Graph Neural Networks in Life Sciences: Opportunities and Solutions

Zichen Wang
Amazon
zichewan@amazon.com

Vassilis N. Ioannidis
Amazon
ivasilei@amazon.com

Huzefa Rangwala
Amazon
rhuzefa@amazon.com

Tatsuya Arai
Amazon
araitats@amazon.com

Ryan Brand
Amazon
brandry@amazon.com

Mufei Li
Amazon
limufe@amazon.com

Yohei Nakayama
Amazon
nak@amazon.com

ABSTRACT

Graphs (or networks) are ubiquitous representation in life sciences and medicine, from molecular interactions maps, signaling transduction pathways, to graphs of scientific knowledge and patient-disease-intervention relationships derived from population studies and/or real-world data, such as electronic health records and insurance claims. Recent advance in graph machine learning (ML) approaches such as graph neural networks (GNNs) has transformed a diverse set of problems relying on biomedical networks that traditionally depend on descriptive topological data analyses. Small- and macro- molecules that were not modeled as graphs also saw a bloom in GNN-based algorithms improving the state-of-the-art performance for learning their properties. Comparing to graph ML applications from other domains, life sciences offer many unique problems and nuances ranging from graph construction to graph-level, and bi-graph-level supervision tasks.

The objective of this tutorial is twofold. First, it will provide a comprehensive overview of the types of biomedical graphs/networks, the underlying biological and medical problems, and the applications of graph ML algorithms for solving those problems. Second, it will showcase four concrete GNN solutions in life sciences with hands-on experience for the attendees. These hands-on sessions will cover: 1) training and fine-tuning GNN models for small-molecule property prediction on atomic graphs, 2) macro-molecule property and function prediction on residue graphs, 3) bi-graph based binding affinity prediction for protein-ligand pairs, and 4) organizing and generating new knowledge for drug discovery and repurposing with knowledge graphs. This tutorial will also instruct the attendees to develop in two extensions of the software library Deep Graph Library (DGL), including DGL-lifesci and DGL-KE, so that they could jumpstart their own graph ML journey to advance life science research and development.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542628

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Knowledge representation and reasoning**; • **Applied computing** → **Life and medical sciences**.

KEYWORDS

GNN, Drug Discovery, Knowledge Graph

ACM Reference Format:
Zichen Wang, Vassilis N. Ioannidis, Huzefa Rangwala, Tatsuya Arai, Ryan Brand, Mufei Li, and Yohei Nakayama. 2022. Graph Neural Networks in Life Sciences: Opportunities and Solutions. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542628

1 TUTORIAL OUTLINE

The tutorial introduces to data science researchers and practitioners GNN based approaches applied to various problems in biomedical sciences and healthcare. The tutorial first provides an overview of the various opportunities in leveraging GNNs for small molecules, macromolecules and biomedical knowledge graphs. The four hands-on activities will provide the participants a diverse set of biomedical problems and in particular how to deploy a GNN-based library for these applications leading to biological phenotype prediction, interaction prediction, affinity prediction and drug discovery. The tutorial will be broken up into the following five sections:

1: Overview of Graph ML in biomedical science. This section describes different types of graphs commonly used in biomedical sciences and how graph-based machine learning approaches like GNNs can be leveraged. In particular, we will cover single-entity biomedical networks including gene regulatory network and protein-protein interaction networks, as well as multi-entity networks such as knowledge graphs of proteins, genes, diseases, symptoms, and drugs. This section also introduces graph representations for small and large molecules such as organic compounds and proteins, which can be modeled as independent graphs of atoms and residues, respectively.

2: Making sense of small molecules with GNNs. This section demonstrates how to develop end-to-end graph-based ML pipeline for molecular property prediction. The pipeline first covers how to

construct features from atom graphs for small organic compounds. Then, it will cover two use cases using DGL-lifesci command-line interface: 1) training a GNN for molecular property prediction from scratch, and 2) fine-tuning a pre-trained GNN for molecular property prediction.

3: Making sense of macro-molecules with GNNs. This section demonstrates how to use GNNs to predict properties for macro-molecules including RNAs and proteins. We will cover two hands-on case studies: 1) Prediction of COVID-19 mRNA vaccine degradation with GCN, and 2) protein function prediction using an equivariant GNN on graphs of amino acid residues.

4: Going beyond single graph, bi-graph based binding affinity prediction for protein-ligand pairs. This section demonstrates a case study for making predictions between a pair of graphs. Protein-ligand binding affinity prediction is important for candidate drug screening during the early stage of drug discovery. We demonstrate how PotentialNet can be used for this task, as well as a novel molecular data anonymization procedure for protecting IP of molecular structures.

