International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

An overview of graph neural networks for molecular biology: Challenges and solutions

Divyansh Choubisa *

Department of Computer Science and Engineering, Wilfrid Laurier University, Ontario, Canada.

International Journal of Science and Research Archive, 2024, 13(01), 2670–2673

Publication history: Received on 08 September 2024; revised on 04 October 2024; accepted on 17 October 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.1.1985

Abstract

This review explores the application of graph neural networks (GNNs) in molecular biology, with a particular focus on their role in drug discovery. Various use cases, including molecule classification for cardiotoxicity detection and the prediction of molecular properties have been examined. A comprehensive analysis compares methodologies, problem statements, datasets, and findings across different studies. A core themes of review is the intuition to regularize a graph neural network which was proven to be reliable in terms of robustness for drug discovery purposes using noise nodes techniques. In the conclusion section, key insights from the reviewed literature and promising future directions for research have been presented. This work aims to provide a foundational understanding and guide future innovations in leveraging GNNs for molecular applications.

Keywords: Graph Neural Networks; Drug Discovery; AI in Medicine; Healthcare AI; Computational Biology;

1. Introduction

The integration of artificial intelligence (AI) in drug discovery is transforming the pharmaceutical industry, making the process faster and more efficient. Traditional methods often require extensive time and resources, but AI technologies are streamlining the identification of new therapeutics. Among these technologies, Graph Neural Networks (GNNs) have emerged as a powerful approach, enabling the representation of molecules and biological interactions as graphs. This capability allows GNNs to capture complex relationships within molecular structures, facilitating the prediction of properties, drug-target interactions, and potential side effects.

Computational biology complements this effort by providing the necessary biological data and insights. By leveraging high-throughput screening, genomic information, and protein structures, computational biology enhances the predictive accuracy of GNNs and other AI methods.

Together, these disciplines are revolutionizing drug discovery, promising to reduce the time and cost of developing new treatments. As research advances, the synergy between AI, GNNs, and computational biology is poised to unlock novel therapeutic strategies for complex diseases, heralding a new era in medicine.

2. Graph Neural Networks

From last many years a field of deep neural networks has emerged massively and gained a lot of popularity and has a lot of powerful applications. It is the “spider web”, of AI, the neural networks who have powerful applications in our real world, these networks are called Graph Neural Networks. It’s not only about our real world or the microscopic world of molecules but the universe itself arranged in a systematic form. Everything is connected with each other forming a

* Corresponding author: Divyansh Choubisa

phenomenal web of existence or we simply say a graph. Maps or social networks are typical examples of graphs. Now when we look at the very existence of nature, i.e., molecules we find that they are also arranged in a form of graph i.e. atoms are connected by chemical bonds forming molecules. Mathematically, a graph is nothing but a system of vertices and edges connected together which can be represented by G = (V, E). [3] mentions that when it comes to biology, graph neural networks can be used for many different tasks that can broadly classify GNN tasks into three categories. 1 – Node Classification: Classifying protein functions in a given protein-protein interaction network. 2 – Graph Classification: Classifying molecules for their quantum-chemistry properties or virtual drug screening. 3 – Link Prediction: Predict links between drug and diseases or drug and targets.

However, graph neural networks also face certain challenges which are regularly addressed by the researchers time to time. In this article I ll talk about few specific challenges GNN encounter and how those challenges were solved, I ll also talk about it in context of molecular biology.

2.1. Trust and reliability of GNN:

GNN provide a great performance in drug discovery process. However, there are quite some issues with graph neural networks when it comes to trust and reliability. One of them is overconfident false negative. Consider a scenario where a graph neural network predicts a toxic molecule as safe. This is minor but a big mistake which can produce harmful results when designing drugs. To solve the problem of reliability of GNN under distributional shifts [2] proposes some techniques. [2] proposes CardioTox, a benchmark dataset based on a real-world drug discovery problem and is compiled from 9K+ drug-like molecules from ChEMBL and NCATS databases to evaluate GNN model reliability. [2] introduces GNN-GP and GNN-SNGP techniques to increase reliability of GNN models and reduce overconfident mispredictions. It was found in the [2] that distance awareness can improve overconfident mispredictions. Hence it was observed in the [2] that GNN-GP and GNN-SNGP together with a new benchmark CardioTox dataset reduce overconfident mispredictions without sacrificing accuracy as displayed in the table below. Below table also compare and describes GNN baseline, GNN-GP and GNN-SNGP models.

