

图神经网络在药物分子性质预测中的研究综述（2022 以来）

药物分子性质预测是 AI 辅助药物发现的核心环节之一，其目标是从分子结构出发高效、准确地估计化合物的理化性质与生物活性。图神经网络（GNN）凭借对分子图结构天然的建模能力，已成为该领域的主流模型范式。本文以 2022 年以来发表的 16 篇代表性论文为对象，基于其题录与摘要，按方法演进脉络梳理图神经网络在药物分子性质预测中的研究现状、主要技术路线与现存问题，以期对相关研究提供系统性的参考。

一、研究现状

2022 年以来，图神经网络在药物分子性质预测领域经历了一轮快速演进，整体呈现出从「标注数据驱动的监督学习」向「大规模自监督预训练 + 下游微调」范式迁移的明显趋势。在自监督表示学习方面，MolCLR 提出分子对比学习框架，利用 GNN 编码器在约千万个无标注分子上做自监督预训练，通过原子掩码、键删除、子图移除三种图增广最大化同一分子不同增广之间的一致性，微调后在多个基准上取得 SOTA 表现^[2]。Noisy Nodes 则从正则化角度出发，通过向输入图注入噪声并施加噪声校正的节点级损失来缓解 GNN 过平滑问题，在量子化学任务与 OGB 数据集上均有显著提升^[3]。KPGT 以 Line Graph Transformer (LiGhT) 为高容量骨干、强调化学键建模，并以知识引导的预训练策略从大规模无标注分子图中捕获结构与语义信息^[5]。面向 SMILES 表示的 K-BERT 与 FP-BERT 则分别以原子/分子特征预测、对比学习等预训练任务提取类化学家信息与化合物指纹语义表示^{[6][7]}。

三维结构信息成为本轮演进的重要方向。Uni-Mol 提出通用 3D 分子表示学习框架，以两个同构的 SE(3) Transformer 分别对分子构象与蛋白口袋进行预训练，在 14/15 个分子性质预测任务上超越当时 SOTA^[1]。Unified 2D and 3D Pre-Training 联合 2D 图与 3D 构象做统一预训练，在 11 个下游任务中 10 个达到 SOTA^[4]；基于去噪的预训练方法则利用大规模平衡态 3D 分子结构学习表示，在 QM9 等基准的多数目标上达到 SOTA^[10]；3D PGT 在 3D 分子图上预训练后再迁移到无 3D 结构图^[11]。

与此同时，扩散模型与大规模语言模型（LLM）正加速渗透到分子表示与生成任务中。EDM 提出 E(3) 等变的扩散生成模型，显著超越此前 3D 分子生成方法^[9]；数据中心视角的扩散模型则利用大规模无标注图生成任务特定的图样本及其标签用于数据增广^[13]。在文本融合方向，GPT-MolBERTa 借助 ChatGPT 收集分子文本描述训练自监督语言模型^[14]，MolXPT 将分子名替换为 SMILES 以统一文本与分子做生成式预训练^[15]，MolFM 则联合分子结构、生物医学文本与知识图谱三种模态构建多模态基础模型^[16]。此外，面向真实药研场景的鲁棒性问题开始受到重视：DrugOOD 系统性策展了分布外（OOD）数据集与基准^[12]，基于消息传递神经网络集成的方法则将校准的概率预测分布引入性质预测^[8]。

二、主要技术路线

（一）预训练与自监督表示学习。这一路线以大规模无标注分子为数据源，通过设计自监督任务学习通用分子表示。MolCLR 采用对比学习框架，用三种图增广构造正样本对并最大化其一致性，显著提升了分类与回归性质预测的性能^[2]；Noisy Nodes 以噪声校正的节点级损失缓解过平滑、鼓励有意义的节点表示^[3]；Unified 2D and 3D Pre-Training 将原子坐标与原子间距编码并融合进 GNN，预训练任务涵盖掩码原子与坐标重建、基于 2D 图的 3D 构象生成等^[4]。在序列/指纹表示一侧，K-BERT 通过原子特征预测、分子特征预测与对比学习三种预训练任务提取类化学家的化学信息^[6]，FP-BERT 则以自监督方式预训练 BERT 编码器获取化合物指纹的语义表示、再接入 CNN 提取高阶特征^[7]。

