MOL-SGCL: Molecular Substructure-Guided Contrastive Learning for Out-of-Distribution Generalization

Andrew Zhou
Department of Biomedical Informatics
Harvard University
Boston, MA
azhou@hms.harvard.edu

Yasha Ektefaie
Eric and Wendy Schmidt Center
The Broad Institute
Cambridge, MA
yektefai@broadinstitute.org

Maha Farhat
Department of Biomedical Informatics
Harvard University
Boston, MA
maha_farhat@hms.harvard.edu

Abstract

Datasets for molecular property prediction are small relative to the vast chemical space, making generalization from limited experiments a central challenge. We present MOL-SGCL – Molecular Substructure-Guided Contrastive Learning – a method that shapes the latent space of molecular property prediction models to align with science-based priors. We hypothesize that engineering inductive biases directly into the representation space encourages models to learn chemical principles rather than overfitting to spurious correlations. Concretely, MOL-SGCL employs a triplet loss that pulls a molecule's representation toward representations of plausibly causal substructures and pushes it away from implausibly causal ones. Plausibility is defined by querying a large language model with the list of extracted substructures. To stress-test out-of-distribution (OOD) generalization under data scarcity, we construct modified Therapeutics Data Commons tasks that minimize train–test similarity and cap the training set at 150 molecules. On these OOD splits, MOL-SGCL outperforms baselines, indicating that MOL-SGCL promotes invariant feature learning and enhances model generalizability in data-limited regimes. We further demonstrate that MOL-SGCL transfers successfully to a D-MPNN backbone, highlighting that the approach is not tied to a specific architecture. We imagine that MOL-SGCL may enhance AI drug discovery by improving molecular property prediction for novel candidate molecules that are out-of-distribution from existing data. Beyond molecular property prediction, we envision that this could be extended to diverse therapeutics tasks, as long as the inputs can be decomposed into substructures whose presence, absence, or configuration has an influence on the target label.

1 Introduction

Molecular space is vast—on the order of 10^{33} drug-like molecules [Maggiora, 2014, Polishchuk et al., 2013], but measured screening datasets are many orders of magnitude smaller and constrained by cost [Walters and Barzilay, 2020]. In AI-driven drug discovery, the objective is often to identify

Published in The NeurIPS 2025 Workshop on AI for Science: The Reach and Limits of AI for Scientific Discovery, San Diego, California, USA, 2025.

structurally novel molecules that are distinct from existing drugs[Zhang et al., 2025]. In such campaigns, target-specific screening data may be limited, and publicly available ADMET datasets are relatively small [Huang et al., 2021]. Thus, it is crucial to develop methods that can improve model performance on out-of-distribution data under data-limited conditions. However, Ektefaie et al. [2024] have shown that generalizing molecular property predictors to out-of-distribution (OOD) compounds remains difficult. Similarly, Antoniuk et al. [2025] report large OOD performance drops of molecular property prediction models across standard benchmarks. A central failure mode is shortcut learning: models exploit spurious correlations that work in-distribution but break OOD [Geirhos et al., 2020], a risk exacerbated in small datasets [Hermann et al., 2024].

In chemistry, many properties are driven by specific substructures (e.g., heteroatom placement, hydrogen-bonding motifs) [Li et al., 2025, Rowley, 2008], and shortcuts manifest as attributing activity to irrelevant fragments. Prior approaches have highlighted the causal role of substructure–property relationships, typically through conditioning and association aiming to disentangle causal features from shortcuts. However, these methods rely on substantial amounts of paired data, making them impractical in data-scarce regimes [Lee et al., 2023]. An alternative line of work encourages substructure awareness in prediction via contrastive learning [Wang et al., 2022b,a, Shen et al., 2024, Wan et al., 2025]. Notably, FragCL aligns molecules with all of their fragments in one of its contrastive views [Kim et al., 2023].

We propose MOL-SGCL , a substructure-guided contrastive learning method that leverages large language models to align a molecule's representation towards its plausibly causal substructure and away from its implausible substructure. MOL-SGCL forms anchor–positive–negative triplets where positives are substructures that plausibly explain the molecule's label and negatives are chemically inconsistent with the molecule's label.

To assess the generalizability of the models, we study deliberately hard OOD regimes adapted from the Therapeutic Data Commons (TDC)[Huang et al., 2021], using split strategies that minimize train–test similarity and cap training sizes. We posit that models that are aware of substructure plausibility generalize better than models that are not. We integrate MOL-SGCL into a state-of-the-art Minimol-fingerprint based model [Kläser et al., 2024] and show increased performance across four therapeutics-related tasks. To demonstrate backbone agnosticism, we show that this framework also improves performance on a Chemprop backbone [Heid et al., 2023]. To ensure gains are attributable to plausibility-guided alignment rather than contrastive training, every experiment is run with two controls: (i) the unmodified backbone (no auxiliary contrastive loss), and (ii) a random-substructure contrast variant that matches our triplet count, loss weight, and margin, but selects positives uniformly at random from substructures. Across settings, MOL-SGCL outperforms both controls on OOD splits, though the magnitude of the gain varies by task. This supports that aligning representations with plausible substructure representations allows the model to learn more generalizable features. Because MOL-SGCL only assumes identifiable substructures, MOL-SGCL could be extended beyond molecules to other domains where meaningful subsets of the input may be causally linked to labels.

