Hybrid Graph-Language Model for Chemical Property Prediction

Andreas Varvarigos
andreas.varvarigos@yale.edu
Yale University
New Haven, CT, USA

Abstract

Understanding the molecular properties of chemical compounds is fundamental in fields such as drug discovery, disease prediction, and personalized medicine. While Graph Neural Networks are effective at learning representations from molecular graphs, they often exhibit limited generalization beyond training distributions. In contrast, Large Language Models benefit from broad pretraining and strong generalization capabilities. In this work, we propose a novel GNN-LLM architecture that unifies graph-structured and textual information by injecting GNN-derived graph embeddings directly into the LLM's token stream using custom special tokens. Additionally, we provide textual context derived from molecular graph statistics to support the LLM's reasoning. This enables the model to accurately generate molecular property descriptors such as weight, polarity, or atom counts in natural language. Our results on multiple chemical descriptor prediction tasks show that this hybrid system successfully leverages both learned molecular structure and pretrained textual knowledge, demonstrating its potential for integration in cheminformatics and genomic pipelines.

1 Introduction and Motivation

Scientific discovery in chemistry and biology increasingly relies on computational tools that can reason about molecules. From designing new drugs to simulating biochemical pathways, the ability to predict how a compound behaves based on its structure is central to progress in fields like personalized medicine, materials engineering, and genomics. Yet, molecular structures are not naturally suited to traditional machine learning pipelines: they are graphs with intricate topologies, not flat feature vectors or plain text.

Graph Neural Networks (GNNs) have become the de facto tool for modeling molecular structures, as they are well-suited to capture the local and global relational patterns in graph-structured data such as atoms and bonds [24, 5]. GNNs exploit inductive biases rooted in neighborhood aggregation and message passing, allowing them to extract structure-aware embeddings that are highly informative for tasks like molecular property prediction. However, GNNs operate purely on graph connectivity and node/edge features, which means they often lack the domain-level understanding necessary to generalize or reason beyond their training distribution, like the underlying physical laws, chemical principles, or terminologies that govern molecular behavior.

In contrast, Large Language Models (LLMs) are pretrained on vast amounts of scientific literature and chemical databases. That way, they have implicitly learned high-level reasoning about chemical classes, quantitative relationships, and the semantics of molecular behavior [3, 4]. For instance, they might associate aromaticity with specific stability trends, or understand that compounds with large polar surfaces tend to be less membrane-permeable. While LLMs can reason about molecular concepts described in text, they do not natively process graphs, and thus cannot leverage the rich topological data embedded in molecular structure.

Project source code: https://github.com/varvarigos/GenLLM-chem-predictor
Video Presentation Link: https://github.com/varvarigos/GenLLM-chem-predictor/blob/main/docs/video_presentation.mp4

This work addresses the gap between structural representation and domain reasoning. We propose a hybrid GNN-LLM system for molecular property prediction. The model uses a GNN to encode a molecular graph into a learned vector representation and introduces this vector into an LLM's input prompt through special tokens. To complement this, we also provide textual context to help the LLM contextualize the graph embedding. This setup enables the LLM to both attend to structural features and apply its latent scientific knowledge to generate descriptor predictions. By bridging symbolic reasoning with geometric understanding, our model aims to improve generalization and interpretability in tasks critical to real-world chemistry, medicine, and molecular research.

2 Background and Related Work

2.1 GNNs for Molecular Property Prediction

GNNs are widely used for molecular property prediction due to their ability to model atoms and bonds as graphs. Early approaches such as GCNs [12] and GATs [23] enabled neighborhood-based message passing and attention over molecular graphs. Spatial and geometric GNNs incorporate 3D information: SchNet [21] used continuous-filter convolutions to model quantum interactions, while DimeNet [13] introduced directional message passing with angular features to capture molecular geometry. Its successor GemNet [10] further integrated multi-hop messaging and equivariant design, achieving SOTA accuracy on quantum chemistry benchmarks. These GNN models demonstrate strong performance on supervised property prediction tasks like QM9.