Table 1 Comparison between different GNN architectures

GNN Varients	Description
GNN Baseline	Used a vanilla Message Passing Neural Network as the baseline.
GNN GP	Improved distance awareness by introducing a *Gaussian Process* layer.
GNN SNGP	Added *Spectral Normalization* along with a *Gaussian Process* layer, outperforming the base architecture not only in accuracy but also in robustness, mainly in reducing overconfident mispredictions.

Table 2 Accuracy (AUROC), robustness (ECE, Brier, NLL) and overconfidence (OFNs) performance for drug cardiotoxicity benchmark. (As mentioned in the corresponding paper)

Test-IID	AUROC (↑)	ECE (↓)	Brier (↓)	NLL (↓)	OFNs% (↓)	DA-AUC (↑)
GNN baseline	0.919±0.003	0.037±0.003	0.194±0.007	0.352±0.018	4.05±0.26	0.500±0.004
GNN-GP	**0.937±0.001**	0.036±0.002	**0.176±0.008**	**0.298±0.015**	**3.51±0.17**	**0.523±0.004**
GNN-SNGP	0.932±0.001	**0.028±0.001**	**0.179±0.005**	**0.295±0.005**	**3.56±0.21**	0.517±0.004
GNN-SNGP Ensemble	**0.942±0.000**	**0.013±0.001**	**0.162±0.000**	**0.264±0.001**	**2.54±0.07**	**0.525±0.002**
Test-OOD1	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**DA-AUC (↑)**
GNN baseline	0.786±0.004	0.102±0.012	0.343±0.071	0.578±0.125	1.68±0.08	0.604±0.006
GNN-GP	0.823±0.003	0.090±0.010	0.327±0.061	0.546±0.107	1.41±0.12	**0.632±0.002**
GNN-SNGP	**0.836±0.003**	**0.074±0.008**	**0.316±0.047**	**0.503±0.072**	**1.31±0.09**	**0.635±0.007**
GNN-SNGP Ensemble	**0.850±0.002**	**0.039±0.002**	**0.277±0.005**	**0.428±0.006**	**1.22±0.08**	**0.643±0.003**
Test-OOD2	**AUROC (↑)**	**ECE (↓)**	**Brier (↓)**	**NLL (↓)**	**OFNs% (↓)**	**DA-AUC (↑)**
GNN baseline	0.831±0.007	0.082±0.012	0.284±0.060	0.492±0.113	1.73±0.20	0.630±0.008
GNN-GP	0.873±0.006	0.074±0.007	0.261±0.047	0.442±0.086	**0.98±0.19**	0.656±0.011
GNN-SNGP	**0.885±0.007**	**0.044±0.006**	**0.238±0.040**	**0.389±0.068**	**1.02±0.11**	**0.678±0.008**
GNN-SNGP Ensemble	**0.896±0.002**	**0.021±0.002**	**0.210±0.003**	**0.333±0.005**	**1.06±0.11**	**0.682±0.003**

2.2. Over-smoothing problem in GNN

It is generally found that performance of GNN does not improve as the number of the layers increases. This effect is called over smoothing. To solve this problem of over smoothing [1] proposes a simple noise regularizer in form of noisy nodes. A straightforward method for regularizing a GNN. Noisy nodes change the original graph prediction problem formulation in several ways to address the issue of oversmoothing by incorporating a diversified noise correction target.

The noise regularization technique introduces a per-node noise correction term and taints the input graph's properties with noise during the training of the graph neural network. By using message passing to the final output, the model is rewarded for upholding and improving unique node representations, which prevents it from becoming oversmoothed. [1] also share the observation that a GNS-based architecture that can be used to predict molecular properties performs well when incorporated with noisy nodes and is comparable to the usage of dedicated architectures. The below table shows the impact of noisy nodes on GNS architecture with respect to QM9 benchmark dataset. The QM9 benchmark dataset contains 134k molecules in equilibrium with up to 9 heavy C, O, N and F atoms, targeting 12 associated chemical properties. Experiment in [1] used 114k molecules for training, 10k for validation and 10k for test. [1] added IID Gaussian noise with mean=0 and $\boldsymbol{\sigma}$ =0.02 to the input atom position. As we can see in the table that how a how simply adding noisy nodes in GNS architectures with different number of layers improves the results. As the experiment in [1] are quite descriptive and broad, to conclude briefly, noisy nodes can help graph neural networks by resolving oversmoothing issues; this improves the performance of deep GNNs in particular. In the domain of molecular biology, it can help with tasks like 3D molecular property prediction and more.