（二）3D 分子表示学习与几何预训练。该路线将分子建模从 2D 图结构拓展到三维空间。Uni-Mol 用两个同构 SE(3) Transformer 分别预训练分子模型与蛋白口袋模型，分子模型以 2.09 亿构象预训练、口袋模型以 300 万蛋白口袋数据预训练，除性质预测外还擅长蛋白-配体结合姿态预测等 3D 任务^[1]；基于去噪的预训练以去噪目标近似学习分子力场（以高斯混合近似玻尔兹曼分布），从而利用平衡态 3D 结构^[10]；3D PGT 基于键长、键角、二面角三种几何描述符设计多任务生成式预训练，并以总能量为代理指标自动搜索三个前置任务的权重分配^[11]。

（三）分子图 Transformer 与知识引导。KPGT 针对预训练任务定义不当与模型容量不足两大问题，以 Line Graph Transformer 为高容量骨干、重点强调化学键建模，并通过知识引导的预训练策略从无标注分子图中捕获结构与语义信息^[5]。

（四）GNN 与大语言模型的文本融合。该路线尝试让分子表示借助自然语言文本获取更丰富的语义。GPT-MolBERTa 用 ChatGPT 收集 32.6 万个分子的文本描述训练自监督大语言模型学习分子表示，注意力分析显示模型能够捕捉文本中的关键信息^[14]；MolXPT 将文本中的分子名替换为 SMILES，使 SMILES 能利用上下文文本信息，在 MoleculeNet 上优于强基线并支持零样本分子生成^[15]；MolFM 通过原子-分子实体-文本间的跨模态注意力增强对三种模态的联合理解，在跨模态检索任务上显著提升^[16]。

（五）扩散模型结合分子图。EDM 用等变网络联合处理原子坐标（连续）与原子类型（类别）特征以对扩散过程去噪，并给出可计算分子似然的概率分析，显著超越此前 3D 分子生成方法^[9]；Data-Centric Learning 则从数据增广角度出发，设计两个新目标以任务标注数据引导去噪过程，生成任务特定的图样本及其标签，在 15 个任务上显著优于 15 种现有方法^[13]。

（六）经典 GNN 表示学习与不确定性量化。在提升预测可靠性的方向上，基于消息传递神经网络集成的方法将 GNN 扩展出校准的概率预测分布，在统一框架下同时考虑任意（aleatoric）与认知（epistemic）不确定性，并在未见数据上重校准预测分布，在 QM9 与 PC9 上实现校准良好的性质预测^[8]。

（七）OOD 基准与鲁棒性评测。DrugOOD 面向 AI 辅助药物发现（ADMET 预测、虚拟筛选）提供系统性的 OOD 数据集策展与基准，聚焦药物-靶点结合亲和力预测，提供自动化策展脚本、领域标注与噪声标注，并揭示出分布内与分布外实验之间存在显著性能差距^[12]。

三、存在问题

综合上述文献，当前图神经网络在药物分子性质预测中仍面临若干突出问题。**其一，分布外泛化能力不足。**DrugOOD 揭示出分布内与分布外实验之间存在显著的性能差距，说明模型在训练数据分布之外的泛化仍不可靠，而这恰是真实药物发现场景的常态^[12]。