2 Related Work

Causal Modeling for Molecular Property Prediction Models. A line of work explicitly models causality to learn predictors whose label–feature relationships are invariant under distribution shift [Wu et al., 2022, Sui et al., 2022, Fan et al., 2022]. In the molecular space, Lee et al. [2023] propose CMRL, a conditional–intervention framework that (i) disentangles causal from shortcut components via learned stochastic masks in representation space and (ii) estimates causal effects under backdoor adjustment by conditioning interventions on the paired molecule. This induces invariance to shortcut distributions while preserving context-dependent effects and yields strong benchmark performance. However, CMRL's gains hinge on sufficient paired data to estimate shortcut distributions, and its causal fragments are proxy signals extracted by the model rather than mechanistically specified moieties. In contrast, while we also target features that remain stable under distribution shift, we inject domain knowledge (LLM-derived cues) to identify plausibly causal substructures and directly shape the representation space; an approach that proves effective in data-poor regimes.

Contrastive Learning for Molecular Property Prediction. Contrastive learning has been widely used to improve molecular representations. For instance, Shen et al. [2024] use a triplet loss to more closely represent molecules with similar activity and separate "activity-cliff" pairs (high structural similarity but dissimilar activity) in representation space. Other approaches incorporate external knowledge through contrastive alignment, such as aligning molecules with their knowledge-graph subgraphs [Jiang et al., 2023], encoding heterogeneous multi-view molecular graphs and aligning across views [Chen et al., 2025], or augmenting molecules with knowledge-graph information and aligning original versus augmented graphs [Fang et al., 2023]. Fragment-aware methods instead focus on aligning molecules with their fragments or introducing fragment-level losses. Such methods include aligning molecules with fused fragment representations [Zhang et al., 2023], aligning atom-level and fragment-level graphs [Tang et al., 2025], or combining molecule- and fragment-level objectives [Wang et al., 2022a]. FragCL, in particular, treats a molecule and all of its fragments as positive examples, with fragments from other molecules in the batch serving as negative examples [Kim et al., 2023]. These methods treat all substructures as equally informative for a task. In contrast, our method uses a large language model to select the most salient substructures to represent based on task-specific plausibility. Crucially, we distinguish plausible from implausible substructures within the same molecule, directly encoding domain knowledge into the representation.

3 Mol-SGCL Overview

Mol-SGCL optimizes the sum of the task loss and a substructure-guided triplet loss. Given a mini-batch, we compute the supervised loss $\mathcal{L}_{\text{task}}$ (BCE for classification; MSE for regression). The triplet loss is computed using the molecule representation and representations of substructures. We use an LLM plausibility evaluator to select *plausible* versus *implausible* substructures, and form anchor–positive–negative triplets (anchor = molecule embedding; positive = plausible fragment embedding; negatives= implausible fragment embedding) (Figure 1a). We train the model using a joint objective that combines triplet loss with a supervised loss on the targets. The triplet loss is computed from the representation layer, which takes in the Minimol fingerprint as the input (Figure 1b)

Mol-SGCL Loss We employ a triplet loss formulation for Mol-SGCL . For a given molecule i, let $a_i \in \mathbb{R}^d$ denote its representation in the latent space of the model, $p_i \in \mathbb{R}^d$ the embedding of a mechanistically plausible substructure, and $n_i \in \mathbb{R}^d$ the embedding of an implausible substructure.

We construct the following triplet loss across N_s molecules for which both a plausible and implausible substructure are identified:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \max\left(0, [(1 - \cos(a_i, p_i)) - (1 - \cos(a_i, n_i)) + m]\right),$$

where $\cos(*)$ denotes the cosine similarity, and $m > 0$ is a margin hyperparameter that enforces separation between positive and negative pairs. We train the model with a hybrid objective that balances predictive accuracy with structural alignment. Specifically, we compute mean binary cross-entropy ($\mathcal{L}_{\text{Mean-BCE}}$) loss (classification tasks) or the mean square error ($\mathcal{L}_{\text{MSE}}$) (regression tasks) over the predictions and labels, and then combine it with the triplet loss:

$$\mathcal{L}_{\text{Mol-SGCL}} = \mathcal{L}_{\text{Mean-BCE}} + \alpha \cdot \mathcal{L}_{\text{triplet}} \text{ (classification)}, \quad \mathcal{L}_{\text{Mol-SGCL}} = \mathcal{L}_{\text{MSE}} + \alpha \cdot \mathcal{L}_{\text{triplet}}\text{(regression)}$$

where $\alpha > 0$ is a tunable weight controlling the influence of representation alignment. Notably, $\mathcal{L}_{\text{Mean-BCE}}$ or $\mathcal{L}_{\text{MSE}}$ is calculated over all molecules, whether substructures are identified or not (Figure 1b).

Molecule Selection We aim to align molecular representations with substructures that are plausibly causal for the prediction task. First, we collect molecules for which the triplet loss will be computed. For binary classification, we collect substructures from both positively and negatively labeled molecules. For regression, we specify thresholds to determine which molecules constitute positive and negative examples. In the LLM prompt, we explicitly state the prediction context (e.g., "Lipophilic Molecule" vs. "Hydrophilic Molecule"), which guides the assignment of plausibility labels (Appendix C.2). We note that all molecules are included in the task-space loss calculation (BCE or MSE loss).