Despite their success, GNNs face significant limitations. They often fail to generalize to novel or out-of-distribution molecules, particularly in scaffold-based splits [11], and tend to memorize local patterns rather than capture higher-level chemical concepts. Moreover, standard GNNs operate purely on structure and lack domain-level chemical reasoning. They do not incorporate explicit chemical knowledge (e.g. functional groups, reactivity rules) and must infer such patterns implicitly. As a result, a GNN might predict a property correctly but cannot explain it in terms a chemist would use. This black-box nature ties into the issue of interpretability; some efforts introduce attention or explanation techniques (e.g. GNNExplainer [25]) to highlight influential atoms or bonds, but these still fall short of providing human-understandable rationales.

2.2 LLMs in Chemistry

LLMs have been applied to chemistry by treating molecular problems as language tasks using textual or linearized representations. Prior work spans molecule captioning – e.g. MolT5 learned to translate between molecular structures and English descriptions, enabling fluent captions of a molecule's features and even text-to-molecule generation [6] – and property extraction from literature, where LLMs parse scientific text to pull out reported compound properties or experimental results. LLMs have also been employed for SMILES generation and completion: models like ChemGPT [8] and DrugGPT [16], built on GPT-style architectures and fine-tuned on millions of SMILES strings, can suggest novel compounds or complete a molecule given a partial sequence. Additionally, domain-specific LLMs such as BioGPT [19] (pre-trained on biomedical text) and even general chat models (GPT-3.5/4) have been used for chemistry question-answering [1], demonstrating an ability to answer queries about drug mechanisms or chemical facts in natural language.

The key strength of LLM-based approaches lies in their broad generalization and semantic reasoning capabilities. Because they are pre-trained on vast text corpora, LLMs possess extensive domain knowledge and can integrate dispersed information to tackle novel questions beyond narrow training data. However, a major limitation is the lack of true structural grounding, since an LLM perceives a molecule only as a sequence of tokens rather than a 3D bonded structure. Subtle structural differences (for example, constitutional isomers) that would be obvious from a graph perspective can be hard for a text-only model to distinguish [9]. Indeed, fine-tuned LLMs have been shown to assign similar likelihoods to original vs. corrupted molecular sequences, revealing their limited understanding of molecular structure [15]. Consequently, purely text-based LLMs may rely on learned associations and sometimes produce plausible-sounding hallucinations or errors about a molecule because they do not truly comprehend its structure [26].

2.3 Hybrid Structure–Language Models in Chemistry

Given the strengths of GNNs and LLMs, recent work has explored hybrid models that combine molecular structure representations with natural language. One direction looks at connecting molecular graphs and textual descriptions by embedding them into a shared space. One such example is MoleculeSTM (Multi-modal Molecule Structure-Text Model) by Liu et al., which combines a GNN for encoding molecular structures with a transformer for text. The two are trained jointly using contrastive learning, encouraging aligned representations across modalities [17]. This enables cross-modal tasks such as retrieving the closest molecule given a text query and even limited text-based editing of molecular structures. However, the fusion in such models is relatively shallow. MoleculeSTM learns parallel representations for two modalities and aligns them, but during generation or prediction it does not deeply intermix structural and textual features since the graph and text components essentially communicate only through a learned embedding similarity.

Other efforts incorporate molecular information into the prompts or architecture of LLMs. MoLFormer [20] is a large-scale transformer trained on SMILES strings of over a billion molecules, achieving powerful "chemical language" understanding. While MoLFormer is not multi-modal (it sees only SMILES, no natural language), the model learns embeddings that correlate with molecular properties and even outperformed some graph-based models on benchmark tasks. However, because it relies on 1D SMILES input, it may still struggle to utilize certain structural information and it does not incorporate general natural language capabilities. Moving toward multimodal integration, MolT5 and related models [27] extend text generation models to handle molecule identifiers. MolT5 was built by pretraining a T5 encoder-decoder on a mix of chemical formulas, SMILES, and textual data, enabling tasks like molecule-to-text translation and vice versa. Such sequence-to-sequence approaches can produce fluent text from molecules, but they typically require the molecule to be encoded as a string (SMILES or InChI), which is an inefficient input format and often leads the LLM to underuse the structural information.

