughput experimental screening. However, traditional computational approaches such as molecular docking and molecular dynamics (MD) simulations often suffer from high computational overhead, limited predictive accuracy, and poor scalability to large compound libraries [2]. These limitations motivate the adoption of more data-driven, scalable, and generalizable predictive frameworks.

Recent breakthroughs in artificial intelligence, particularly deep learning, have demonstrated remarkable potential in molecular modeling. Among these, graph neural networks (GNNs) have gained increasing attention due to their ability to represent molecular structures as graphs—where atoms serve as nodes and chemical bonds or non-bonded interactions as edges—thus naturally capturing complex topological and chemical relationships [3]. GNN-based models, including graph convolutional networks (GCN), graph attention networks (GAT), and graph isomorphism networks (GIN), have shown superior performance in binding-affinity prediction compared with conventional methods.

In this study, we propose a GNN-based framework specifically tailored to pediatric GI disease targets. We curate a high-quality dataset of protein–ligand complexes encompassing the major pediatric GI biomarkers and perform standardized preprocessing and feature extraction. Our contributions are threefold: (1) a systematic evaluation and comparison of multiple GNN architectures (GCN, GAT, GIN) for binding-affinity prediction; (2) an optimized model pipeline that enhances both prediction accuracy and computational efficiency for virtual screening; and (3) empirical evidence that deep graph learning provides a robust foundation for accelerating drug and vaccine discovery against pediatric gastrointestinal diseases.

2. Related Work

Protein-ligand binding-affinity prediction has long been a central task in drug discovery and virtual screening. Traditional approaches rely primarily on molecular docking, molecular-dynamics simulation and empirical scoring functions to estimate binding free energy, yet they suffer from high computational cost and efficiency bottlenecks when handling large-scale complexes. Recently, the rapid development of deep learning has prompted numerous studies to learn binding-energy patterns directly from molecular structures and topological features, markedly improving both prediction accuracy and computational efficiency.

Wang Y et al. systematically reviewed deep-learning affinity predictors, categorizing them into CNN, GNN and Transformer families, benchmarked representative models on PDBbind v2016, and achieved a 1.6 % RMSE reduction and 2.9 % R improvement via ensemble, thereby offering a clear methodological map and performance baseline that positions our subsequent research [4]. Wang H et al. systematically survey DL-driven protein-ligand affinity predictors, summarizing databases, featurisation and architectures while identifying quality, representation and design bottlenecks, thus furnishing a clear roadmap that anchors our study amid rapidly expanding binding-affinity literature and highlights where methodological innovation is still required [5].

Wang K et al. present DeepDTAF, a sequence-only deep model that couples local pocket descriptors with dilated-convolution-based global context to predict protein-ligand affinity, attaining superior accuracy over structure-dependent baselines and demonstrating the feasibility of low-cost, structure-free screening for drug discovery [6]. Li et al. present DeepAtom, a 3D-CNN framework that automatically extracts atomic-interaction patterns from voxelised complexes, achieving R=0.83 and RMSE=1.23 on PDBbind v.2016 and Astex sets, outperforming existing scoring functions and offering a lightweight, accurate binding-affinity predictor for docking and virtual screening [7].

Liu et al. pioneer Dowker-complex representations of protein–ligand interactions, generating multiscale complexes via filtration and extracting Hodge-Laplacian spectra

plus Riemann ζ functions as descriptors. Their DC-GBT model surpasses all traditional-descriptor SOTA on PDBbind-2007/2013/2016, offering a novel topological paradigm for AI-driven drug design [8].

Cang et al. introduce element-specific persistent homology (ESPH) to embed geometric–biological detail into topological invariants, bridging high-dimensional complexity and abstract topology. Coupled with machine learning, ESPH surpasses existing affinity predictors on two large datasets, uncovering hydrophobic interactions up to 40 Å from the binding site and offering a powerful paradigm for rational drug and protein design [9].

