3D Denoisers are Good 2D Teachers: Molecular Pretraining via Denoising and Cross-Modal Distillation

Sungjun Cho[1], Dae-Woong Jeong[2], Sung Moon Ko[2], Jinwoo Kim[3], Sehui Han[2], Seunghoon Hong[3], Honglak Lee[2], Moontae Lee[2,4]

[1]University of Wisconsin-Madison
[2]LG AI Research
[3]KAIST
[4]University of Illinois Chicago

sungjuncho@cs.wisc.edu, dw.jeong@lgresearch.ai, sungmoon.ko@lgresearch.ai, jinwoo-kim@kaist.ac.kr, hansse.han@lgresearch.ai, seunghoon.hong@kaist.ac.kr, honglak@lgresearch.ai, moontae.lee@lgresearch.ai

Abstract

Pretraining molecular representations from large unlabeled data is essential for molecular property prediction due to the high cost of obtaining ground-truth labels. While there exist various 2D graph-based molecular pretraining approaches, these methods struggle to show statistically significant gains in predictive performance. Recent work have thus instead proposed 3D conformer-based pretraining under the task of denoising, leading to promising results. During downstream finetuning, however, models trained with 3D conformers require accurate atom-coordinates of previously unseen molecules, which are computationally expensive to acquire at scale. In this paper, we propose a simple solution of denoise-and-distill (D&D), a self-supervised molecular representation learning method that pretrains a 2D graph encoder by distilling representations from a 3D denoiser. With denoising followed by cross-modal knowledge distillation, our approach enjoys use of knowledge obtained from denoising as well as painless application to downstream tasks with no access to 3D conformers. Experiments on real-world molecular property prediction datasets show that the graph encoder trained via D&D can infer 3D information based on the 2D graph and shows superior performance and label-efficiency against previous methods.

Introduction

Molecular property prediction has gained much interest across the machine learning community, leading to breakthroughs in various applications such as drug discovery (Guvench 2016; Kanakaveti et al. 2017) and material design (Suh et al. 2020; Pyzer-Knapp, Li, and Aspuru-Guzik 2015; Schmidt et al. 2019; Pyzer-Knapp et al. 2022). As molecules can be represented as a *2D graph* with nodes and edges representing atoms and covalent bonds, many graph neural networks have been developed with promising results (Duvenaud et al. 2015; Defferrard, Bresson, and Vandergheynst 2016; Bruna et al. 2013; Coley et al. 2017; Scarselli et al. 2009; Jin et al. 2018). However, achieving high precision requires accurate ground-truth property labels which are very expensive to obtain. This limitation has motivated adaptation of self-supervised pretraining widely used in natural language processing (Devlin et al. 2018; Brown et al. 2020) and computer vision (He et al. 2020; Bachmann et al. 2022) onto molecular graphs with proxy objectives developed to instill useful knowledge into neural networks with unlabeled data. But existing 2D graph-based pretraining frameworks face a fundamental challenge: while the model is trained to learn representations that are invariant under various data augmentations, augmenting 2D graphs can catastrophically disrupt its topology, which renders the model unable to fully recover labels from augmented samples (Trivedi et al. 2022). As a result of this limitation, recent work has shown that existing 2D pretraining approaches do not show statistically meaningful performance improvements in downstream tasks (Sun 2022).

As an alternative, recent work have proposed incorporating 3D information to the pretraining objective, leveraging large unlabeled datasets of *3D conformers*, or point clouds of atoms floating in the physical space. While a natural task would be to reconstruct the input conformer, this may not induce generalizable knowledge as each conformer only represents a single local minima in a distribution of 3D configurations. On the other hand, the force field that controls the overall stabilization process provides significant chemical information that can be used across many different molecular properties (Mezey 2001). This naturally translates to pretraining via denoising conformers under perturbations, an approach that has shown state-of-the-art performance in diverse molecular property prediction benchmarks (Zaidi et al. 2022; Liu, Guo, and Tang 2022).

Despite great performance, a model trained with denoising requires conformers downstream as well, and obtaining accurate conformers require expensive quantum mechanical computations. While there exist many rule-based (Riniker and Landrum 2015; Landrum 2016) as well as deep learning-based approaches (Ganea et al. 2021; Xu et al. 2022; Jing et al. 2022) for generating conformers, previous work have shown that existing methods fail to generate conformers quickly and accurately enough to be used in a large scale (Stärk et al. 2022).

