A Semi-supervised Molecular Learning Framework for Activity Cliff Estimation

Fang Wu

Stanford University & Westlake University

fangwu97@stanford.edu

Abstract

Machine learning (ML) enables accurate and fast molecular property predictions, which is of interest in drug discovery and material design. Their success is based on the principle of similarity at its heart, assuming that similar molecules exhibit close properties. However, activity cliffs challenge this principle, and their presence leads to a sharp decline in the performance of existing ML algorithms, particularly graph-based methods. To overcome this obstacle under a low-data scenario, we propose a novel semi-supervised learning (SSL) method dubbed SemiMol, which employs predictions on numerous unannotated data as pseudo-signals for subsequent training. Specifically, we introduce an additional instructor model to evaluate the accuracy and trustworthiness of proxy labels because existing pseudo-labeling approaches require probabilistic outputs to reveal the model's confidence and fail to be applied in regression tasks. Moreover, we design a self-adaptive curriculum learning algorithm to progressively move the target model toward hard samples at a controllable pace. Extensive experiments on 30 activity cliff datasets demonstrate that SemiMol significantly enhances graph-based ML architectures and outpasses state-of-the-art pretraining and SSL baselines.

1 Introduction

The Similar-Structure, Similar-Property Principle (SSP) [Bender and Glen, 2004] has been a central premise in medicinal chemistry that similar molecules tend to exhibit similar properties, which can be physical (*e.g.*, boiling points) or biological (*e.g.*, activity). This fundamental assertion is validated by long experience suggesting that rules of thumb such as phenethylamines are likely to have activity in the central nervous system and that β-lactams frequently possess antibacterial activity [Martin *et al.*, 2002]. Computational chemists have also exploited this premise in their analysis of the molecular diversity of compound libraries and the selection of compounds for high-throughput screening (HTS) [Johnson *et al.*, 1989].
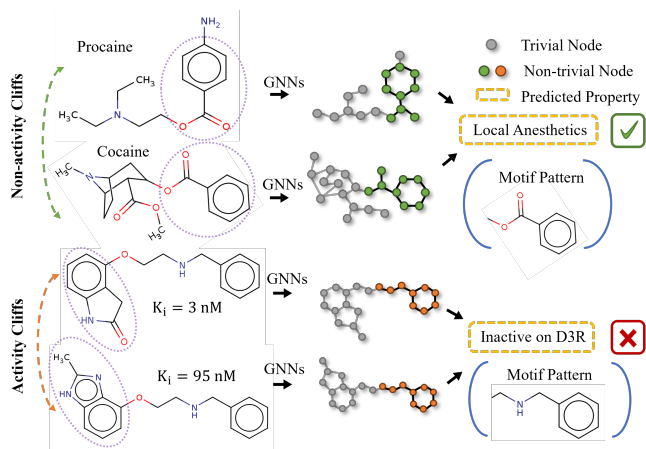

Figure 1: For non-activity cliffs, GNNs successfully discover the motif of `COC(=O)c1ccccc1` and group molecules with this pattern as local anesthetics. For activity cliffs, GNNs are misguided by the shared motif pattern of `COCCNCc1ccccc1` and wrongly predict the inactivity of relevant molecules.

This concept of chemical similarity (or molecular similarity) rationalizes the thriving deployment of machine learning (ML) models in discovering and designing new drugs with expected functions. Noticeably, ML algorithms usually work by recognizing the statistically important patterns and assigning close values to neighbor samples in the high-dimensional feature space [Mitchell and Mitchell, 1997]. They learn motifs that appear repeatedly in the training set, cluster molecules that carry the same motif, and make similar predictions for them [Wu *et al.*, 2023b]. Therefore, SSP lays down a solid scientific foundation for and highly aligns with the inner operation logic of ML mechanisms.

Despite the triumph that ML has achieved in forecasting molecular properties, one particular exception to this principle holds great insights into the underlying structure-activity (or structure-property) relationships [Stumpfe *et al.*, 2020]. This exception is constituted by activity cliffs [Maggiora, 2006], which are pairs of structurally similar molecules that exhibit a large difference in their biological activity. These outliers can impose a detrimental effect on ML models by misguiding them to seriously mispredict the activity of certain molecules, even with an overall high model predictivity

(see Figure 1). Although numerous studies have focused on defining activity cliffs and investigating their influence [van Tilborg *et al.*, 2022], how to ameliorate this obstacle to the development of ML remains unexplored.

In particular, the quantity and quality of the data are one of the most pivot factors in the success of deep learning (DL) algorithms, but the evaluation of DL methods on activity cliff estimation is always in a low-data scenario. Meanwhile, the graph research community for molecule modeling has been trying to replicate the victory of self-supervised pretraining in NLP and claims substantial improvements by pretraining on large-scale datasets [Rong *et al.*, 2020; Zhou *et al.*, 2022]. Consequently, it provokes the question that *"how much benefit self-supervised learning can bring to estimate activity cliffs?"*. As an answer, we observe a negligible reward gained by pretraining in many cases. Our first important finding is that self-supervised graph pretraining does not always have statistically significant advantages over non-pretraining methods with an average improvement ratio of merely 6.35%. Secondly, although several recent studies leverage the 3D spatial information of molecules during the pretraining stages [Fang *et al.*, 2022], the incorporation of 3D geometry does not exhibit considerable advantages in recognizing molecular cliffs. Third, the design of the backbone GNNs has a larger impact on the accuracy of activity cliff estimation than the choice of pretraining mechanisms.