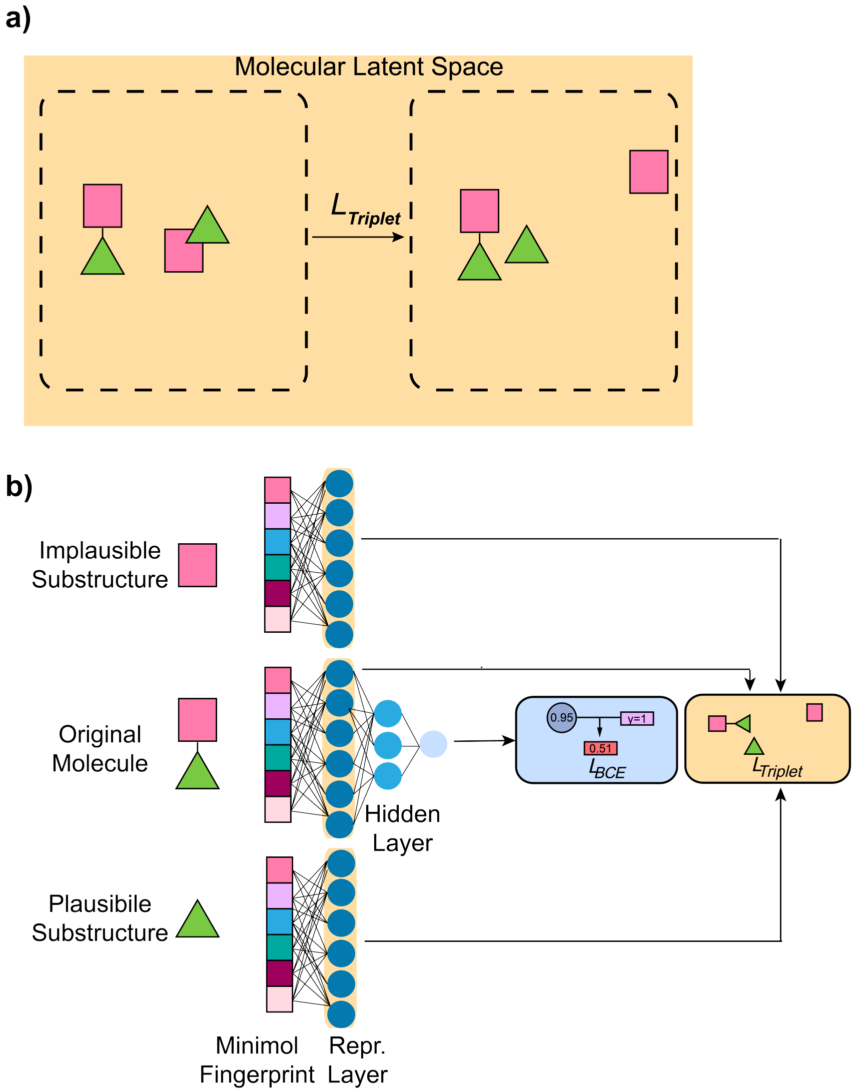

Figure 1: Mol-SGCL Overview. (a) The triplet loss enforces minimal representation space distance between the "plausible" substructure (green triangle) and molecule (green triangle attached to pink square), while maximizing representation distance between the molecule and the "implausible" substructure (pink square). (b) An overview of the Minimol architecture. The triplet loss is calculated from the Representation Layer, and takes in the representation of the original molecule as well as the representation of the substructures. The total loss is the sum of the triplet loss and a supervised loss (MSE for regression, BCE for classification)

Plausibility labeling. We view plausibility labeling as a method to incorporate domain priors into the model. For each molecule x, we aim to identify exactly one causally plausible substructure and one causally implausible substructure. With the natural language description of the molecule, we directly prompt the large language molecule to select one causally plausible substructure for given task, $r^{+}(x)$, and one causally implausible substructure, $r^{-}(x)$. We enforce the size of each substructure to be within 20%-40% of the full molecule's size by heavy-atom count. The full promopt is available in Appendix C.2. The training triplets are thus obtained as:

$$\big(x,\ r^{+}(x),\ r^{-}(x)\big).$$

.

MOL-SGCL uses a large language model, GPT-5 [OpenAI, 2025], to assign plausibility of substructures (See Appendix C.2 for prompt). Since large language models struggle with understanding of SMILES strings[Jang et al., 2025], we provide the textual description of the molecule and each substructure, in addition to the task (Appendix C.1). A summary of the framework for MOL-SGCL is shown in Figure 2.

After substructures are obtained, we check that the substructures are valid molecules and that they are substructures of the input molecule. If either of these checks fail, the substructure inference LLM call is repeated up to three times. The triplet loss is only calculated if valid substructures are obtained; otherwise the molecule's loss function consists of only the target-space loss.

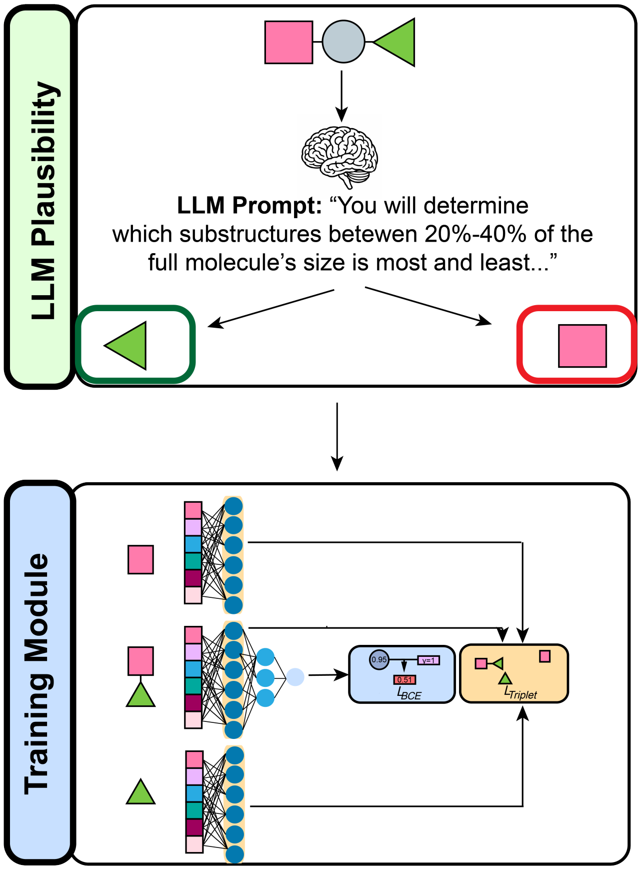
LLM Plausibility
LLM Prompt: "You will determine which substructures betewen 20%-40% of the full molecule's size is most and least..."
Training Module
0.95
y=1
0.51
L_{BCE}
$L_{Triplet}$

