Knowledge-Guided Graph Neural Networks for Low-Data Molecular Property Prediction

Yenuli Bimanya Indigahawela Gamage
School of Computing
Informatics Institute of Technology (IIT)
Colombo, Sri Lanka
yenuli.20221090@iit.ac.lk

Banuka Athuraliya
School of Computing
Informatics Institute of Technology (IIT)
Colombo, Sri Lanka
banu.a@iit.ac.lk

***Abstract*—Molecular property prediction plays a central role in drug discovery, toxicity assessment and computational chemistry. However, most real world datasets in these domains are extremely small and recent large foundation models struggle to generalize in low-data and geometry-sensitive scenarios. To address this problem, we propose a knowledge-guided graph neural network (GNN) framework designed specifically for data-efficient molecular learning. The first contribution is a lightweight architecture that incorporates fundamental chemical priors. These priors such as functional groups and molecular descriptors to guide representation learning without relying on large scale pretraining. The second contribution is transfer learning embedding level knowledge distillation strategy where a pretrained teacher GNN transfers high-level structural knowledge to a compact student model suitable for small molecule datasets. Preliminary expert feedback from 16 domain specialists indicates that the proposed design is both feasible and relevant for drug discovery use cases although experimental validation is planned as future work. The framework is expected to deliver improved performance in cold-start and low data settings. The conceptual study outlines a promising direction for efficient molecular property prediction using knowledge-augmented GNNs.**

***Index Terms*—graph neural networks, molecular property prediction, data-efficient molecular learning, transfer learning, knowledge distillation, chemical priors**

I. Introduction

Molecular property prediction is a fundamental component of drug discovery, toxicity screening and chemical design. Machine learning approaches have shown substantial promise in automating these tasks; however, real-world molecular datasets are often extremely small, with only a limited number of labeled compounds available. In such low-data settings, large pretrained molecular language models and SMILES-based transformers frequently struggle to generalize, particularly when geometric interactions and subtle structural variations dominate molecular behavior [1], [2], [3].

Graph neural networks (GNNs) have emerged as a powerful alternative by representing molecules as graphs and explicitly modeling atomic connectivity and local chemical environments [4], [5]. Despite their favorable inductive bias, GNNs trained from scratch on small datasets often fail to learn robust structure–property relationships due to insufficient supervision. Recent studies suggest that incorporating chemical knowledge, transfer learning and distilled representations can significantly improve learning efficiency under data-scarce conditions [6], [7], [8].

Motivated by these observations, this paper proposes a knowledge-guided GNN framework tailored for low-data molecular property prediction. The framework integrates lightweight chemical priors into the representation process and employs embedding-level knowledge distillation to transfer structural information from a pretrained teacher GNN to a compact student model. Preliminary expert feedback from chemistry and bioinformatics researchers indicates that the proposed design is both feasible and relevant for practical early-stage drug discovery applications.

II. Background and Motivation

Accurate molecular property prediction plays a central role in modern drug discovery pipelines, where experimental assays are costly and time consuming. While deep learning models have achieved strong performance on large molecular datasets, their effectiveness degrades significantly when only limited labeled data are available. This challenge is particularly pronounced for tasks that require sensitivity to subtle structural differences between molecules.

Graph neural networks provide a more suitable inductive bias for molecular learning by explicitly encoding atomic connectivity and local chemical environments [4], [5]. However, when trained on small datasets, GNNs often lack sufficient supervision to learn stable and generalizable representations. This limitation motivates the use of transfer learning strategies that reuse structural knowledge learned from large, unlabeled molecular corpora.

Pretrained molecular GNNs such as GROVER and Chemprop have demonstrated that transferable structural priors can improve downstream molecular property prediction [7], [8]. Building on this idea, embedding-level knowledge distillation offers an effective mechanism for compressing rich pretrained representations into lightweight student models, improving data efficiency while preserving essential structural information [6]. Together, these considerations motivate frameworks that combine transfer learning, distillation and chemically meaningful priors to address the challenges of low-data molecular prediction.

III. Related Work

Molecular property prediction has been extensively stud- ied using both graph-based and language-based architectures. Early work on message passing neural networks (MPNNs) demonstrated that GNNs can effectively model atom–bond interactions and outperform traditional descriptor-based methods on quantum chemistry tasks [5]. The Graph Isomorphism Network (GIN) further established the expressive power of GNNs for distinguishing subtle structural variations critical to chemical activity prediction [4].