Several recent systems attempt a tighter integration of graphs with LLMs. MolXM and MolXPT [18] introduce graph-aware transformers where a molecular graph encoder is paired with a language model for downstream tasks like text-based molecule design and property prediction. Typically, these use a two-stage approach: first encode the graph, then feed the encoded vector or sequence into the LLM. While this indeed allows the LLM to condition on structural data, without careful design the LLM can ignore the graph input. Fang et al. [7] observed that a naively fine-tuned multi-modal model often relies predominantly on the textual channel and neglects the graph modality, unless the model is explicitly forced to attend to it. This suggests that many current hybrids achieve only a superficial fusion: the structural features are present but not deeply integrated in the model's reasoning. Another limitation of early hybrids is that they sometimes treat the graph input in an ad-hoc way, for example converting a molecule's graph into a large enumeration of descriptors or an image, and then feeding that to the LLM. ChemCrow [2] takes a systems approach rather than a model architecture approach; specifically, it uses an LLM to decide which external chemistry tool to call and, then incorporates the result into its final answer. This approach successfully tackles complex tasks like multi-step synthesis planning by leveraging external graph-based computations; however, the LLM itself remains a pure text generator and relies on external modules for structural reasoning.

2.4 From Structure-Language Alignment to Deep Integration

The limitations in existing GNN- and LLM-based approaches, as well as in their prior combinations, point toward a need for deeper and more intentional fusion of structural and linguistic modalities. While contrastive models like MoleculeSTM [17] and retrieval-based pipelines offer useful alignment between molecules and their textual representations, they fall short in enabling generative, structure-conditioned reasoning. Similarly, SMILES-centric transformer architectures [20, 27, 16] often process chemical information in a syntactic rather than structural way, lacking explicit awareness of molecular topology or geometry. Attempts to bridge the gap via prompt engineering or external feature injection, such as in MolXPT [18] or ChemCrow [2], frequently underutilize structural signals due to shallow integration or reliance on post-hoc tools.

This work advances molecular representation learning by moving beyond post-hoc alignment or auxiliary conditioning, toward a model in which structural and linguistic modalities are integrated during language generation. We condition a language model directly on GNN-derived graph embeddings and concise structural summaries, introduced at the token level. This setup enables the model to

generate property descriptions while attending explicitly to molecular structure, rather than relying solely on memorized patterns or superficial textual signals. The result is a unified architecture that could improve both predictive accuracy and interpretability, particularly in descriptor prediction tasks where structural grounding is essential.

3 Methodology

We propose a hybrid GNN-LLM architecture that combines molecular graph structure with natural language prompts to generate chemical descriptors. This section outlines our data processing pipeline, model design, and training setup.

3.1 Dataset Construction

We use the AIDS Antiviral Screen dataset [22], which provides molecular graphs where atoms are nodes and bonds are edges. Each node is represented as a one-hot vector over 31 atom types, and each edge as a one-hot vector over 3 bond types (single, double, triple). These graph features are passed directly to our GNN model.

To generate natural language targets, we translate each graph into a canonical SMILES string using an internal conversion script. These SMILES strings are fed into RDKit [14] to compute 30 chemical descriptors spanning physicochemical, electronic, and topological properties. The descriptors are formatted as a natural language sentence and used as generation targets. The full list of descriptors appears in Table 1.

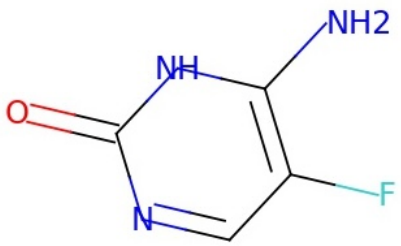

Figure 1: Example of a molecular graph from the AIDS dataset.

3.2 Model Architecture

The architecture, shown in Figure 2, consists of three components: a GNN to encode the graph, a projection layer, and a pretrained LLM for autoregressive generation.

Given a molecular graph $G = (V, E)$, node features $X \in \mathbb{R}^{N\times 31}$ and edge features $E \in \mathbb{R}^{M\times 3}$ are passed through a multi-layer GAT. After L layers of message passing, we apply mean pooling to obtain a graph embedding $\mathbf{h}_G \in \mathbb{R}^{d_g}$:

$$\mathbf{h}_G = \frac{1}{|V|} \sum_{v\in V} \mathbf{h}_v^{(L)}.$$

This embedding is projected to the LLM's embedding dimension d_{LLM} using a learned linear layer:

$$\tilde{\mathbf{g}} = W\mathbf{h}_G + b, \quad \tilde{\mathbf{g}} \in \mathbb{R}^{d_{\text{LLM}}}.$$

To condition the LLM on structural input, we introduce a special pseudo-token <|GRAPH_EMBEDDING|> with embedding $\tilde{\mathbf{g}}$. We wrap this token between <|GRAPH_START|> and <|GRAPH_END|> to help the LLM identify and attend to the graph representation.