Figure 2: MOL-SGCL training framework. Two substructures and their task-specific plausibility labels (one plausible and one implausible) for each molecule are assigned by an LLM-based module. These triplets are taken to the training module, guided by a combination of the triplet and supervised loss.

4 Experiments

4.1 Datasets

We used four ADMET property datasets from the Therapeutic Data Commons [Huang et al., 2021]. Though TDC tasks already are built with scaffold splits, these splits have been shown to sub-optimally separate similar molecules[Guo et al., 2024]. We posited that a model that trained on invariant features would perform better than models that are not when assessed on highly distinct train-test splits, especially when the training dataset is small. To simulate such setting, we created more challenging train-test splits with small training sets and lower cross-split overlap. All train sets had 150 molecules, and the pairwise Tanimoto Similarity to the test set molecules was minimized. We assess the median maximum pairwise Tanimoto-Similarity (MMPTS) between the train and test splits (Morgan Fingerprint), and show that the train-test overlap is substantially lower than the original TDC splits. The validation set was sampled randomly out of the test set (20%). The dataset properties are summarized in Table 1

Table 1: Dataset characteristics. MMPTS denotes the median maximum pairwise Tanimoto similarity between the train and test splits. Val: validation dataset

Dataset	Train Size	Val Size	Test Size	Task Type	MMPTS TDC Splits	MMPTS Our Splits
Lipophilicity	150	810	3240	Regression	0.521	0.262
Free Energy of Solvation	150	98	394	Regression	0.400	0.333
Blood-Brain Barrier	150	377	1512	Classification	0.412	0.250
AMES	150	1425	5703	Classification	0.500	0.268

4.2 MOL-SGCL Can Improve Generalization of Molecular Property Prediction Models

We assessed the performance of MOL-SGCL on these four ADMET tasks. We compare MOL-SGCL to two baselines: the base Minimol model and MOL-SGCL -Random Plausibility. The base model provides a measure of model performance without the triplet loss term, as it is constructed by setting the triplet weight to zero. MOL-SGCL -Random Plausibility has the same hyperparameters as MOL-SGCL , but the plausible and implausible labels are determined randomly. We intend this to be a negative control, as the triplet loss term causes random and unstructured changes to the representation space.

The tasks selected have varying complexity. For instance, lipophilicity is governed by simpler physical processes (i.e. hydrogen bonding between the molecule and water). On the other hand, the AMES assay necessitates synthesizing complex biological and chemical interactions (i.e. molecular response to rat liver enzymes, whether the product is a mutagen to bacteria). We conducted independent hyperparameter sweeps on the validation set (epochs, learning rate, batch norm, hidden layer depth, hidden layer dimensions), and selected the optimal hyperparameters for each method (Appendix A). We trained 20 models with different random seeds and evaluated on the test set. Mean and standard deviation shown in Table 2

Applying MOL-SGCL yields improvements in the OOD test split on the tasks assessed (Table 2). The magnitude of the benefit varies from task-to-task. We speculate that these differences in gains are driven by the LLM labeling. Simpler tasks (i.e. lipophilicity) are more likely to have LLM plausibility labels that match physical reality and thus demonstrate more substantial gains. Tasks with less substantial gains (i.e. AMES) require more complex reasoning over biological systems. We imagine that with a stronger task-specific knowledge base (for instance by using retrieval-augmented generation [Lewis et al., 2020]), and/or quantification of LLM uncertainty, performance on these datasets will improve. Importantly, the substantial degradation in performance with random labels indicates that the improvement arises from meaningful substructure-based alignment, not from the mere presence of a triplet loss.

4.3 MOL-SGCL Improves Generalizability on a D-MPNN Backbone

To ensure the results are not specific to the Minimol architecture, we apply MOL-SGCL to a D-MPNN (Chemprop)[Heid et al., 2023]. MOL-SGCL -Chemprop details can be found in Appendix

Table 2: Performance of MOL-SGCL across tasks on the challenging test sets. Hyperparameters selected on the validation set. Mean ± standard deviation over 20 replicates.

Task (metric)	**Base Minimol**	**MOL-SGCL - Random Plaus.**	**MOL-SGCL - Minimol**
Lipophilicity (RMSE ↓)	0.967 ± 0.009	0.984 ± 0.015	$\mathbf{0.947} \pm 0.011$
FreeSolv (RMSE ↓)	1.923 ± 0.058	1.878 ± 0.078	$\mathbf{1.849} \pm 0.078$
AMES (AUROC ↑)	0.681 ± 0.017	0.646 ± 0.037	$\mathbf{0.690} \pm 0.025$
Blood-Brain Barrier (AUROC ↑)	0.789 ± 0.009	0.780 ± 0.009	$\mathbf{0.796} \pm 0.016$

B. We find that there are similar performance gains on the Chemprop backbone as on the Minimol backbone(Table 3), suggesting the portability of this method to diverse architectures for molecular property prediction.