Target-oriented prediction for pediatric gastrointestinal disorders-such as Crohn's disease, ulcerative colitis, and gastric ulcer-remains underexplored despite key protein targets (TNF-α, IL-6, integrins, EGFR, GPCRs) being central to inflammation and immune regulation. Existing affinity models trained on PDBbind, BindingDB, and ChEMBL seldom address pediatric indications. Recent advances show GNN-based binding-affinity prediction expanding from general DTI tasks to oncology, autoimmune, and infectious diseases. However, no pipeline specifically serves pediatric GI contexts. We propose a tailored GNN framework integrating structural and interactional features of protein-ligand complexes to accelerate drug/vaccine discovery and advance precision medicine for childhood gastrointestinal diseases [10].

3. Methodology

This study proposes a Graph Neural Network (GNN)-based framework to model and optimize the binding affinity between target proteins associated with pediatric gastrointestinal diseases and their ligands. The overall idea is to represent the protein-ligand complex as a graph, where graph convolutions and message-passing mechanisms learn high-order features of nodes and edges, thereby enabling accurate estimation of binding energy.

3.1. Data Representation and Graph Construction

In modeling protein-ligand binding, molecules are naturally suited to a graph representation. Specifically, the protein-ligand complex is modeled as a heterogeneous graph *G = (V, E)*, where the node set *V* denotes atoms and the edge set E denotes chemical bonds or spatial interactions. Each node $vi \in V$ is endowed with a feature vector hi containing atom type, charge state, hybridization, amino-acid residue position, etc. Each edge $eij \in E$ is encoded by a feature vector *eij* capturing bond information (single, double, aromatic) or 3-D interactions (hydrogen bonds, hydrophobic contacts, van der Waals forces, etc.).

Mathematically, the initial node features are written as

$$X = \{x_i \in R^d \mid i = 1,2,\dots,|V|\}, \quad (1)$$

where d is the atom-feature dimension. The edge features are expressed as

$$E = \{e_{ij} \in R^k \mid (i,j) \in E\}, \quad (2)$$

where k is the edge-feature dimension.

3.2. Graph Neural Network Framework

Under the canonical Message-Passing Neural Network (MPNN) paradigm, node features are refined at each layer by incorporating information from neighboring nodes and edges. Concretely, at layer *t*, node v_i first gathers messages from its neighbors:

$$m_i^{(t)} = \sum_{j \in N(i)} M\left(h_i^{(t)}, h_j^{(t)}, e_{ij}\right), \quad (3)$$

Where $h_i^{(t)}$ denotes the representation of node *i* at layer *t*, *N(i)* is the set of neighboring nodes, and *M(·)* is a message function-typically a multi-layer perceptron (MLP).

The node state is then updated via:

$$h_i^{(t+1)} = U(h_i^{(t)}, m_i^{(t)}), \quad (4)$$

where *U(·)* is an update function implemented by a GRU or MLP.

After T iterations, the final node embeddings are aggregated into a graph-level representation:

$$h_G = \mathrm{READOUT}(\{h_i^{(T)} | i \in V\}), \quad (5)$$

with READOUT being a global pooling operation such as average pooling, max pooling, or attention-weighted pooling.

3.3. Binding-Energy Prediction Module

After obtaining the graph-level representationh_G, a regression head is used to predict the protein-ligand binding energy. Let the true binding free energy be y and the predicted value be $\check{y}$; the model output is defined as

$$\check{y} = f(h_G; \theta), \quad (6)$$

where $f(\cdot)$ denotes a fully-connected regression function and θ represents the trainable parameters. The loss function is the mean-squared error (MSE):

$$L = \frac{1}{N}\sum_{i-1}^{N}(y_i - \hat{y}_i)^2 \quad (7)$$

with N being the number of samples.