The final input prompt has the form:

```
<BOS> <|GRAPH_START|> <|GRAPH_EMBEDDING|> <|GRAPH_END|>
Contextual text ...  Natural language query <EOS>
```

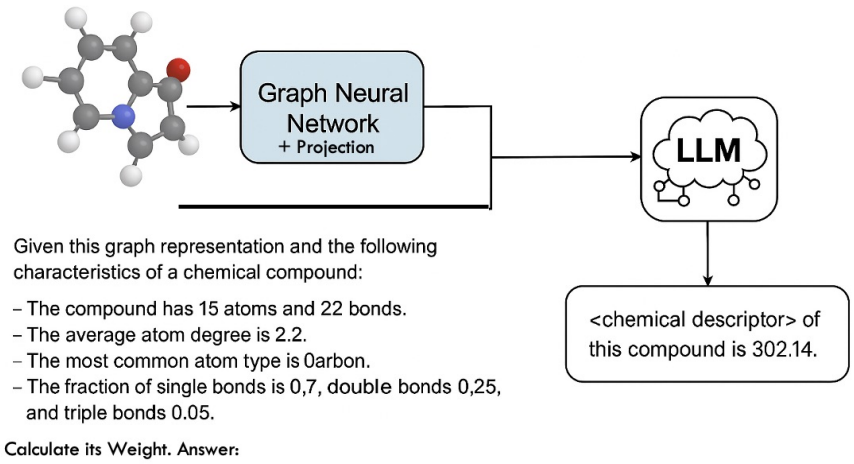

Figure 2: Overview of the GNN-LLM architecture.

3.3 Training Objective

The model is trained using a causal language modeling objective. Let $x_{1:P}$ be the tokenized prompt (including the graph token) and $y_{1:T}$ the tokenized descriptor. We minimize the negative log-likelihood of the target tokens:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x_{1:P}).$$

All model parameters—the GNN, projection, and LLM—are trained end-to-end. This objective encourages the model to align the injected graph embedding with the output descriptor tokens.

4 Results

Each test instance in our evaluation consists of a molecular graph paired with a specified descriptor (one of 30 possible descriptors). We treat each descriptor as an independent prediction task: the model is prompted to output the scalar value of that descriptor in textual form given the graph. The predicted text is parsed to extract a numeric value, and we compute the mean absolute error (MAE) between this predicted scalar and the ground-truth value. The final reported MAE is the average over all test samples and all descriptors.

We first consider a purely graph-based baseline. A GNN regressor (no language component) is trained end-to-end with a standard MSE loss on the scalar targets. This baseline achieves a **test MAE of 67.67** on our descriptor dataset.true descriptor values by 67.67 units. By contrast, our proposed GNN–LLM hybrid model (where the GNN output is converted to a prompt for an LLM to generate the descriptor text) yields a much higher **test MAE of 99.15**. Crucially, this MAE is computed only over the subset of outputs that produced a valid numeric string: in fact, only 29.95% of the model's outputs were successfully parseable as numbers. The remaining 70.05% of outputs were invalid (e.g. containing words or nonsensical symbols), and thus could not be interpreted as a numeric prediction. In summary, the LLM-based model not only has a higher average error on its valid outputs, but it also fails to produce a numeric prediction in the majority of cases. This low validity of outputs substantially inflates the overall MAE metric.

During training we monitored the LLM's causal language-model (cross-entropy) loss on both the training and validation sets. At convergence, the hybrid model's training cross-entropy loss was 0.2877, and the test (validation) cross-entropy loss was 0.4833. The loss curves for both the training and test splits are plotted in Appendix Figure 3. As shown there, the two curves converge without a large gap, indicating that the model does not overfit despite its size. In particular, the moderate increase from train to test loss suggests reasonable generalization given the limited data.

The LLM we use is Microsoft's Phi-1.5, a 1.3-billion-parameter Transformer pre-trained on a "textbook-quality" corpus of reasoning and STEM-focused text, selected for its strong logical reasoning capabilities and near SOTA performance on smaller-model benchmarks. However, like other powerful generative LLMs, it is prone to "hallucinating" plausible-sounding but incorrect or irrelevant content. In our task, such hallucinations typically manifest as invalid descriptor strings (e.g. spelling out numbers in words, adding context, or simply guessing a wrong format), which explains the low validity rate. These erroneous outputs directly degrade numeric accuracy: if the model "hallucinates" instead of outputting a clean number, the extracted value is either missing or highly off, which in turn inflates the MAE. In short, while the LLM can generate descriptive text, its lack of strict numeric fidelity limits its utility for precise descriptor prediction.