Table 3: Performance of different methods on the lipophilicity prediction task. Mean ± standard deviation over 5 training replicates.

Method	**Lipophilicity RMSE ↓**
Baseline D-MPNN	1.146 ± 0.018
Random Plausibility	1.168 ± 0.052
MOL-SGCL D-MPNN	$\mathbf{1.066 \pm 0.021}$

5 Conclusion

We introduce MOL-SGCL , a substructure-guided contrastive framework that incorporates plausibility signals into molecular representation learning. By aligning molecules with mechanistically plausible substructures and separating representations from implausible substructures, MOL-SGCL improves performance of molecular property prediction models when evaluated on test sets that are highly dissimilar to the training data. Unlike prior fragment-aware contrastive methods that treat all fragments as equally informative, our approach selectively emphasizes fragments with plausible causal influence as defined by large-language models. This principled distinction between plausible and implausible substructures makes the learned representations more scientifically grounded. We hope that this method will be useful to therapeutics researchers who operate in data-limited regimes and need generalizable molecular property prediction models. Furthermore, MOL-SGCL could be used to incorporate specific domain priors if expert-defined rules are used for substructure labeling instead of LLMs. Beyond molecular property prediction, our approach is general and potentially may be applied to any domain where inputs decompose into parts with differential causal relevance, such as protein motifs or functional groups in materials.

Our study has several limitations. First, the triplet loss requires that both a plausible and an implausible substructure can be identified for each molecule; cases where such pairs cannot be reliably extracted are excluded, which may limit coverage. Second, we demonstrated benefits on custom-defined small data splits with lower train-test overlap because we expected that this is the domain where models that best learn invariant features would excel. Additional work is needed to evaluate MOL-SGCL in larger datasets and datasets where there is more train-test similarity. Third, the performance of MOL-SGCL depends on the accuracy and consistency of language model plausibility labeling. If in a regime where the LLM does not provide useful information, MOL-SGCL may worsen performance. Solutions to quantify LLM uncertainty or improve LLM performance may be employed to alleviate this issue. Despite these caveats, we envision that contrastive methods like these, coupled with strong priors, can flexibly help diverse models learn more generalizable representations.

6 Data and Code Availability

Code and custom data-splits can be found at: `https://github.com/azhou5/MolSGCL_AI4Science` Full datasets are all public and available at `https://tdcommons.ai/`

Acknowledgments and Disclosure of Funding

A.Z. is supported by the Dunleavy Foundation and award Number T32GM144273 from the National Institute of General Medical Sciences. Y.E. is supported by the Eric and Wendy Schmidt Fellowship. The authors gratefully acknowledge National Institute of Allergy and Infectious Diseases / National Institutes of Health grant R01AI155765. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

Evan R. Antoniuk, Shehtab Zaman, Tal Ben-Nun, Peggy Li, James Diffenderfer, Busra Demirci, Obadiah Smolenski, Tim Hsu, Anna M. Hiszpanski, Kenneth Chiu, Bhavya Kailkhura, and Brian Van Essen. Boom: Benchmarking out-of-distribution molecular property predictions of machine learning models. *arXiv preprint arXiv:2505.01912*, 2025. doi: 10.48550/arXiv.2505.01912. URL https://doi.org/10.48550/arXiv.2505.01912.

Mukun Chen, Jia Wu, Shirui Pan, Fu Lin, Bo Du, Xiuwen Gong, and Wenbin Hu. Knowledge-aware contrastive heterogeneous molecular graph learning. *arXiv preprint arXiv:2502.11711*, 2025. doi: 10.48550/arXiv.2502.11711. URL https://arxiv.org/abs/2502.11711.

Yasha Ektefaie, Andrew Shen, Daria Bykova, Maximillian G. Marin, Marinka Zitnik, and Maha Farhat. Evaluating generalizability of artificial intelligence models for molecular datasets. *Nature Machine Intelligence*, 6:1512–1524, 2024. doi: 10.1038/s42256-024-00931-6. URL https://doi.org/10.1038/s42256-024-00931-6.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. In *Advances in Neural Information Processing Systems*, pages 24934–24946. Curran Associates, Inc., 2022.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553, 2023. doi: 10.1038/s42256-023-00654-0.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Qianrong Guo, Saiveth Hernandez-Hernandez, and Pedro J Ballester. Scaffold splits overestimate virtual screening performance. *arXiv preprint arXiv:2406.00873*, 2024. URL https://arxiv.org/abs/2406.00873.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):78–89, 2023. doi: 10.1021/acs.jcim.3c01237. URL https://doi.org/10.1021/acs.jcim.3c01237. Open Access, CC-BY 4.0.

Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228v2*, 2024. doi: 10.48550/arXiv.2310.16228. URL https://doi.org/10.48550/arXiv.2310.16228. Published as a conference paper at ICLR 2024.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021. doi: 10.48550/arXiv.2102.09548. URL https://arxiv.org/abs/2102.09548. Published at NeurIPS 2021 Datasets and Benchmarks.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. Improving chemical understanding of llms via smiles parsing. *arXiv preprint arXiv:2505.16340*, 2025. doi: 10.48550/arXiv.2505.16340. URL https://doi.org/10.48550/arXiv.2505.16340.

Pengcheng Jiang, Cao Xiao, Tianfan Fu, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Bi-level contrastive learning for knowledge-enhanced molecule representations. *arXiv preprint arXiv:2306.01631*, 2023. doi: 10.48550/arXiv.2306.01631. AAAI 2025, version 6 (last revised Feb 16, 2025).