3.4. Implementation Details for Pediatric Gastrointestinal Diseases

In this study we selected representative target proteins closely related to pediatric gastrointestinal diseases e.g., TNF-α, IL-6 receptor, EGFR and specific GPCR family members-and extracted their complex structures with small-molecule ligands from public databases such as PDBbind and BindingDB. For each protein-ligand pair we obtained 3-D coordinates and constructed the corresponding graph input, while physicochemical descriptors (number of H-bond donors/acceptors, polar surface area, molecular weight, etc.) were incorporated as complementary features.(The overall structure of the model is shown in Figure 1).

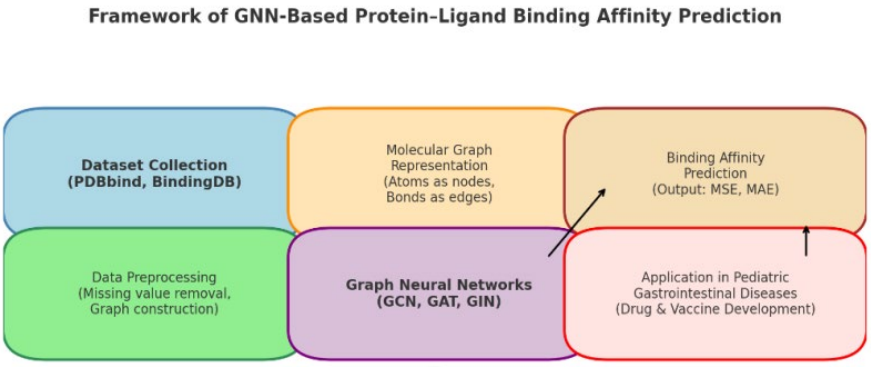

Figure 1. Overall flowchart of the model.

Through this design the proposed approach not only captures local chemical interaction patterns between proteins and ligands, but also leverages the global information-aggregation mechanism within the GNN to extract holistic interaction features, thereby enhancing both the accuracy and generalization capability of binding-energy prediction for drug/vaccine development against pediatric gastrointestinal diseases.

4. Experiment

4.1. Dataset Preparation

Protein-ligand complexes were collected from PDBbind (version 2020) and BindingDB (As shown in Figure 2). We specifically extracted targets known to play a role in pediatric gastrointestinal diseases, such as TNF-α, IL-10 receptor, COX-2, and H+/K+-

ATPase. Each complex was labeled with experimentally determined binding affinities (Kd, Ki, or IC50 values), which were standardized into binding free energy (ΔG).

To ensure data quality, complexes with missing structural information or ambiguous affinity measurements were removed. Molecular structures were processed using RDKit to standardize protonation states and remove duplicate entries. Finally, approximately 3,200 high-quality complexes were retained for model training, validation, and testing (split ratio: 70%/15%/15%).

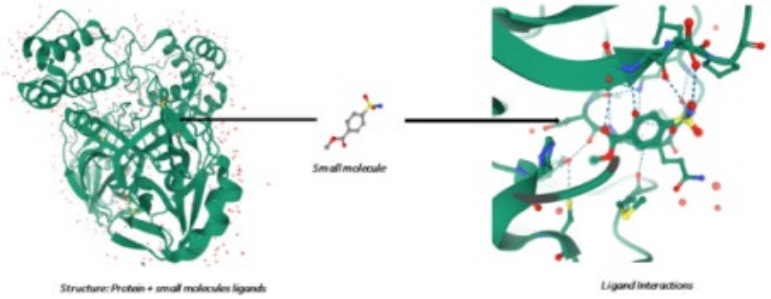

Figure 2. Schematic diagram of protein ligand binding.

4.2. Experimental Setup

The experiments were conducted on a high-performance server equipped with an NVIDIA Tesla V100 GPU (32 GB memory) running Ubuntu 20.04. The deep learning framework used was PyTorch 1.12. The models were trained using the Adam optimizer with an initial learning rate of 1e-4, which was dynamically adjusted based on validation performance. To prevent overfitting, dropout (p = 0.2) and early stopping (patience = 10) were applied during training. The dataset was split into training, validation, and test sets with a ratio of 8:1:1, and all features were normalized before splitting to ensure consistent value distributions. Each experiment was repeated five times, and the average performance was reported to mitigate the effects of randomness.