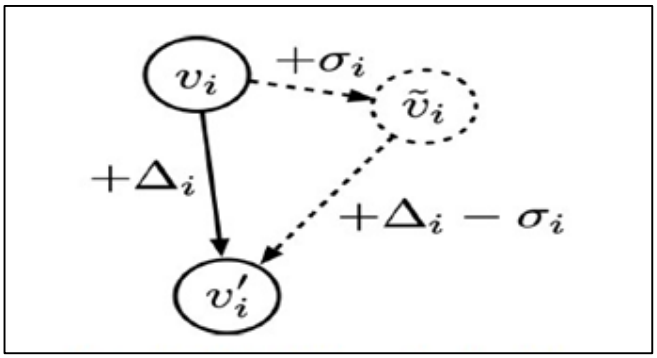

Figure 1 Noisy Node mechanics during training. Input position are corrupted with noise $\boldsymbol{\sigma}$, and the training objective is the node- level difference between target positions and the noisy inputs

Table 3 QM9 Dataset. Impact of Noisy Nodes on GNS architecture

	Layers	std. MAE	% Change	logMAE
GNS	10	1.17	-	-5.39
GNS + Noise But No Node Target	10	1.16	-0.9%	-5.32
GNS + Noisy Nodes	10	0.90	-23.1%	-5.58
GNS-10 + Noisy Nodes	20	0.89	-23.9%	-5.59
GNS-10 + Noisy Nodes + Invariance	30	0.92	-21.4%	-5.57
GNS-10 + Noisy Nodes	30	**0.88**	**-24.8%**	**-5.60**

3. Conclusion

3.1. Comparison Between the Papers, Datasets, Problem Statement, Methods and Promising Future Directions

In this article I presented a brief overview of graph neural networks, what are some challenges that GNN encounter and what are the different techniques proposed to solve those problems and how GNN can be useful for tasks in molecular biology. In the very beginning we get to know about graph neural networks and some of its usage in molecular biology. We also saw that we can improve the trust, reliability and robustness of GNNs by introducing spectral normalization and gaussian process layer in a GNN on a benchmark dataset CardioTox. At the same time in we saw that by adding

noisy nodes in graph neural networks we can see improvement in performance of GNNs in tasks such as 3D molecular property prediction. Although these all experiments are themselves complex and big in nature but they provide a good glimpse on how to build more powerful graph neural networks that are more robust, reliable and powerful to solve very complex tasks on large datasets of molecular biology such as 3d molecular property prediction or molecule classification task to detect cardiotoxicity caused by binding hERG target, a protein associated with heart beat rhythm. We also came across how these different approaches formed different benchmark datasets for the experimentation and evaluation. We found a new benchmark dataset CardioTox to evaluate the trust and reliability of GNN under distributional shifts, however when it comes to further scope of development in, GP and SNGP techniques can be incorporated into GAT and PNA architectures to same benchmark dataset to further study its effectiveness. We also learned how the performance of GNN (with different number of depths) is impacted by addition of noisy nodes on QM9 dataset. However, there are certain limitations when it comes to noisy nodes, like careful selection of kind of noise we are adding, amount of noise. To do that we further require to understand the type of problem we are working on, dataset, distribution and some hyperparameter tuning to gauge the correct amount and selection of noise. This review shows a brief overview of what has been done to solve few challenges encountered by GNN with respect to molecular biology, to understand the experiments in depth, please refer the corresponding papers to capture full understanding.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Jonathan Godwin, Michael Schaarschmidt, Alexander Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veli˘ckovi´c, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction beyond. 2022.

[2] Kehang Han, Balaji Lakshminarayanan, and Jeremiah Liu. Reliable graph neural networks for drug discovery under distributional shift. 2021.

[3] Petar Veli˘ckovi´c. Everything is connected: Graph neural networks.