其二，**预训练任务设计与模型容量的权衡**。KPGT 明确指出预训练任务定义不当与模型容量不足是制约自监督学习效果的两大瓶颈^[5]；即便强如去噪预训练，其表示质量仍高度依赖大规模 3D 结构数据的获取与力场假设^[10]。其三，**图表示的过平滑问题**。Noisy Nodes 的工作表明，深层 GNN 的节点隐表示易趋于相似、出现过平滑，需要专门的正则化手段加以缓解^[3]。其四，**预测不确定性的量化仍待完善**。尽管集成方法已能给出校准良好的概率预测分布^[8]，但不确定性估计在实际高价值决策（如先导化合物筛选）中的应用仍处于探索阶段。其五，**三维数据依赖与计算成本**。多条 3D 技术路线均以大规模构象或平衡态结构为预训练前提，如 Uni-Mol 以 2.09 亿构象预训练分子模型^[1]、去噪预训练依赖平衡态 3D 结构^[10]、3D PGT 需在 3D 分子图上预训练^[11]，此类数据获取与计算开销较大，且如何将 3D 预训练知识高效迁移到无 3D 结构的分子上仍是开放问题^[11]。其六，**多模态与文本融合的可靠性**。基于 LLM 的分子表示依赖训练文本的质量与规模，如 GPT-MolBERTa 的文本描述由生成式模型产出^[14]，其噪声与偏差对下游性质预测的影响尚需更充分的验证。

参考文献

1. Zhou G, et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. NeurIPS 2022. DOI: 10.26434/chemrxiv-2022-jjm0j-v4.
2. Wang Y, et al. Molecular Contrastive Learning of Representations via Graph Neural Networks (MolCLR). Nature Machine Intelligence, 2022. DOI: 10.1038/s42256-022-00447-x.
3. Godwin J, et al. Simple GNN Regularisation for 3D Molecular Property Prediction and Beyond. ICLR 2022. DOI: 10.48550/arxiv.2106.07971.
4. Zhu J, et al. Unified 2D and 3D Pre-Training of Molecular Representations. KDD 2022. DOI: 10.1145/3534678.3539368.
5. Li H, et al. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction. KDD 2022. DOI: 10.1145/3534678.3539426.
6. Wu Z, et al. Knowledge-based BERT: a Method to Extract Molecular Features Like Computational Chemists. Briefings in Bioinformatics, 2022. DOI: 10.1093/bib/bbac131.
7. Wen N, et al. A Fingerprints Based Molecular Property Prediction Method Using the BERT Model. Journal of Cheminformatics, 2022. DOI: 10.1186/s13321-022-00650-3.
8. Busk J, et al. Calibrated Uncertainty for Molecular Property Prediction Using Ensembles of Message Passing Neural Networks. Machine Learning: Science and Technology, 2022. DOI: 10.1088/2632-2153/ac3eb3.
9. Hoogeboom E, et al. Equivariant Diffusion for Molecule Generation in 3D (EDM). ICML 2022. DOI: 10.48550/arxiv.2203.17003.
10. Zaidi S, et al. Pre-training via Denoising for Molecular Property Prediction. ICLR 2023. DOI: 10.48550/arxiv.2206.00133.
11. Wang X, et al. Automated 3D Pre-Training for Molecular Property Prediction (3D PGT). KDD 2023. DOI: 10.1145/3580305.3599252.
12. Ji Y, et al. DrugOOD: Out-of-Distribution Dataset Curator and Benchmark for AI-Aided Drug Discovery. AAAI 2023. DOI: 10.1609/aaai.v37i7.25970.
13. Liu G, et al. Data-Centric Learning from Unlabeled Graphs with Diffusion Model. NeurIPS 2023. DOI: 10.48550/arxiv.2303.10108.

14. Balaji S, et al. GPT-MolBERTa: GPT Molecular Features Language Model for Molecular Property Prediction. arXiv (CoRR abs/2310.03030), 2023. DOI: 10.48550/arxiv.2310.03030.
15. Liu Z, et al. MolXPT: Wrapping Molecules with Text for Generative Pre-training. ACL 2023. DOI: 10.48550/arxiv.2305.10688.
16. Luo Y, et al. MolFM: A Multimodal Molecular Foundation Model. arXiv (CoRR abs/2307.09484), 2023. DOI: 10.48550/arxiv.2307.09484.