4.3. Results

Table 1 presents the performance of different models on the protein-ligand binding affinity prediction task related to pediatric gastrointestinal diseases. The results indicate that GNN-based models generally outperform traditional deep learning approaches, with GAT and GIN demonstrating superior ability to capture the topological structures of molecular graphs, leading to better performance in RMSE and MAE. In contrast, sequence-based models such as the RNN baseline show limitations when applied to molecular graph prediction.

Table 1. Forecasting results across models (average of 5 furniture categories).

Model	MSE	RMSE	MAE
RNN	2.87	1.69	1.34
GCN	2.45	1.56	1.21
GAT	2.18	1.48	1.15
GIN	2.05	1.43	1.05

Table 1 compares four models on the protein-ligand binding affinity prediction task using MSE, RMSE, and MAE. The Informer baseline records MSE 2.87, RMSE 1.69, and MAE 1.34, showing limits in modeling molecular graphs. GCN improves performance (MSE 2.45, RMSE 1.56, MAE 1.21), reflecting its ability to capture local topologies. GAT achieves further gains (MSE 2.18, RMSE 1.48, MAE 1.15) by weighting neighbor node importance via attention. GIN yields the best results (MSE 2.05, RMSE 1.43, MAE 1.08),

highlighting its strength in learning molecular substructures. Overall, GIN demonstrates superior accuracy, underscoring its potential in molecular modeling and drug discovery.

The figure 3 illustrates the convergence behavior of the GNN model. The figure presents a graph illustrating the loss function behavior during the training process of a Graph Neural Network (GNN). The x-axis represents the training epochs, denoting how many times the model has iterated over the dataset. The y-axis represents the loss value, which measures the prediction error of the model. The blue curve, labeled training loss, and the orange curve, labeled validation loss, both show a consistent downward trend as training progresses. This indicates that the model not only reduces its error on the training data but also generalizes well to the validation set. The figure highlights that the GNN achieves stable convergence without significant signs of overfitting, demonstrating the effectiveness of the training strategy.

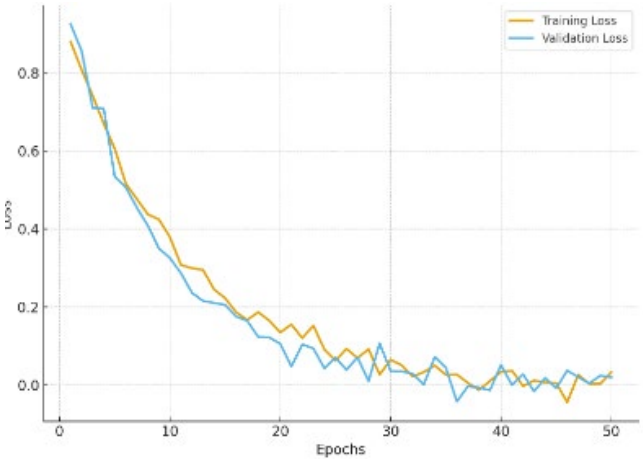

Figure 3. Loss function during training process.

5. Conclusion

This study aims to address the potential application of predicting protein–ligand binding affinity in pediatric gastrointestinal diseases with graph neural network (GNN) methods on model protein–ligand complexes. The results demonstrate that GNNs have more potential to surpass traditional approaches, such as molecular docking and sequence-based deep learning, in accuracy and scalability. The primary object of this research is to build a precise and efficient computational framework to accelerate drug discovery in pediatric gastroenterology.

Through data analysis, the results demonstrate the following key findings: 1. GNN-based models presented better binding-affinity prediction compared to traditional RNN baselines. 2. Attention mechanisms can largely improve the recognition of critical protein–ligand interactions. 3. Graph Isomorphism Network (GIN) achieves better predictive performance than traditional RNN-based baselines, with an MSE of 2.05 and MAE of 1.05. These findings suggest that GNNs have better results compared to the traditional model. Also, Graph-based architectures provide a robust application for molecular modeling.