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 4849–4859. PMLR, 2020.

Seojin Kim, Jaehyun Nam, Junsu Kim, Hankook Lee, Sungsoo Ahn, and Jinwoo Shin. Fragment-based multi-view molecular contrastive learning. In *From Molecules to Materials: ICLR 2023 Workshop on Machine Learning for Materials (ML4Materials)*, May 2023. Poster; available on OpenReview.

Kerstin Kläser, Błażej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew Fitzgibbon. Minimol: A parameter-efficient foundation model for molecular learning. *arXiv preprint arXiv:2404.14986*, 2024. doi: 10.48550/arXiv.2404.14986. URL `https://arxiv.org/abs/2404.14986`.

Knowladgator Engineering. Smiles2iupac-canonical-base: Translating smiles chemical names to iupac standards. `https://huggingface.co/knowledgator/SMILES2IUPAC-canonical-base`, 2025. Model based on MT5 with encoder-decoder attention mechanism. License: Apache 2.0.

Namkyeong Lee, Kanghoon Yoon, Gyoung S. Na, Sein Kim, and Chanyoung Park. Shift-robust molecular relational learning with causal substructure. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, page 13, Long Beach, CA, USA, 2023. ACM. doi: 10.1145/3580305.3599437. URL `https://doi.org/10.1145/3580305.3599437`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020. doi: 10.48550/arXiv.2005.11401. URL `https://arxiv.org/abs/2005.11401`.

Jianmin Li, Tian Zhao, Qin Yang, Shijie Du, and Lu Xu. A review of quantitative structure-activity relationship: The development and current status of data sets, molecular descriptors and mathematical models. *Chemometrics and Intelligent Laboratory Systems*, 256:105278, jan 2025. doi: 10.1016/j.chemolab.2024.105278.

Gerald Maggiora. Introduction to molecular similarity and chemical space. In Karina Martinez-Mayorga and José Medina-Franco, editors, *Foodinformatics*, pages 1–23. Springer, Cham, 2014. ISBN 978-3-319-10225-2. doi: 10.1007/978-3-319-10226-9_1. URL `https://doi.org/10.1007/978-3-319-10226-9_1`.

OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. URL `https://openai.com/index/gpt-5-system-card/`. Describes GPT-5's unified model architecture (including gpt-5-main, ...-thinking, etc.), real-time model routing, safe-completions, performance safety enhancements.

Pavel G. Polishchuk, Timur I. Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of Computer-Aided Molecular Design*, 27 (8):675–679, 2013. doi: 10.1007/s10822-013-9672-4. URL `https://doi.org/10.1007/s10822-013-9672-4`.

Michael Rowley. The discovery of raltegravir, an integrase inhibitor for the treatment of hiv infection. *Progress in Medicinal Chemistry*, 46:1–28, 2008. doi: 10.1016/S0079-6468(08)00001-6. URL `https://doi.org/10.1016/S0079-6468(08)00001-6`. Available online March 2008.

Wanxiang Shen, Chao Cui, Xiaorui Su, Zaixi Zhang, Alejandro Velez-Arce, Jianming Wang, Xiangcheng Shi, Yanbing Zhang, Jie Wu, Yu Zong Chen, and Marinka Zitnik. Activity cliff-informed contrastive learning for molecular property prediction. Working Paper, Version 2, posted 06 November 2024; preprint, not peer-reviewed, 2024. URL `https://chemrxiv.org/engage/chemrxiv/article-details/6703d9c351558a15ef5b9e06`.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pages 1696–1705, Washington, DC, USA, 2022. doi: 10.1145/3534678.3539366.

Xiang Tang, Qichang Zhao, Jianxin Wang, and Guihua Duan. Molfcl: predicting molecular properties through chemistry-guided contrastive and prompt learning. *Bioinformatics*, 41(2):btaf061, February 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf061. URL `https://doi.org/10.1093/bioinformatics/btaf061`.

Patrick Walters, W. and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, 54(2):263–270, December 2020. doi: 10.1021/acs.accounts.0c00699.

Yue Wan, Jialu Wu, Tingjun Hou, Chang-Yu Hsieh, and Xiaowei Jia. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nature Communications*, 16:413, 2025. doi: 10.1038/s41467-024-54711-y.

Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *arXiv preprint arXiv:2202.09346*, 2022a.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b. doi: 10.1038/s42256-022-00447-x.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=hGXij5r9fN`.

Kang Zhang, Xin Yang, Yifei Wang, Yunfang Yu, Niu Huang, Gen Li, Xiaokun Li, Joseph C. Wu, and Shengyong Yang. Artificial intelligence in drug development. *Nature Medicine*, 31:45–59, 2025. doi: 10.1038/s41591-024-03165-2. URL `https://doi.org/10.1038/s41591-024-03165-2`.

Ziqiao Zhang, Ailin Xie, Jihong Guan, and Shuigeng Zhou. Molecular property prediction by semantic-invariant contrastive learning. *Bioinformatics*, 39(8):btad462, jul 2023. doi: 10.1093/bioinformatics/btad462.

A Hyperparameters

Hyperparameters specific to MOL-SGCL were preset as follows: The triplet margin was set to $m = 0.2$.