This phenomenon inspires us to explore a more efficient schema to employ the abundant unlabeled molecules and relieve the data scarcity in activity cliff identification. Towards this direction, we present a novel and effective semi-supervised learning (SSL) mechanism dubbed Semi-Mol, which leverages predictions on unannotated data as supervised signals for subsequent training. However, pseudo-labels can be unreliable because labeled and unlabeled samples are usually drawn from different data distributions and DL models often struggle to detect this discrepancy. Besides that, existing pseudo-labeling approaches require probabilistic outputs to reveal the model's confidence and fail to be applied in regression tasks [Rizve *et al.*, 2021; Wu *et al.*, 2023a]. To surmount this barrier, we introduce an additional instructor model to evaluate the accuracy and trustworthiness of proxy labels. These confidence scores then instruct the target molecular model to selectively focus on different data points in a progressive way. In SemiMol, rather than using fixed thresholds or manually tuning thresholds with percentile scores [Cascante-Bonilla *et al.*, 2021], we propose a self-adaptive criterion for curriculum learning to determine which subset of pseudo-labeled samples should be incorporated into the training set. With these delicately designed controllable learning paces, the target molecular model can take advantage of the unlabeled data assuredly without suffering from over-reliance on labeled data and the severe inaccuracy of some pseudo-labels.

We conducted extensive experimentation and visualization on MoleculeACE [van Tilborg *et al.*, 2022] to verify the effectiveness of our method. It can be found that our SemiMol significantly promotes the representation ability of graph-based DL models in low-data scenarios and it outperforms state-of-the-art (SOTA) pretrained GNNs, descriptor-based ML mod-els, and sequenced-based language models as well as prevailing SSL methods on all 30 activity cliff datasets. These results provide encouraging evidence that SemiMol can accurately capture adequate chemical and biological information to recognize cliffs in bioactivity. Our work positions itself in a broader movement within the graph-learning community for bioinformatics and aims to explore the optimal solution for low-data molecular property prediction systematically.

2 Method

2.1 Preliminaries and Background

Task formulation. Assume that there exists a set of accessible molecular data $\mathcal{D}^\star = \{(x_i^\star, y_i^\star)\}_{i=1}^N$, which consists of labeled data points $x_i^\star$. They can be in different formats such as 1D sequences, 2D graphs, and 3D structures, and $y_i^\star$ can be any discrete (*e.g.*, toxicity, and drug reaction) or continuous properties (*e.g.*, water solubility, and free energy). Here we only consider the continuous property to align with the setting of the activity cliff estimation problem. Still, our approach can be easily extended to binary or multi-label circumstances. The target molecular model $f : \mathcal{X} \to \mathcal{Y}$ aims to capture the mapping of molecules and their properties. It can be any type of architecture such as Transformers, GNNs, or geometric neural networks. We also have another set of unseen data $\mathcal{D}^{test}$ to evaluate the model's performance. Typically, $\mathcal{D}^\star$ is divided into the training and validation sets, denoted as $\mathcal{D}^{train} = \{(x_i^{train}, y_i^{train})\}_{i=1}^{N_1}$ and $\mathcal{D}^{val} = \{(x_i^{val}, y_i^{val})\}_{i=1}^{N_2}$, respectively. Furthermore, there is a set of unlabeled data points $\mathcal{D}^* = \{x_i^*\}_{i=1}^M$, where the number of unlabeled data M is much larger than that of labeled data N (*e.g.*, 100 or 10000 times larger).

Semi-supervised learning. SSL has attracted increasing interest in overcoming the need for large annotated datasets. Existing semi-supervised algorithms can be roughly separated into three sorts: consistency regularization, proxy-label methods, and generative models. The consistency regularization is based on the simple concept that randomness within the neural network (*e.g.*, with dropout) or data augmentation transformations should not modify model predictions given the same input and imposes an auxiliary loss. This line of research includes π-model [Laine and Aila, 2016], temporal ensembling [Laine and Aila, 2016], and mean teachers [Tarvainen and Valpola, 2017] The proxy-label methods regard proxy labels on unlabeled data as targets and consist of two groups: self-training [Yarowsky, 1995], where the model itself produces the proxy labels, and multiview learning [Zhao *et al.*, 2017], where proxy labels are produced by models trained on different views of the data. The generative models rely on variational autoencoders (VAE) [Ehsan Abbasnejad *et al.*, 2017] and generative adversarial networks (GAN) [Odena, 2016] to capture the joint distribution $p(y|x)$ more accurately.