5: Organizing and generating new knowledge for drug discovery and repurposing with knowledge graphs (KGs). This section showcases another application of graphs in life sciences by employing large-scale KGs to organize the information from diverse medical sources and make prediction on these KGs. KG is a directed heterogeneous multigraph whose node and relation types have domain-specific semantics. We will review three approaches to construct such medical KGs 1) mining medical documents and publications 2) processing and stitching together different KGs coming from various medical databases 3) converting relational databases to KGs. We will show examples detailing how to construct such KGs. The resulting KGs store information efficiently and can be used for KG completion, drug repurposing, and question answering among other tasks. We will review notebooks showcasing how to use the KGs and graph ML to make predictions in these KGs. We also will explain common objectives used for KG completion.

2 SIMILAR/HIGHLY RELATED TUTORIALS

- Deep Graph Learning: Foundations, Advances and Applications[1]
- Scalable Graph Neural Networks with Deep Graph Library [25]
- All You Need to Know to Build a Product Knowledge Graph[2]

REFERENCES

[1] Ziqi Chen, Bo Peng, Vassilis N Ioannidis, Mufei Li, George Karypis, and Xia Ning. 2021. CTKG: A Knowledge Graph for Clinical Trials. *medRxiv* (2021).

[2] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* 10, 2 (2019), 370–377.

[3] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems* 28 (2015).

[4] Evan N Feinberg, Debnil Sur, Zhenqin Wu, Brooke E Husic, Huanghao Mai, Yang Li, Saisai Sun, Jianyi Yang, Bharath Ramsundar, and Vijay S Pande. 2018. PotentialNet for molecular property prediction. *ACS central science* 4, 11 (2018), 1520–1530.

[5] Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. 2021. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics* 22, 6 (2021), bbab159.

[6] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.

[7] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nature communications* 12, 1 (2021), 1–14.

[8] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. 2017. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603* (2017).

[9] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).

[10] Vassilis N Ioannidis, Antonio G Marques, and Georgios B Giannakis. 2019. Graph neural networks for predicting protein functions. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 221–225.

[11] Vassilis N Ioannidis, Da Zheng, and George Karypis. 2020. Few-shot link prediction via graph neural networks for covid-19 drug-repurposing. *arXiv preprint arXiv:2007.10261* (2020).

[12] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. 2020. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411* (2020).

[13] Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. 2021. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS omega* 6, 41 (2021), 27233–27238.

[14] Sanaa Mansoor, Minkyung Baek, Umesh Madan, and Eric Horvitz. 2021. Toward More General Embeddings for Protein Design: Harnessing Joint Representations of Sequence and Structure. *bioRxiv* (2021).

[15] Hannes Stärk, Octavian-Eugen Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. 2022. Equibind: Geometric deep learning for drug binding structure prediction. *arXiv preprint arXiv:2202.05146* (2022).

[16] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 4 (2020), 688–702.

[17] Raphael JL Townshend, Stephan Eismann, Andrew M Watkins, Ramya Rangan, Maria Karelina, Rhiju Das, and Ron O Dror. 2021. Geometric deep learning of RNA structure. *Science* 373, 6558 (2021), 1047–1051.

[18] Zichen Wang, Steven A Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O Salawu, Colby J Wise, Sri Priya Ponnapalli, et al. 2022. Lm-gvp: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific reports* 12, 1 (2022), 1–12.

[19] Hannah K Wayment-Steele, Wipapat Kladwang, Andrew M Watkins, Do Soon Kim, Bojan Tunguz, Walter Reade, Maggie Demkin, Jonathan Romano, Roger Wellington-Oguri, John J Nicol, et al. 2021. Predictive models of RNA degradation through dual crowdsourcing. *ArXiv* (2021).

[20] Colby Wise, Vassilis N Ioannidis, Miguel Romero Calvo, Xiang Song, George Price, Ninad Kulkarni, Ryan Brand, Parminder Bhatia, and George Karypis. 2020. COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv:2007.12731* (2020).

[21] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* 59, 8 (2019), 3370–3388.

[22] Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, and Chee Keong Kwoh. 2022. Graph representation learning in bioinformatics: trends, methods and applications. *Briefings in Bioinformatics* 23, 1 (2022), bbab340.

[23] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. 2021. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics* 12 (2021).

[24] Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. Dgl-ke: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 739–748.

[25] Da Zheng, Minjie Wang, Quan Gan, Xiang Song, Zheng Zhang, and George Karypis. 2021. Scalable graph neural networks with deep graph library. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1141–1142.

[26] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.

[1] https://ai.tencent.com/ailab/ml/KDD-Deep-Graph-Learning.html

[2] https://naixlee.github.io/Product_Knowledge_Graph_Tutorial_KDD2021/