In light of such limitations, we propose D&D (Denoise-and-Distill), a self-supervised molecular representation learning framework that enjoys the best of both worlds. Fig-

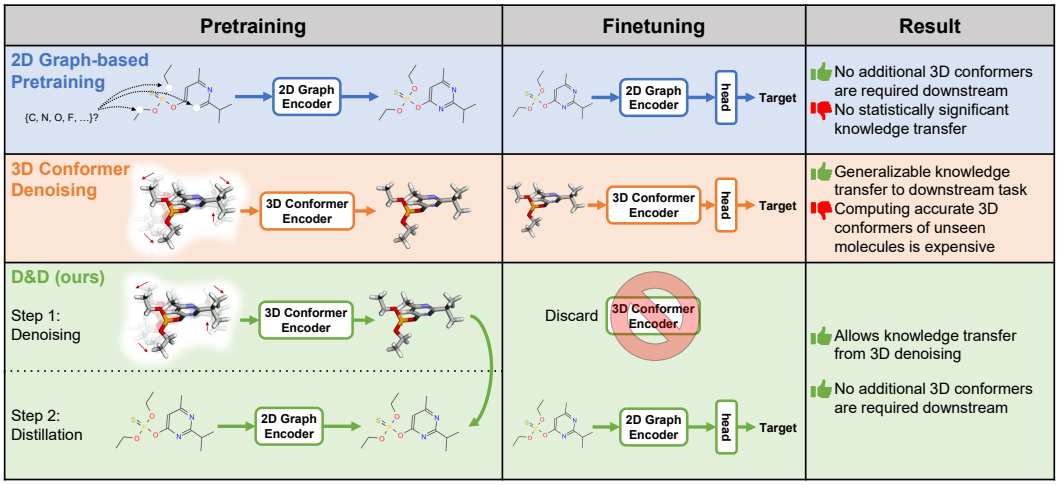


Figure 1: Comparison between D&D and existing molecular pretraining frameworks. **Top:** 2D graph-based pretraining methods fail to bring significant benefit to downstream molecular property prediction. **Middle:** 3D denoising is effective in predicting molecular properties by approximately learning the force field in the physical space, but cannot be easily applied to downstream tasks where only 2D graphs are available. **Bottom:** Our method D&D allows practitioners to leverage knowledge from 3D denoising in downstream scenarios where only 2D molecular graphs are available without the need to generate 3D conformer via expensive computations or machine learning approaches.

ure 1 shows the overall pipeline of our work. D&D sequentially performs two steps: 1) we pretrain a 3D teacher model that denoises conformers artificially perturbed with Gaussian noise and 2) freeze the 3D teacher encoder and distill representations from the 3D teacher onto the 2D student. When given a downstream task with access to 2D molecular graphs only, the 3D teacher is discarded and the 2D student is finetuned towards the given task. As a result of distillation, D&D encourages the 2D graph encoder to exploit the topology of the molecular graph towards encoding the input molecule similarly to the 3D conformer encoder without any explicit supervision from property labels. Surprisingly, experiments on various molecular property prediction datasets indicate that the 2D graph representations from D&D can generalize to unseen molecules. To the best of our knowledge, our method is the first self-supervised molecular representation learning framework that adopts cross-modal knowledge distillation to transfer knowledge from a 3D denoiser onto a 2D graph encoder. We summarize our main contributions as follows:

- We propose D&D, a two-step self-supervised molecular representation pretraining framework that performs 3D-to-2D cross-modal distillation.
- Pretraining results show that under D&D, the 2D student model can closely mimic representations from the 3D teacher model using graph features and topology. Further analysis shows that the intermediate representations of the 2D student also aligns well with 3D geometry.
- Experiments on the MoleculeNet benchmark and curated molecular property regression datasets show that D&D leads to significant knowledge transfer, and also performs well in downstream scenarios where the number of labeled training data points is limited.

Related Work

In this section, we first discuss previous work on knowledge distillation that inspired our approach. We also cover existing self-supervised pretraining approaches for molecular representation learning.