2.2 Motivation of SemiMol

Proxy-labeling with confidence. Labeled data $\mathcal{D}^\star$ and unlabeled data $\mathcal{D}^*$ in most cases follow significantly different data distributions, *i.e.*, $\mathbb{P}(x_i^\star, y_i^\star) \neq \mathbb{P}(x_i^*, y_i^*)$. How-
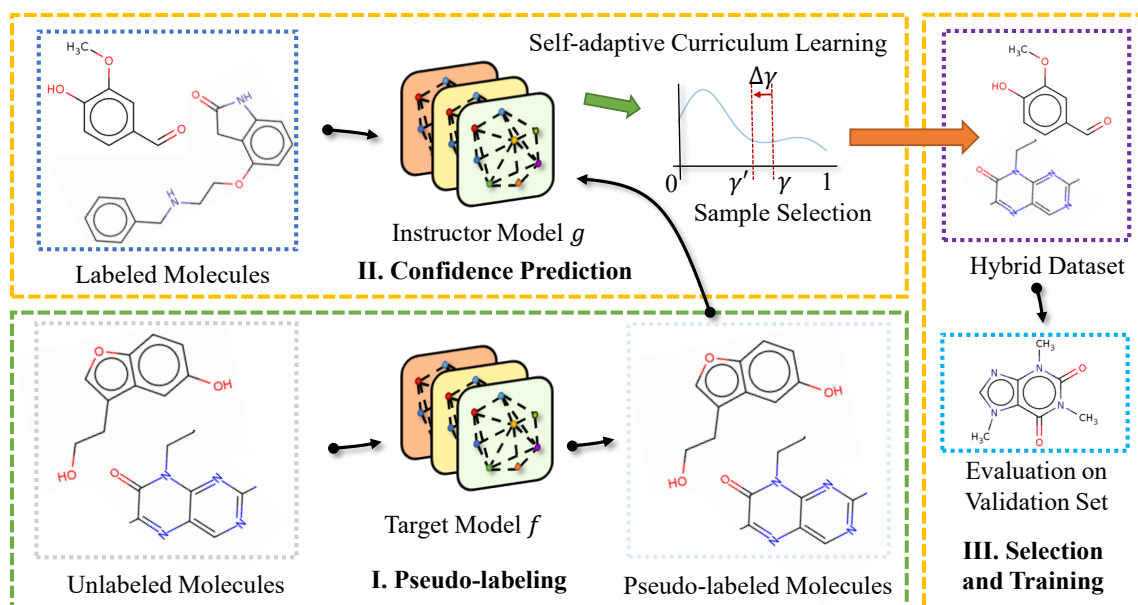
Figure 2: Illustration of SemiMol. The target model first assigns predictions for unlabeled molecular data. Then the instructor model analyzes the confidence of those proxy labels. After that, a self-adaptive curriculum learning schema is adopted to move the target model towards hard samples progressively. By iteratively repeating these processes, the target model can extract task-specific information from vast unannotated molecules to the greatest extent.

ever, on the one hand, many existing semi-supervised algorithms [Yarowsky, 1995; Laine and Aila, 2016; Sohn *et al.*, 2020; Xie *et al.*, 2020] assume that there is no distributional shift between $\mathcal{D}^\star$ and $\mathcal{D}^*$. They are likely only to reinforce the consistent information in the labeled data $\mathcal{D}^\star$ to unlabeled examples $\mathcal{D}^*$ instead of mining auxiliary information from $\mathcal{D}^*$, without exception for proxy labeling [Lee and others, 2013]. On the other hand, despite the versatility and modality-agnostic of proxy labeling, it achieves relatively poor performance compared to recent semi-supervised algorithms [Rizve *et al.*, 2021]. This arises because some pseudo-labels $\hat{y}_i^*$ can be severely incorrect during training due to the poor generalization ability of classic DL models [Hendrycks *et al.*, 2021]. If we directly utilize pseudo-labels that are predicted by a previously learned model for subsequent training, the conformance-biased information in the proceeding epochs could increase confidence in erroneous predictions and eventually lead to a vicious circle of error accumulation [Arazo *et al.*, 2020]. The situation can be even worse when the labeled data $\mathcal{D}^\star$ contain noises due to unavoidable experimental errors. Consequently, it is necessary to understand the quality and reliability of pseudo-annotations and intelligently select a subset to diminish the noise present in training [Rizve *et al.*, 2021].

Confidence for regression tasks. Even though confidence is crucial for pseudo-label selection and confidence-based selection reduces pseudo-label error rates, the poor calibration of neural networks renders this solution insufficient. Explicitly, incorrect predictions in poorly calibrated networks can also have high confidence scores (*i.e.*, $\hat{y}_i^* \to 0$ or $\hat{y}_i^* \to 1$) [Rizve *et al.*, 2021]. More importantly, the majority of existing approaches such as uncertainty-aware pseudo-label se-

lection (UPS) [Rizve *et al.*, 2021] resort to the target model's output $\hat{y}_i$ as the indicator of confidence. They tend to produce hard labels by $y_i^* = \mathbb{1}[\hat{y}_i^* \geq \gamma_1]$ or $y_i^* = \mathbb{1}[\hat{y}_i^* \leq \gamma_2]$, where γ_1 and γ_2 are pre-defined confidence thresholds. However, this selection mechanism becomes inapplicable if $\mathcal{Y}$ is a continuous label space because networks no longer output class probabilities. In other words, for regression tasks, $\hat{y}_i^*$ does not disclose any confidential information. Noticeably, a great number of biochemical problems are regression-based, including molecular property prediction [Wu *et al.*, 2018], 3D structure prediction [Jumper *et al.*, 2021], and binding affinity prediction [Wang *et al.*, 2005]. This presents an unavoidable but essential challenge for probabilistic output-based proxy-labeling algorithms [Rizve *et al.*, 2021].