Large-scale molecular pretraining has significantly advanced the field. Models such as GROVER introduced self-supervised graph transformers trained on millions of molecules and achieved strong performance across multiple MoleculeNet benchmarks [7], [8]. Similarly, Chemprop leveraged directed message passing and ensemble strategies to set competitive baselines for molecular property prediction. Despite these successes, pretrained GNNs still require moderate amounts of labeled data for fine-tuning and often exhibit unstable performance in low-data regimes [2].

Transformer-based molecular foundation models have also gained attention. ChemBERTa-2 adapted masked language modeling to SMILES representations [3], while MolE introduced multiscale architectures incorporating both 2D and 3D structural information [9]. However, benchmarking studies consistently show that large language models struggle to generalize reliably under data-scarce and structure-sensitive molecular settings [2].

Knowledge distillation has been widely explored as a model compression and representation learning technique, including for graph-structured data [6]. In parallel, integrating domain knowledge such as functional group indicators and physicochemical descriptors has been shown to improve molecular prediction performance [10]. In contrast to prior work, the proposed framework combines embedding-level distillation from pretrained molecular GNNs with lightweight chemical priors specifically designed for low-data molecular property prediction. This integration aims to preserve transferable structural knowledge while guiding learning with chemically meaningful inductive biases, addressing a gap between general-purpose distillation methods and practical early-stage molecular discovery constraints.

IV. Proposed Framework

This section describes the proposed knowledge-guided GNN framework designed to improve molecular property prediction under low-data conditions. The approach integrates three key components: (A) a pretrained teacher GNN used for transfer learning, (B) a lightweight student GNN optimized for data-efficient inference and (C) embedding-level knowledge distillation to transfer structural priors alongside (D) chemical priors that guide representation learning. The overall design aims to capture transferable structural information while keeping the student model compact, interpretable and stable under limited supervision.

A. Pretrained Teacher GNN (Transfer Learning)

Large molecular GNNs such as GROVER and Chemprop have been trained on millions of unlabeled compounds and learn rich structural representations useful across downstream tasks [7], [8]. In the proposed framework, one such pretrained GNN acts as the teacher model. Its parameters remain fixed and it generates latent embeddings for molecules in the low- data target dataset. These embeddings encode transferable chemical knowledge related to atomic environments, molec- ular substructures and topology. Using a frozen teacher also ensures computational efficiency, as no gradient updates are required for the high-capacity model during training. This setup allows the student model to benefit from broad chemical knowledge even when the downstream dataset contains only a few hundred samples.

B. Lightweight Student GNN

A compact GNN such as Graph Convolutional Network (GCN) or Graph Isomorphism Network (GIN) serves as the student model. This model is intentionally small to avoid overfitting in low-data scenarios. Without guidance, such lightweight architectures struggle to learn robust structure-property relationships. However, when paired with a strong teacher model, they can approximate more expressive molecular representations while maintaining data efficiency. The reduced parameter count also leads to faster training, lower memory usage and enhanced stability across small random initialization seeds, which is an important property when dealing with datasets that contain limited structural diversity.

C. Embedding-Level Knowledge Distillation

To transfer structural priors from the teacher to the student, we employ embedding-level knowledge distillation [6]. During training, the student is encouraged to align its intermediate representations with the teacher's embeddings using a combination of L1 distance and cosine similarity losses. This encourages the student model to mimic the teacher's internal representation space while still optimizing directly for the downstream prediction objective. The dual-objective training setup helps the student capture both local atomic interactions and broader molecular patterns. By aligning embeddings at intermediate layers rather than only at the prediction layer, the framework preserves more nuanced structural information that is often lost in output-level distillation.

D. Integration of Chemical Priors

To improve inductive bias and stabilize learning under limited supervision, the proposed framework incorporates lightweight chemical priors derived directly from molecular structure. These priors are designed to be task-agnostic, inexpensive to compute, and free from information leakage.

Specifically, the framework considers functional group indicators extracted using cheminformatics toolkits such as RDKit, including the presence of common moieties (e.g., aromatic rings, hydroxyl groups, amines). These signals can be represented as binary or count-based features and integrated either

as auxiliary prediction targets or as additional inputs following graph-level pooling.

In addition, simple physicochemical descriptors such as molecular weight, hydrogen bond donors and acceptors, and topological polar surface area may be concatenated with learned graph embeddings to provide coarse-grained chemical context. Unlike learned representations, these descriptors encode well-established chemical heuristics and can help anchor the model's latent space in low-data regimes.

Finally, graph-native structural attributes including atom degree, aromaticity flags, and ring membership are retained as node-level features, reinforcing chemically meaningful message passing without increasing model complexity. Importantly, all priors are derived solely from molecular structure and do not rely on external knowledge graphs or bioactivity annotations, ensuring robustness and generalizability to unseen compounds.