Knowledge Distillation. Knowledge distillation (KD) was developed under the motivation of transferring knowledge learned by a large *teacher* model to a much more compact *student* model, thereby reducing the computational burden while preserving the predictive performance (Hinton et al. 2015). Example approaches in computer vision include distilling class probabilities as a soft target for classification models (Ba and Caruana 2014) or transferring intermediate representations of input images (Tian, Krishnan, and Isola 2019). For dense prediction tasks such as semantic segmentation, it has been shown that a *structured* KD approach that distills pixel-level features instead leads to improvements in performance (Liu et al. 2020). Another extension that is more closely related to our approach is *cross-modal* KD on unlabeled modality-paired data (e.g. RGB and Depth images), which was proposed to cope with modalities with limited data (Gupta, Hoffman, and Malik 2016). Inspired by this work, D&D performs 3D-to-2D cross-modal KD to allow downstream finetuning on 2D molecular graphs while utilizing the feature space refined by 3D conformer denoising. Further information on KD can be found in a recent survey by Gou et al. (2021).

Pretraining for Molecular Property Prediction. Inspired by previous work in the NLP domain, there exist many self-supervised pretraining approaches for learning representations of molecular graphs. Similar to masked token prediction in BERT (Devlin et al. 2018), Hu et al. (2019)

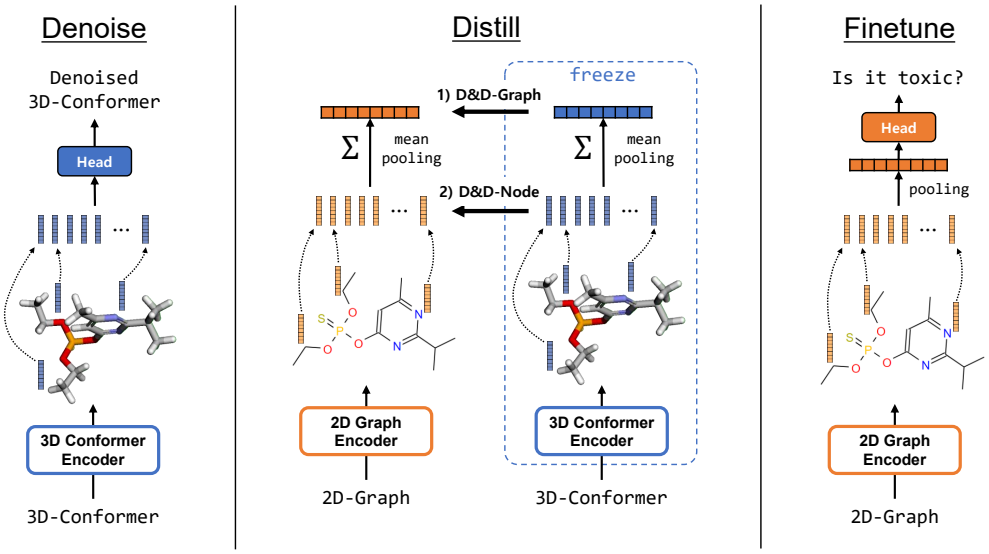


Figure 2: Illustration of our D&D framework. First we pretrain a 3D conformer encoding module by denoising perturbed conformers. Next we pretrain a 2D graph encoder by distilling representations from the 3D teacher. We propose two variants: D&D-GRAPH distills mean-pooled graph representations while D&D-NODE distills node representations in a more fine-grained manner. During finetuning, we tune the 2D graph encoder only with the given downstream data.

proposed node-attribute masking and context prediction to reconstruct topological structures or predict attributes of masked nodes. GROVER (Rong et al. 2020) proposed predicting what motifs exist in the molecular graph, under the insight that functional groups determine molecular properties. Contrastive approaches were also proposed, in which the task is to align representations from two augmentations of the same molecule and repel representations of different ones (Hassani and Khasahmadi 2020; You et al. 2020). Despite promising results, it has been shown that obtaining significant gains in performance with existing 2D pretraining methods is non-trivial, as empirical improvements rely heavily on the choice of hyperparameters and other experimental setups (Sun 2022).

As molecules lie in the 3D physical space, some work have deviated from the 2D graph setting and instead proposed 3D pretraining via denoising conformers perturbed with Gaussian noise, from which empirical results have shown significant knowledge transfer to diverse property prediction tasks (Zaidi et al. 2022; Liu, Guo, and Tang 2022). Despite great downstream performance, such 3D approaches necessitate access to accurate 3D conformers of molecules under concern, which are difficult to obtain as it requires expensive quantum mechanical computations such as density functional theory (DFT) (Parr and Weitao 1995).