According to these two motivations, instead of depending on the output of the model $\hat{y}_i^*$ to judge the reliability of the proxy labels, we accompany the target molecular model f with an additional instructor model g. It plays the role of a critic and predicts label observability, *i.e.*, whether the label is true or fake. The introduction of g disentangles the confidence prediction and the property prediction and can greatly reduce the noise introduced by the pseudo-labeling process.

2.3 Framework of SemiMol

Instructor-guided SSL. To circumvent the obstacle of over-reliance and reduce the noise of proxy labels, we present an instructor-guided semi-supervised framework, dubbed SemiMol (see Figure 2), to intelligently select a subset of pseudo-labels with less noise. SemiMol is made up of two components: one target molecular model f and one instructor model g. The former f is responsible for predicting the properties $\hat{y}$. In contrast, the latter g plays a role in mea-

suring the reliability p of supervised signals, which can be interpreted as the confidence probabilities of pseudo-labels. We separate the integral procedure of SemiMol into two phases. In the first step, we retain pseudo-labels $\{\hat{y}_i^*\}_{i=1}^M$. There are several approaches to creating proxy labels, such as label propagation via neighborhood graphs [Iscen *et al.*, 2019]. Here we follow [Lee and others, 2013] and require the molecular model f to directly annotate samples in the unlabeled dataset $\mathcal{D}^*$. Then in the following step, we construct a new dataset with both labeled and pseudo-labeled samples as $\mathcal{D}' = \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*)\}_{i=1}^M$ and proceed training the target molecular model f and the instructor model g based on this new set. These two operations are iteratively repeated until f reaches the optimal performance on the validation set $\mathcal{D}^{val}$.

To be specific, the instructor model $g : (\mathcal{X} \times \mathcal{Y} \times \mathcal{H}_f) \rightarrow \mathcal{P}$ forecasts the confidence measure p_i ($0 \leq p \leq 1$) of whether the given label y_i' belongs to the ground-truth label set $\{y_i^\star\}_{i=1}^N$ or the pseudo-label set $\{y_i^*\}_{i=1}^M$. It digests three items: the data sample with its label $(x_i', y_i') \in \mathcal{D}'$ and an additional loss term $\mathcal{H}_f(f(x_i'), y_i')$, where $\mathcal{H}_f$ is traditionally selected as a root-mean-squared-error (RMSE) loss or a mean absolute error (MAE) loss for regression tasks and cross-entropy (CE) loss for classification problems. Here we regard $\mathcal{H}_f(.)$ as the ingredient of g's input to provide more information about the main molecular property prediction task. Finally, the instructor model g is supervised via a binary CE loss (BCE) as

$$
\begin{aligned}
\mathcal{L}_g\left(\mathcal{D}', \{\hat{y}_i'\}_{i=1}^{N+M}\right) &= \sum_{(x_i', y_i') \in \mathcal{D}'} \mathrm{BCE}(p_i, c_i) \\
&= \sum_{(x_i', y_i') \in \mathcal{D}'} \mathrm{BCE}\Big(g\big(x_i', y_i', \mathcal{H}_f\left(\hat{y}_i', y_i'\right)\big), c_i\Big),
\end{aligned}
\tag{1}
$$

where $c_i \in [0, 1]$ is an integer and represents the observability mask. It indicates whether y_i is pseudo-labeled ($c_i = 0$) or not ($c_i = 1$).

Curriculum-based pseudo-labeling. Meanwhile, the target molecular model f receives discriminative information $\{p_i\}_{i=1}^{N+M}$ from the instructor model g and uses it to efficiently select pseudo-labeled samples to backpropagate its gradient. Remarkably, the criterion leveraged to determine which subset of unlabeled samples $\mathcal{D}^*$ to be incorporated into the training in each round is vital to the success of pseudo-labeling. Various uncertainty metrics have been explored in the previous literature. For instance, several approaches choose instances with the highest-confidence [Zhu, 2005] or retrieve the nearest samples in the feature space [Shi *et al.*, 2018]. Afterward, some studies adopt techniques of fixed thresholds, tuning thresholds manually, or using percentile scores [Cascante-Bonilla *et al.*, 2021]. Here, we propose a self-adaptive curriculum learning algorithm to enable pseudo-annotated samples to enter or leave the new training set.

Specifically, we first initialize a percentage threshold γ and use it to produce hard labels and constitute the hybrid set $\mathcal{D}'' = \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*) \,|\, p_i^* \geq \gamma\}$ to train the target model f.