Taken together, these results indicate that the hybrid approach faces significant challenges. The GNN–LLM model's average error is more than twice that of the GNN-only baseline, primarily because the LLM often fails to produce valid numeric outputs. By contrast, the GNN baseline reliably outputs a number and thus achieves a lower MAE. Future work will be needed to address these hallucination errors if such GNN–LLM hybrids are to match the numeric accuracy of direct regression models. Note that all reported experiments were carried out end-to-end by a single graduate student (covering data preparation, model implementation, training, and evaluation).

5 Conclusion and Future Work

We presented a hybrid GNN–LLM architecture for molecular descriptor prediction that integrates learned structural embeddings from a graph neural network with a pretrained large language model (LLM). While the GNN-only baseline achieved superior quantitative performance, the GNN–LLM model successfully learned to generate textual outputs that were often semantically aligned with the target descriptors. Despite its higher MAE and lower output validity, the hybrid model demonstrates that pretrained LLMs can be conditioned on graph-structured molecular data to produce context-aware predictions. Notably, qualitative inspection of many outputs revealed that the model often produced values close to the ground truth, even when the exact format was invalid—suggesting that it has internalized meaningful structure-to-text mappings.

We emphasize that the current results were obtained without extensive hyperparameter tuning or prolonged training, due to resource and time constraints. With more systematic optimization, better LLM initialization, and improvements in output decoding, we expect the model's performance to significantly improve. A promising direction for future work is to reframe the architecture to mitigate hallucination: for example, by assigning a learnable token to each descriptor type, allowing the LLM to act as a context encoder rather than a generator. This would enable a final prediction head—such as a lightweight MLP—to produce the numeric output in a controlled and interpretable fashion, while still leveraging the LLM's rich latent knowledge of chemical properties. More broadly, we view this hybrid modeling approach as a step toward bridging symbolic reasoning and structural understanding in computational chemistry.

Appendix

Descriptor	Short Description
Molecular weight	Total mass of a molecule in Daltons.
Lipophilicity (LogP)	Estimate of molecule's solubility between lipids and water.
Number of hydrogen donors	Count of NH or OH groups capable of donating hydrogen bonds.
Number of hydrogen acceptors	Number of nitrogen or oxygen atoms that can accept hydrogen bonds.
Heavy atom count	Number of atoms in the molecule excluding hydrogens.
Topological polar surface area (TPSA)	Surface area contributed by polar atoms, useful for predicting permeability.
Fraction of sp^3-hybridized carbons	Proportion of carbon atoms with sp^3 hybridization.
Number of rings	Total number of ring systems present in the molecule.
Balaban's J index	Connectivity-based topological index reflecting graph complexity.
Molar refractivity	Estimate of molecular volume and electronic polarizability.
Bertz complexity index	Quantitative measure of molecular structural complexity.
NH or OH count	Number of NH or OH groups, related to hydrogen bonding potential.
NO count	Combined count of nitrogen and oxygen atoms.
Number of aliphatic rings	Count of saturated non-aromatic rings.
Number of aromatic rings	Number of rings with delocalized π electrons.
Number of saturated rings	Count of ring structures with only single bonds.
Number of heteroatoms	Atoms in the molecule that are not carbon or hydrogen.
Number of rotatable bonds	Single non-ring bonds that allow free rotation.
Valence electron count	Total number of valence electrons in the molecule.
Labute ASA	Approximate solvent-accessible surface area.

Table 1: Chemical descriptors calculated for each molecular graph using RDKit, spanning physico-chemical, electronic, and topological properties.