E. Preliminary Practical Validation

Informal feedback collected from 16 researchers in cheminformatics and computational chemistry indicated that the proposed combination of transfer learning, distillation and chemical priors is feasible, interpretable and well aligned with the constraints of low-data molecular tasks. Many experts emphasized that the separation between a fixed teacher and a lightweight student is particularly useful in early-stage drug discovery where computational resources and labeled data are limited. This early validation supports the practical relevance of the design and its applicability to drug discovery workflows.

V. Future Evaluation Plan

Although the proposed framework is conceptual, future work will involve systematic empirical evaluation on standard molecular property prediction benchmarks such as ESOL, BACE and ClinTox using scaffold-based splits to prevent overly optimistic performance estimates. In addition to full benchmark datasets, reduced-size subsets from MoleculeNet will be explored to explicitly study how model performance degrades as the amount of labeled data decreases [7].

To assess robustness, activity-cliff-oriented evaluation will be conducted. Activity cliffs, which consist of structurally similar molecules exhibiting large differences in properties, are known to expose limitations in many machine learning models for chemistry. Evaluating student models in these scenarios will help determine whether embedding-level knowledge distillation improves sensitivity to subtle structural variations. Performance will be quantified using standard metrics such as RMSE for regression tasks and ROC-AUC for classification tasks, together with activity-cliff-specific indicators suggested in recent studies [6].

Further experiments will focus on cold-start settings, where only a small number of labeled samples are available for fine-tuning. This scenario is common in early-stage drug discovery and directly reflects the target use case of the proposed framework. Comparisons will be performed against multiple baselines, including (i) a student-only GNN trained from scratch, (ii) a frozen teacher embedding model without distillation and (iii) standard pretrained GNN baselines. These comparisons will enable a clear analysis of the contribution of transfer learning, distillation and chemical priors. To evaluate training stability and reproducibility, experiments will be repeated across multiple random seeds and data subsampling ratios, addressing known concerns regarding performance variance in low-data molecular learning [2].

VI. Conclusion

This paper presented a knowledge-guided graph neural network framework for molecular property prediction in low-data settings. By combining pretrained teacher GNNs, lightweight student architectures, embedding-level knowledge distillation and simple chemical priors, the approach aims to improve structural representation learning without relying on large datasets.

Although the work is conceptual, implementation of the proposed framework has been initiated, with experimental validation planned as ongoing work. Future efforts will focus on completing the prototype, benchmarking against existing GNN and transformer-based models, and assessing robustness under challenging conditions such as scaffold splits and activity cliffs.

References

[1] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, D. Jensen, and R. Barzilay, "Analyzing learned molecular representations for property prediction," *Journal of Chemical Information and Modeling*, vol. 59, no. 8, pp. 3370–3388, 2019. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.9b00237

[2] Z. Zhong, K. Zhou, and D. Mottin, "Benchmarking large language models for molecule prediction tasks," *arXiv preprint*, 2024, arXiv:2403.05075. [Online]. Available: https://arxiv.org/abs/2403.05075

[3] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta-2: Towards chemical foundation models," *arXiv preprint*, 2022, arXiv:2209.01712. [Online]. Available: https://arxiv.org/abs/2209.01712

[4] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint*, 2019, arXiv:1810.00826. [Online]. Available: https://arxiv.org/abs/1810.00826

[5] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *arXiv preprint*, 2017, arXiv:1704.01212. [Online]. Available: https://arxiv.org/abs/1704.01212

[6] Y. Tian, S. Pei, X. Zhang, C. Zhang, and N. V. Chawla, "Knowledge distillation on graphs: A survey," *arXiv preprint*, 2023, arXiv:2302.00219. [Online]. Available: https://arxiv.org/abs/2302.00219

[7] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. S. Pande, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2018/sc/c7sc02664a

[8] Y. Rong, Y. Bian, T. Xu, W. Chen, and J. Huang, "Grover: Self-supervised graph transformer on large-scale molecular data," *arXiv preprint*, 2020, arXiv:2007.02835. [Online]. Available: https://arxiv.org/abs/2007.02835

[9] O. Me´ndez-Lucio *et al.*, "Mole: A foundation model for molecular graphs using disentangled attention," *Nature Communications*, vol. 15, p. 53751, 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-53751-y

[10] S. Riniker and G. A. Landrum, "Open-source platform to benchmark fingerprints for ligand-based virtual screening," *Journal of Cheminformatics*, vol. 5, p. 26, 2013. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-5-26