Then, as the training progresses, we adaptively adjust the confidence cut-off γ based on the model's performance in the validation set. If the model is assessed to be more powerful

---

Algorithm 1 SemiMol Algorithm

1: **Input:** target model f, instructor model g, labeled data $\mathcal{D}^\star$, unlabeled data $\mathcal{D}^*$, pseudo-label update frequency k, threshold percent γ, stepping threshold percent $\Delta\gamma$
2: Initialize and pretrain a target model f_0 and an instructor model g_0
3: **for** epochs $n = 0, 1, 2, ...$ **do**
4:     **if** $n \mod k == 0$ **then**
5:         $\hat{y}_i^* \leftarrow f(x_i^*) \quad \forall x_i^* \in \mathcal{D}^*$ ▷ Iteratively assign pseudo-labels every k epochs
6:     **end if**
7:     $\mathcal{D}' \leftarrow \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*)\}_{i=1}^M$
8:     $\hat{y}_i' \leftarrow f(x_i') \quad \forall x_i' \in \mathcal{D}'$
9:     $p_i \leftarrow g(x_i', y_i', \mathcal{H}_f(.)) \quad \forall (x_i', y_i') \in \mathcal{D}'$ ▷ Predict the confidence scores
10:     $\mathcal{L}_g\left(\mathcal{D}', \{\hat{y}_i'\}_{i=1}^{N+M}\right) \leftarrow$ Equation 1
11:     $\mathcal{D}'' \leftarrow \mathcal{D}^\star \cup \{(x_i^*, \hat{y}_i^*) \,|\, p_i^* \geq \gamma\}$ ▷ Build the training set
12:     $\mathcal{L}_f(\mathcal{D}'') \leftarrow$ Equation 2
13:     Update the parameters of f and g based on $\mathcal{L}_g(.)$ and $\mathcal{L}_f(.)$
14:     Evaluate f on the validation set and attain performance s
15:     **if** $n > 0$ **and** $s < s'$ **then**
16:         $\gamma \leftarrow \gamma - \Delta\gamma$ ▷ Turn down the threshold automatically
17:     **end if**
18:     $s' \leftarrow s$ ▷ Record the validation metric of previous epoch
19: **end for**

(*e.g.,* the evaluation metric like RMSE becomes smaller), we propose to lower the threshold as $\gamma = \gamma - \Delta\gamma$, where $\Delta\gamma$ is the stepping percent. This leads to a larger hybrid set $\mathcal{D}''$ and the loss of the target model f can be written as:

$$
\mathcal{L}_f(\mathcal{D}'') = \sum_{(x_i'', y_i'') \in \mathcal{D}''} \mathcal{H}_f(f(x_i''), y_i''),
\tag{2}
$$

Our self-adaptive curriculum learning strategy discourages concept drift or confirmation bias since it can prevent erroneous annotations predicted by an undertrained network during the early stages of training to be accumulated [Cascante-Bonilla *et al.*, 2021]. The pseudo-code of our SemiMol is depicted in Algorithm 1.

Tips for implementing SemiMol. Before executing SemiMol, it is natural to first obtain a well-trained molecular target model f_0 through regular supervised learning on the labeled dataset $\mathcal{D}^\star$ and then initialize an instructor model g_0 by discriminating pseudo-labels generated by that f_0. This is empirically proven to achieve higher training stability and robustness. Moreover, we choose to assign pseudo-labels every k epochs, where a proper setting of k is critical to the success of SemiMol. If pseudo-labels are updated too frequently, the training procedure tends to be volatile.
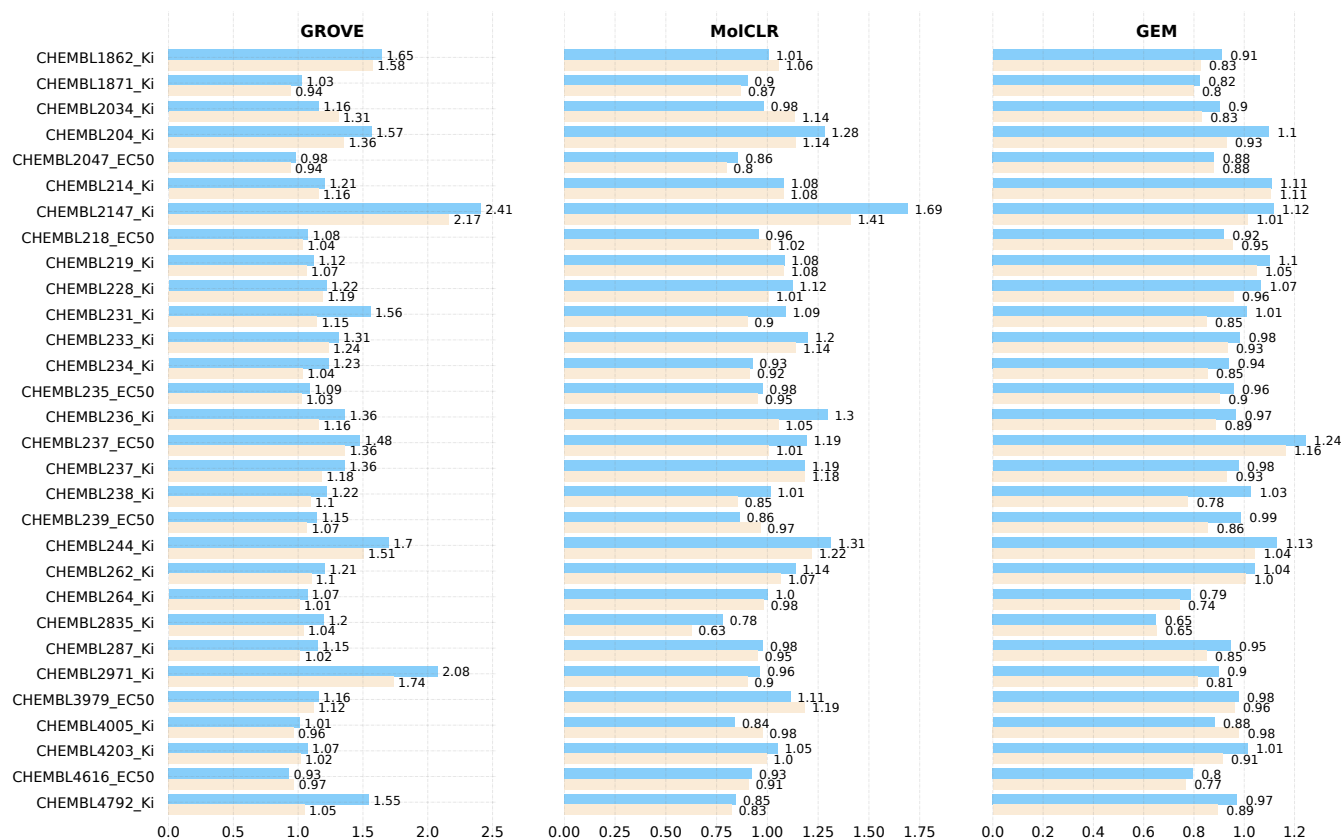
Figure 3: Comparison of different GNNs with and without pretraining on 30 activity cliff datasets. The light blue corresponds to the performance without pretraining, while the antique white represents the performance with pretraining.