For all Minimol experiments, we performed a grid search on the validation set over the following hyperparameters (two replicates per configuration):

$$\begin{gathered}\text{Learning rate (lr)} \in \{3 \times 10^{-4},\ 2 \times 10^{-3}\},\\ \text{Epochs} \in \{200\ 500\},\\ \text{Hidden size} \in \{512,\ 1024\},\\ \text{Triplet Weight} \in \{10, 100, 400\}\end{gathered}$$

We additionally swept Boolean settings for batch normalization and whether the raw Minimol fingerprint is concatenated to the final layer (as is done in the Minimol implementation) Kläser et al. [2024].

$$\text{BatchNorm} \in \{\text{true},\ \text{false}\}, \qquad \text{Combine fingerprint with final rep.} \in \{\text{true},\ \text{false}\}.$$

Other hyperparameters were held fixed across the sweep:

$$\text{Representation size} = 512, \quad \text{Dropout} = 0.1,$$

We conducted two replicates for each hyperparameter combination, and selected the best set on the validation set (by average performance) for testing. The hyperparameters for the "random" branch were the same as those for the MOL-SGCL branch. For every experiment, except AMES, we sampled collected molecules for the triplet loss from both the "hits" and "non-hits". For AMES, we only collected "hits" due to the the domain knowledge that there are likely not structures that make molecules non-mutagenic; only substructures that make molecules mutagenic.

The validation set was sampled as a 20% split from the test set.

B D-MPNN Mol-SGCL Implementation

For the D-MPNN implementation, the triplet loss was applied in the layer directly after the message passing encoder. For this implementation, we also built a training pipeline that supplemented the BRICS and Murcko fragmentation with a Monte Carlo Tree Search to identify substructures [Jin et al., 2020]. Just as in the Minimol implementation, substructures had a minimum of five heavy atoms and a maximum of eleven. The substructures taken for the contrastive loss were the top two substructures by prediction and two random substructures. Substructures were recollected every five times throughout the training. The summary of the Mol-SGCL -D-MPNN is provided in Figure 3

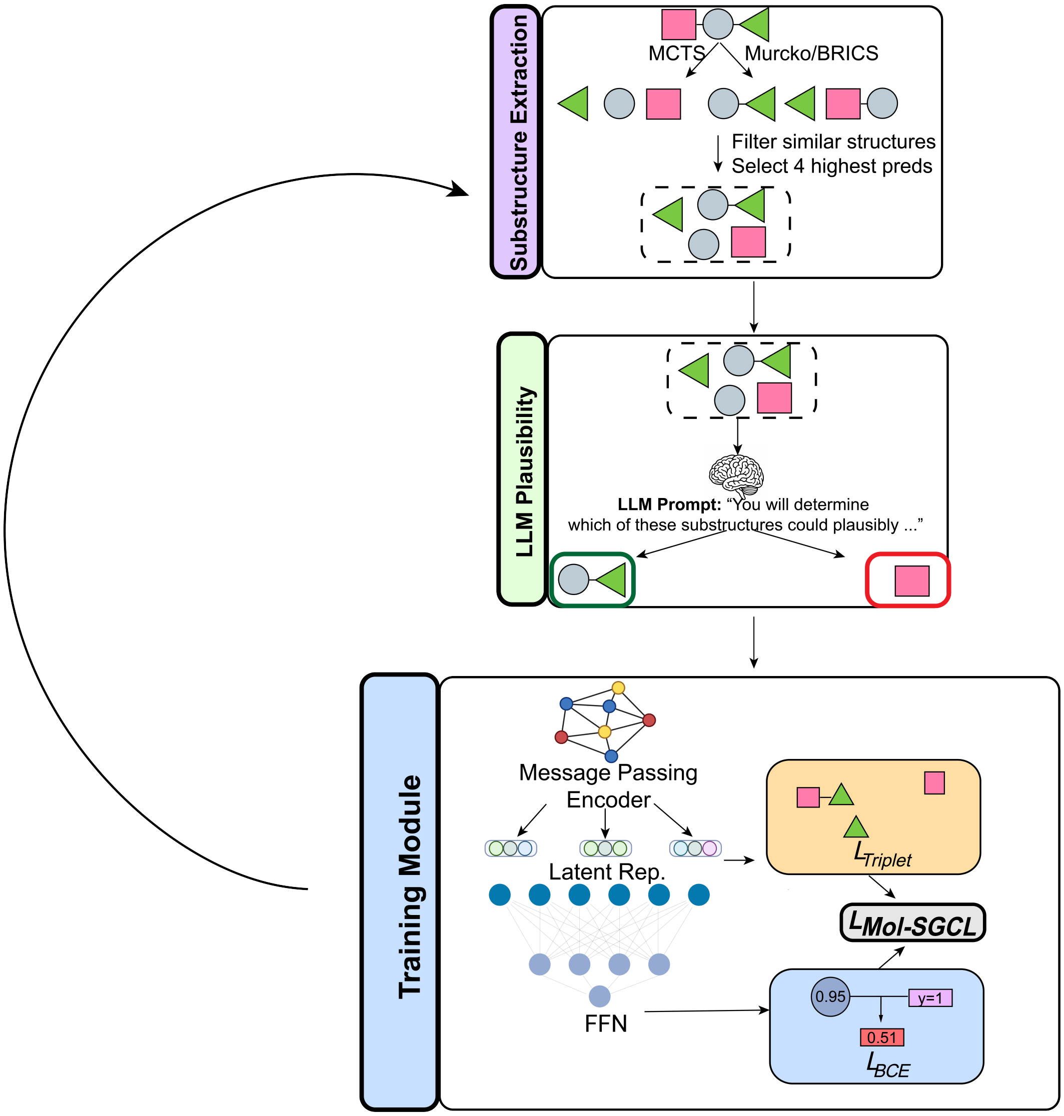

Figure 3: Mol-SGCL -DMPNN overview. The triplet loss is conducted in the layer directly after the message passing encoder. Unlike the Mol-SGCL -Minimol implementation, Mol-SGCL -DMPNN proceeds in cycles of training and substructure collection.