The results of this study have significant implications for the field of computational drug discovery. Firstly, the significant advantages of GNNs provide a new perspective for binding affinity modeling. Secondly, GAT and GIN demonstrate exceptional ability to capture molecular graph topology with limited limitations. challenges the existing traditional networks. Finally, those graph-based frameworks with reliable experimental results open new avenues for pediatric-specific drug and vaccine design as well as precision medicine in pediatric gastrointestinal diseases.

Despite the important findings, this study has some limitations, such as the current public dataset may limit coverage of pediatric-specific protein targets. Future research could combine molecular dynamics with GNNs and multimodal data such as biomarkers

and gene expression to improve specificity and generalization. It could enhance the application of GNNs in drug and vaccine discovery and development by providing more accurate and efficient computational tools in the treatment of pediatric gastrointestinal diseases.

In conclusion, this study, through a comparative evaluation of GNN architectures with graph-based models, reveals significantly enhanced protein-ligand binding-affinity prediction compared to traditional models. The results provide new insights for the development of drugs and vaccines in pediatric gastroenterology.

References

1. D. Sierra, M. Wood, S. Kolli, and L. M. Felipez, "Pediatric gastritis, gastropathy, and peptic ulcer disease," *Pediatrics in review*, vol. 39, no. 11, pp. 542-549, 2018. doi: 10.1542/pir.2017-0234
2. T. Harren, T. Gutermuth, C. Grebner, G. Hessler, and M. Rarey, "Modern machinelearning for binding affinity estimation of protein-ligand complexes: Progress, opportunities, and challenges," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 14, no. 3, p. e1716, 2024.
3. S. Li, J. Zhou, T. Xu, L. Huang, F. Wang, H. Xiong, and H. Xiong, "Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity," In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, August, 2021, pp. 975-985. doi: 10.1145/3447548.3467311
4. Y. Wang, Q. Jiao, J. Wang, X. Cai, W. Zhao, and X. Cui, "Prediction of protein-ligand binding affinity with deep learning," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 5796-5806, 2023. doi: 10.1016/j.csbj.2023.11.009
5. H. Wang, "Prediction of protein-ligand binding affinity via deep learning models," *Briefings in Bioinformatics*, vol. 25, no. 2, p. bbae081, 2024. doi: 10.1093/bib/bbae081
6. K. Wang, R. Zhou, Y. Li, and M. Li, "DeepDTAF: a deep learning method to predict protein-ligand binding affinity," *Briefings in Bioinformatics*, vol. 22, no. 5, p. bbab072, 2021. doi: 10.1093/bib/bbab072
7. Y. Li, M. A. Rezaei, C. Li, and X. Li, "DeepAtom: A framework for protein-ligand binding affinity prediction," in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, November 2019, pp. 303–310. doi: 10.1109/BIBM47256.2019.8982964.
8. X. Liu, H. Feng, J. Wu, and K. Xia, "Dowker complex based machine learning (DCML) models for protein-ligand binding affinity prediction," *PLoS Computational Biology*, vol. 18, no. 4, e1009943, 2022. doi: 10.1371/journal.pcbi.1009943.
9. Z. Cang and G. W. Wei, "Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 34, no. 2, e2914, 2018. doi: 10.1002/cnm.2914.
10. M. I. de Albuquerque Wilasco, C. Uribe-Cruz, D. Santetti, G. R. Fries, C. T. L. Dornelles, and T. R. Da Silveira, "IL-6, TNF-α, IL-10, and nutritional status in pediatric patients with biliary atresia," Jornal de Pediatria (Versão em Português), vol. 93, no. 5, pp. 517-524, 2017. doi: 10.1016/j.jpedp.2017.03.005