3 Experiments

3.1 Data and Experimental Setups

All our evaluations in this section are performed on datasets from MoleculeACE (Activity Cliff Estimation), which is an open-access benchmarking platform and available on Github https://github.com/molML/MoleculeACE. It contains more than 35,000 molecules over 30 macromolecular targets, where each target corresponds to a dataset. In particular, 12 of 30 datasets in MoleculeCAE have no more than 1K molecules in the training set, indicating a standard low-data regime. In addition, we take Graph Isomorphic Network (GIN) [Xu *et al.*, 2018] along with a Graph Multi-set Transformer (GMT) [Baek *et al.*, 2021] to aggregate node features as the backbone. More experimental details and dataset statistics are elaborated on Appendix.

3.2 Does GNN Pretraining Help Mitigate the Activity Cliff Problem?

Data quantity and quality are always regarded as one of the most pivot factors in the success of DL algorithms. However, assessing DL methods on activity cliffs is in a low-data scenario. It is worth noting that the past few years have witnessed the prevailing of self-supervised learning across multiple areas [Devlin *et al.*, 2018; Brown *et al.*, 2020;

Dosovitskiy *et al.*, 2020], and the graph research community in molecule modeling has been trying to replicate its success in natural language processing. Various works attempt to excavate information on large-scale unlabeled molecules [Rong *et al.*, 2020; Fang *et al.*, 2022; Wang *et al.*, 2022; Zhou *et al.*, 2022], and claim significant performance improvements [Rong *et al.*, 2020; You *et al.*, 2020; Zhu *et al.*, 2021; Zhou *et al.*, 2022]. Nevertheless, the efficacy of those pretraining frameworks is only validated on datasets of non-activity cliff molecules and it remains unknown whether they are also powerful in alleviating the data scarcity trouble in activity cliff estimation.

Here, we investigate several SOTA pretraining GNNs, containing GROVE [Rong *et al.*, 2020], MolCLR [Wang *et al.*, 2022], and GEM [Fang *et al.*, 2022], and examine their effectiveness in activity cliff datasets. To be explicit, GROVE adopts a Transformer-style [Vaswani *et al.*, 2017] architecture, significantly enlarging the representation ability and application scope of molecular representation learning schemes. MolCLR applies data augmentation to molecular graphs at both node and graph levels and uses a contrastive learning strategy to generalize GNNs to a more giant chemical space. GEM proposes a geometry-based GNN with dedicated geometry-level self-supervised learning techniques to capture molecular geometry knowledge. Figure 3 exhibits

Figure 4: Performance of various ML models measured by the overall RMSE on 30 activity cliff datasets.

the empirical results. It can be found that pretraining generally benefits the activity cliff estimation problem in all three selected pretraining frameworks. It leads to average improvements of 8.57%, 3.92%, and 6.54% for GROVE, MolCLR, and GEM on all 30 datasets. However, the improvements are not always convincing and can be negligible or even negative. For instance, GROVE performs worse with pretraining in two datasets, MolCLR becomes less competitive with pretraining in six datasets, and GEM is negatively influenced with pretraining in three datasets.

This phenomenon is consistent with the conclusion in one recent study [Sun, 2022] that self-supervised graph pretraining does not always behave statistically more powerfully over non-pretraining methods in many settings. Besides, the local and global molecular 3D structures may not be the key to tackling our problem since GEM has incorporated them for self-supervised learning but shows little benefit. Finally, we compare the performance of different GNNs before and after self-supervised learning. GROVE reaches an average RMSE of 1.3107 and 1.1871 before and after pretraining. MolCLR attains an average RMSE of 1.0556 and 1.0071 before and after pretraining. GEM realizes an average RMSE of 0.9699 and 0.9035 before and after pretraining. The design of GNNs' architecture is a more foundational factor in achieving a better performance in our activity cliff problem. GEM without pretraining can easily outpace pretrained MolCLR and pretrained GROVE.