On the validation set, we conducted an abbreviated hyperparameter grid search on learning rate and epochs.

$$\text{Learning rate (lr)} \in \{1 \times 10^{-3},\ 3 \times 10^{-4}\}, \qquad \text{Epochs} \in \{100,\ 200,\ 300\}.$$

All other hyperparameters were the default as implemented in Chemprop v2.1 Heid et al. [2023].

C Prompts

C.1 Molecular Describer

To improve the molecular comprehension of LLMs, we created a function that maps SMILES strings to the corresponding natural language description. We sanitize SMILES with RDKit, remove atom-map numbers and stereochemistry, and canonicalize/kekulize to stabilize the representation. We then generate a candidate IUPAC name using a pretrained converter[Knowladgator Engineering, 2025]. An LLM is prompted with the following template:

"You are an expert medicinal chemist and it is crucial that you get this correct. In one to two sentences, describe the molecule including its core and what relevant side groups it may have. Think carefully about the structure and go step by step internally. I will give you the smiles string.I will also give you an image of the molecule. Lastly, I will give a posssible IUPAC name, which has been generated by a LLM. It may or may not be correct. Use the IUPAC name as a starting point to generate a description of the molecule.Use common names for substituents wheneve possible. Output only the final description.

Example Prompt and output. You are an expert medicinal chemist and it is crucial that you get this correct. In one to two sentences, describe the molecule including its core and what relevant side groups it may have. Think carefully about the structure and go step by step internally. I will give you the smiles string.I will also give you an image of the molecule. Lastly, I will give a posssible IUPAC name, which has been generated by a LLM. It may or may not be correct.Use the IUPAC name as a starting point to generate a description of the molecule. Use common names for substituents whenever possible. Output only the final description IUPAC name: 7-[(4aS,7aS)-1,2,3,4,4a,5,7,7a-octahydropyrrolo[3,4-b]pyridin-6-yl]-1-cyclopropyl-6-fluoro-8-methoxy-4-oxoquinoline-3-carboxylic acid SMILES: "COC1=C2C(=CC(=C1N3C[C@@H]4CCCN[C@@H]4C3)F)C(=O)C(=CN2C5CC5)C(=O)O"

Here is the SMILES: smiles Here is the IUPAC name: iupac

The LLM used was OpenAI GPT-5-mini [OpenAI, 2025].

C.2 Prompts for Plausibility Analysis

The prompt for the plausibility module is specified below: You are an expert medicinal chemist. Given the SMILES string and natural language description of a molecule, you must identify the single most plausible and the single least plausible substructure (as SMILES) that could causally explain the molecule being {task description} Each substructure must be a connected fragment of the molecule containing between 20%-40% of the heavy atoms of the molecule. The minimum number of heavy atoms is 6; override the minimum bound if 20% of the heavy atoms is less than 6. Return ONLY JSON with the keys 'most plausible smiles' and 'least plausible smiles'. The plausible and implausible substructures should always be distinct from each other functionally (i.e. they should have different functional groups, structures, etc). They will be used to guide a triplet loss with one as a positive example and one as a negative example. If the positive and negative examples are very similar, you should return empty strings for both cases. If you are unsure, return empty strings for both values. Return VALID SMILES strings.

Molecule SMILES: {parent smiles} Molecule Description: {parent description or 'N/A'} Task: {task description or 'N/A'} Dataset: {dataset description or 'N/A'}

Please provide: 1. The SMILES of the most plausible substructure (between 20%-40% of the heavy atoms of the molecule), which is most likely to be causally responsible for the molecule being a {task description}. 2. The SMILES of the least plausible substructure (between 20%-40% of the heavy atoms of the molecule), which is least likely to be causally responsible for the molecule being a {task description}.

Output strictly JSON: {"most plausible smiles": "...", "least plausible smiles": "..."}. If you are not confident, return empty strings for both values.

The task description, task description for the negative labeled data, and dataset descriptions are given as follows:

Lipophilicity

- **Task Description:** Lipophilic Molecule
- **Task Description Negative:** Hydrophilic Molecule
- **Dataset Description:** This dataset, curated from ChEMBL database, provides experimental results of octanol/water distribution coefficient (logD at pH 7.4) of 4200 compounds

Free Energy of Solvation

- **Task Description:** Insoluable Molecule
- **Task Description Negative:** Soluable Molecule
- **Dataset Description:** The Free Solvation Database, FreeSolv(SAMPL), provides experimental hydration free energy of small molecules in water. Energies cover a range of approximately 29 kcal/mol, from 3.43 kcal/mol . . . to -25.47 kcal/mol . . .

Ames

- **Task Description:** Mutagen (By AMES Assay)
- **Dataset Description:** The Ames test is a short-term bacterial reverse mutation assay detecting a large number of compounds which can induce genetic damage and frameshift mutations. The dataset is aggregated from four papers. A positive test indicates that the chemical is mutagenic either by itself or when activated with S9 rat liver fraction. The Ames test uses several strains of the bacterium Salmonella typhimurium that carry mutations in genes involved in histidine synthesis.

Blood-Brain Barrier

- **Task Description:** Blood Brain Barrier Permeator
- **Task Description Negative:** Blood Brain Barrier Non-Permeator (Does not Pass)
- **Dataset Description:** The BBB, or blood–brain barrier, is the specialized physiological barrier that separates circulating blood from the brain's extracellular fluid. The positive label in this dataset are molecules that can cross the BBB, while the negative label are molecules that cannot cross the blood brain barrier.'