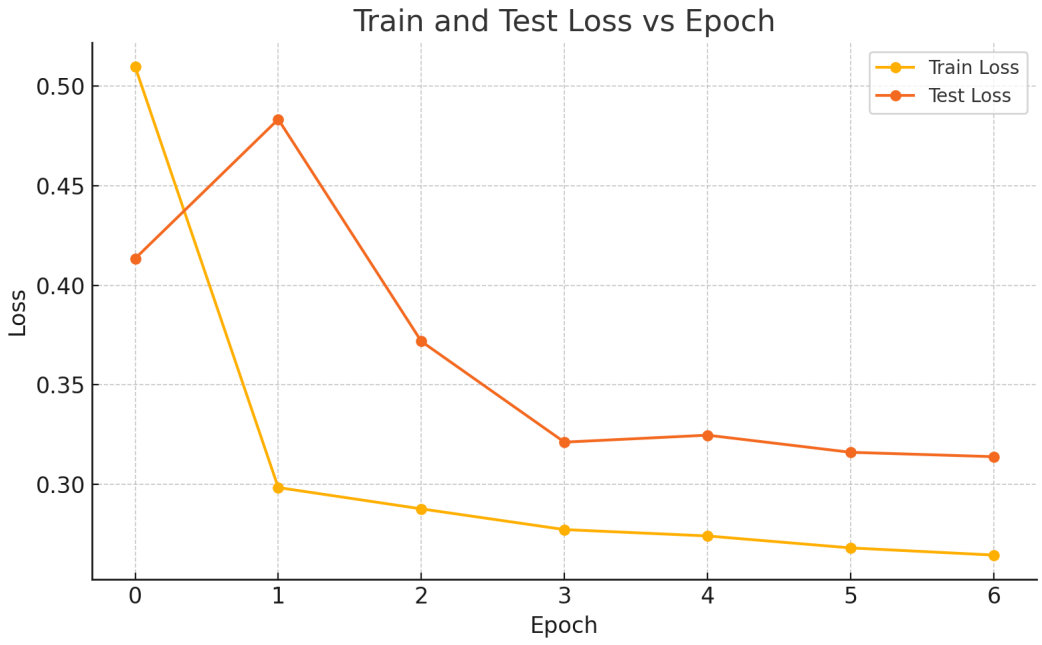

Figure 3: Train and test cross-entropy loss curves across epochs for the GNN-LLM hybrid model.

References

[1] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

[2] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2024.

[3] Andres M Bran, Théo A Neukomm, Daniel Armstrong, Zlatko Jončev, and Philippe Schwaller. Chemical reasoning in llms unlocks steerable synthesis planning and reaction mechanism elucidation. *arXiv preprint arXiv:2503.08537*, 2025.

[4] Andres M Bran, Théo A Neukomm, Daniel Armstrong, Zlatko Joncev, and Philippe Schwaller. Revealing chemical reasoning in llms through search on complex planning tasks. In *Proc. of the 2025 International Conference on Learning Representations (ICLR)*, 2025.

[5] Roman Bresson, Giannis Nikolentzos, George Panagopoulos, Michail Chatzianastasis, Jun Pang, and Michalis Vazirgiannis. Kagnns: Kolmogorov-arnold networks meet graph learning. *arXiv preprint arXiv:2406.18380*, 2024.

[6] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 375–413, 2022.

[7] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

[8] Nathan C. Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Stephan Hemweg, Steven L. Stokes, Heather Van Meter, Sergey Maslov, Vijayet Gadepally, Nathaniel Thomas, and H. Sebastian Seung. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5:1297–1305, 2023.

[9] Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013, 2024.

[10] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 6790–6802, 2021.

[11] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 22118–22133, 2020.

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

[14] Greg Landrum. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2016.

[15] Chanhui Lee, Yuheon Song, YongJun Jeong, Hanbum Ko, Rodrigo Hormazabal, Sehui Han, Kyunghoon Bae, Sungbin Lim, and Sungwoon Kim. Mol-llm: Generalist molecular llm with improved graph utilization, 2024. arXiv preprint arXiv:2502.02810.

[16] Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Wenkai Wang, Zhenling Peng, Jianyi Yang, and Hong Wei. Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins. *bioRxiv preprint*, 2023.

[17] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing, 2023. arXiv preprint arXiv:2212.10789.

[18] Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. Molxpt: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 1606–1616, 2023.

[19] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 2022.

[20] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022.

[21] Kristof Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6543–6552, 2017.

[22] US National Cancer Institute. Aids antiviral screen data. `https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data`, 2017.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[24] Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14:2585, 2023.

[25] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 9244–9255, 2019.

[26] Shuzhou Yuan and Michael Färber. Hallucinations can improve large language models in drug discovery. *arXiv preprint arXiv:2501.13824*, 2024.

[27] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13:862, 2022.