Why is GNN pretraining not Beneficial? We fail to reproduce the gain of pretraining in conventional molecular datasets like MoleculeNet due to several potential reasons. Firstly, the target labels of most unsupervised pretraining are not appropriately aligned with or even contradictory to our activity cliff estimation task. For instance, GROVE [Rong *et al.*, 2020] requires GNNs to decide whether a molecule contains a motif. However, these 85 sorts of pre-defined motifs might be deceptive in estimating the property of activity cliff molecules. Secondly, as suggested by [Sun, 2022], pretraining is found to typically help when high-quality hand-crafted features are absent. In our experiments, we inject rich features into GNNs, including the additional node features such as hydrogen acceptor match, acidic match, and bond features such as ring information. Last but not least, some pretraining methods, such as masked node label prediction, might be too easy to transfer enough knowledge for our activity cliff estimation task.

3.3 Does SemiMol Improve the Activity Cliff Estimation?

Overall performance. We compare SemiMol with diverse ML and DL algorithms on MoleculeACE and document the results of activity cliff molecules and all molecules in Figure 4 and Appendix, separately. It can be observed that our SemiMol outperforms other baselines with the lowest RMSE on 30 activity cliff datasets. More essentially, we discover an average improvement of 26.53% brought by SemiMol, which is significantly larger than the benefits of SOTA pretraining approaches. This phenomenon declares that GNNs empowered by SSL can be a more influential toolkit than existing algorithms to overcome the obstacle of activity cliffs and accelerate the process of new drug discovery.

Among all approaches, non-pretrained GNNs, including MPNN [Gilmer *et al.*, 2017], GAT [Veličković *et al.*, 2017], GCN [Kipf and Welling, 2016], and AFP [Baek *et al.*, 2021], generally perform worse than other methods. Meanwhile, the performance of pretrained GNNs varies dramatically. For instance, GEM outweighs GROVE by a large margin. However, as analyzed in Section 3.2, the benefit of pretraining is proven limited. This difference is more related to the capability of the backbone GNNs than to the pretraining strategies. Ad-

Datasets	CHEMBL1862_Ki	CHEMBL1871_Ki	CHEMBL2034_Ki	CHEMBL204_Ki	CHEMBL2047_EC50
GIN	1.019	0.854	0.951	1.179	0.895
+ π-model	0.984	0.793	0.902	1.060	0.880
+ Semi-GAN	0.955	0.781	0.834	0.927	0.854
+ SemiMol	**0.714**	**0.671**	**0.703**	**0.801**	**0.629**

Table 1: Performance of different semi-supervised learning algorithms on 5 datasets of MoleculeACE.

Datasets	CHEMBL214_Ki	CHEMBL2147_Ki	CHEMBL218_EC50	CHEMBL219_Ki	CHEMBL228_Ki
SemiMol					
+ Fixed Threshold	0.788	0.704	0.781	0.914	0.821
+ Percentile Scores	0.739	0.683	0.744	0.850	0.779
+ Self-adaptive Paces	**0.726**	**0.672**	**0.723**	**0.835**	**0.758**

Table 2: Comparison with different curriculum learning methods on 5 datasets of MoleculeACE.

ditionally, sequence-based models such as LSTM [Hochreiter and Schmidhuber, 1997] and Transformer [Vaswani *et al.*, 2017] are conspicuous for their outstanding performance. They are empirically a better choice for addressing the activity cliff problem than graph-based and descriptor-based models if no additional unlabeled data are provided in the fine-tuning stage.

Comparison with other semi-supervised methods. Diverse SSL algorithms have been invented to employ the abundant unlabeled resources in a task-specific manner. Approximately, they can be divided into three categories: consistency regularization, proxy-label methods, and generative models. Here, we pick up one baseline instance from each kind and investigate their improvements for molecular property prediction. Notably, random node dropping and edge perturbation are biochemically improper, and data augmentation may alter the semantics of molecular graphs [Sun *et al.*, 2021]. Thus, we employ the π-model [Laine and Aila, 2016] for comparison, which applies different stochastic transformations (*i.e.*, dropout) to the networks instead of the input graphs. Moreover, we select Semi-GAN [Odena, 2016], which introduces a discriminator to classify whether the input is labeled or not. As mentioned previously, UPS [Rizve *et al.*, 2021] leverages the prediction uncertainty to guide the pseudo-label selection procedure and is therefore inapplicable in our situation. We select five datasets from MoleculeACE for comparison and use GIN as the backbone architecture. As shown in Table 1, SemiMol outperforms all other semi-supervised methods in overall RMSE. In addition to that, SemiMol overcomes the drawbacks of UPS and can be used for regression tasks. All the above evidence clarifies that SemiMol is a cutting-edge proxy labeling algorithm in comparison to existing SSL mechanisms with stronger generalization and broader applications.

Ablation study. We also implement an extra experiment to illustrate the effectiveness of our self-adaptive curriculum learning mechanism and report the results in Table 2. It can be found that our self-adaptive schedule achieves stronger performance in overall RMSE than a fixed threshold and percentile score-based curriculum learning paradigm [Cascante-Bonilla *et al.*, 2021].

Performance for classification tasks. Notably, datasets in MoleculeACE are all regression tasks, and one major advan-

tage of our SemiMol over other proxy-labeling algorithms is that SemiMol can deal with continuous property space. In spite of that, SemiMol can also be employed for classification tasks. In order to better demonstrate its superiority, we use a classification dataset, CYP3A4 [Rao *et al.*, 2022], and examine its performance. CYP3A4 consists of the property cliffs in the Cytochrome P450 3A4 inhibitions experimentally measured by [Veith *et al.*, 2009], including 3,626 active inhibitors/substrates and 5,496 inactive compounds. We report the corresponding performance of different approaches in Table 3, where the best performance is in bold and the second best is underlined. It can be observed that SemiMol outperforms all semi-supervised baselines, which rely on MPNN as the backbone architecture.

Models	CYP3A4	
	ROC-AUC	Cliff ROC-AUC
GCN	0.7659	0.6801
GAT	0.7734	0.6828
MPNN	0.7811	0.6867
π-model	0.7842	0.6903
Semi-GAN	0.8015	0.7122
UPS	<u>0.8438</u>	<u>0.7546</u>
SemiMol	**0.8568**	**0.7703**

Table 3: Performance in the classification dataset, CYP3A4.

4 Conclusion

The past decade has witnessed the increasing employment of machine learning techniques for drug discovery, but the topic of activity cliffs (AC) receives little interest from the scientific community. In this work, we identify the best practice to enhance models' predictivity in the presence of activity cliffs and propose an effective semi-supervised learning (SSL) method named SemiMol. It introduces an instructor model to understand the confidence of proxy-label and exploits a self-adaptive curriculum training paradigm to move the target model toward hard samples more efficiently. We demonstrate that graph-based DL models can hold up against simpler ML algorithms for drug discovery even in low-data scenarios. We envision more efficient and better-suited SSL methods to propel the frontier of AC identification.

References

[Arazo *et al.*, 2020] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Baek *et al.*, 2021] Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. *arXiv preprint arXiv:2102.11533*, 2021.

[Bender and Glen, 2004] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218, 2004.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Cascante-Bonilla *et al.*, 2021] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6912–6920, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Ehsan Abbasnejad *et al.*, 2017] M Ehsan Abbasnejad, Anthony Dick, and Anton van den Hengel. Infinite variational autoencoder for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2017.

[Fang *et al.*, 2022] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

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

[Hendrycks *et al.*, 2021] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Iscen *et al.*, 2019] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079, 2019.

[Johnson *et al.*, 1989] M Johnson, M Lajiness, and G Maggiora. Molecular similarity: a basis for designing drug screening programs. *Progress in clinical and biological research*, 291:167–171, 1989.

[Jumper *et al.*, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Laine and Aila, 2016] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[Maggiora, 2006] Gerald M Maggiora. On outliers and activity cliffs why qsar often disappoints, 2006.

[Martin *et al.*, 2002] Yvonne C Martin, James L Kofron, and Linda M Traphagen. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

[Mitchell and Mitchell, 1997] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[Odena, 2016] Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.

[Rao *et al.*, 2022] Jiahua Rao, Shuangjia Zheng, Yutong Lu, and Yuedong Yang. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns*, 3(12):100628, 2022.

[Rizve *et al.*, 2021] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

[Rong *et al.*, 2020] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

[Shi *et al.*, 2018] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315, 2018.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[Stumpfe *et al.*, 2020] Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath. Advances in exploring activity cliffs. *Journal of Computer-Aided Molecular Design*, 34:929–942, 2020.

[Sun *et al.*, 2021] Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. Mocl: Data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3585–3594, 2021.

[Sun, 2022] Ruoxi Sun. Does gnn pretraining help molecular representation? *arXiv preprint arXiv:2207.06010*, 2022.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[van Tilborg *et al.*, 2022] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Veith *et al.*, 2009] Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P Austin, David G Lloyd, et al. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. *Nature biotechnology*, 27(11):1050–1055, 2009.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[Wang *et al.*, 2005] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

[Wang *et al.*, 2022] 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, 2022.

[Wu *et al.*, 2018] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[Wu *et al.*, 2023a] Fang Wu, Huiling Qin, Wenhao Gao, Siyuan Li, Connor W Coley, Stan Z Li, Xianyuan Zhan, and Jinbo Xu. Instructbio: A large-scale semi-supervised learning paradigm for biochemical problems. *arXiv preprint arXiv:2304.03906*, 2023.

[Wu *et al.*, 2023b] Fang Wu, Dragomir Radev, and Stan Z Li. Molformer: motif-based transformer on 3d heterogeneous molecular graphs. *Rn*, 1:1, 2023.

[Xie *et al.*, 2020] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[Yarowsky, 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When does self-supervision help graph convolutional networks? In *international conference on machine learning*, pages 10871–10880. PMLR, 2020.

[Zhao *et al.*, 2017] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[Zhou *et al.*, 2022] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. *International Conference on Learning Representations*, 2022.

[Zhu *et al.*, 2021] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, pages 2069–2080, 2021.

[Zhu, 2005] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.