There exist solutions to avoid these drawbacks. 3DInfomax (Stärk et al. 2022) proposed a cross-modal contrastive pretraining framework that aligns 2D and 3D representations. In addition to contrastive learning, GraphMVP (Liu et al. 2021) also incorporates generative pretraining which trains the model to reconstruct the 2D graph representation from its 3D counterpart, and vice versa. Similar to GraphMVP, MoleculeSDE (Liu et al. 2023) pretrains representations using two group-symmetric stochastic differential equation-based generative processes, which led to state-of-the-art downstream performance. While all aforementioned methods use 3D conformers during pretraining only, they either do not capture information from molecular force fields which we conjecture to be helpful for forward knowledge transfer, or involve extensive tuning on loss weighting and data augmentation due to simultaneously optimizing contrastive and generative objectives. In contrast, our D&D framework is remarkably simple due to the disentangled two-step pretraining procedure with each step optimizing a single objective, yet enjoys generalizable knowledge obtained through conformer denoising.

Preliminaries

We first introduce preliminary information on learning representations of 2D molecular graphs and 3D molecular conformers alongside notations that we use in later sections.

2D Molecular Graphs. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the 2D molecular graph with N atoms represented as nodes in $\mathcal{V}$, and M bonds represented as edges in $\mathcal{E}$. In addition to the graph-connectivity, each node is assigned features based on chemical attributes such as atomic number and aromaticity, and similarly for each edge with features based on bond type and stereo configurations. Given the graph $\mathcal{G}$, a 2D graph encoder f^{2D} typically first returns representations for each node:

$$f^{\text{2D}}(\mathcal{G}) = \boldsymbol{Z}^{\text{2D}} \text{ where } \boldsymbol{Z}^{\text{2D}} \in \mathbb{R}^{N \times d}. \quad (1)$$

In molecular property prediction settings we need a single representation for each molecular graph. Typical operators used to extract graph-level representations include mean-pooling all node representations or adding a virtual node to the input graph and treating its representation as the graph representation (Hamilton 2020).

3D Molecular Conformers. Each molecule can also be represented as a 3D conformer $\mathcal{C} = (\mathcal{V}, \boldsymbol{R})$ with 3D spatial coordinates of each atom stored in $\boldsymbol{R} \in \mathbb{R}^{N\times 3}$. Note that unlike 2D graphs, 3D conformers have no information of the graph connectivity via edges in $\mathcal{E}$, and are instead treated as point cloud data. As with 2D graphs, let f^{3D} denote the 3D conformer encoder that takes the conformer $\mathcal{C}$ and returns representations of each atom:

$$f^{\text{3D}}(\mathcal{C}) = \boldsymbol{Z}^{\text{3D}} \text{ where } \boldsymbol{Z}^{\text{3D}} \in \mathbb{R}^{N\times d}. \tag{2}$$

Note that how the molecule is oriented on the 3D Euclidean space naturally does not affect its chemical property. Thus, the encoder f_{3D} must return representations that are invariant under rotations and translations on $\boldsymbol{R}$ (i.e. $f^{\text{3D}}((\mathcal{V}, \boldsymbol{R})) = f^{\text{3D}}((\mathcal{V}, g(\boldsymbol{R}))$ for $g \in \text{SE(3)}$) for efficient weight-tying across SE(3) roto-translations. Note that because molecular properties can change under chiral orientations, we only respect rotations and translations, but not reflections. There exist many architectures that respect SE(3) symmetry as an inductive bias (Fuchs et al. 2020; Bronstein et al. 2021; Gasteiger, Becker, and Günnemann 2021; Satorras, Hoogeboom, and Welling 2021; Thölke and De Fabritiis 2022), and any such architecture can be used for f^{3D}.

D&D: Denoise and Distill

Here we describe D&D, a molecular pretraining framework that transfers generalizable knowledge from 3D conformer denoising to a 2D graph encoder via cross-modal distillation, thereby allowing painless downstream applications without computing accurate conformers of unseen graphs. The two major steps are as follows: 1) Denoising perturbed conformers with a 3D conformer encoder f^{3D}, and 2) Distilling representations from the 3D teacher to the 2D graph encoder f^{2D}. An illustration of the overall pipeline can be found in Figure 2. As our first step of D&D is based upon previous work on conformer denoising (Zaidi et al. 2022; Liu et al. 2022), we provide a brief outline of the task and refer readers to corresponding papers for further details and theoretical implications.

Step 1: Pretraining via denoising. Given a stabilized ground-truth conformer $\mathcal{C} = (\mathcal{V}, \boldsymbol{R})$, f^{3D} is given as input a perturbed version of the same conformer $\tilde{C} = (\mathcal{V}, \tilde{\boldsymbol{R}})$, produced by slightly perturbing the coordinates of each atom with Gaussian noise as

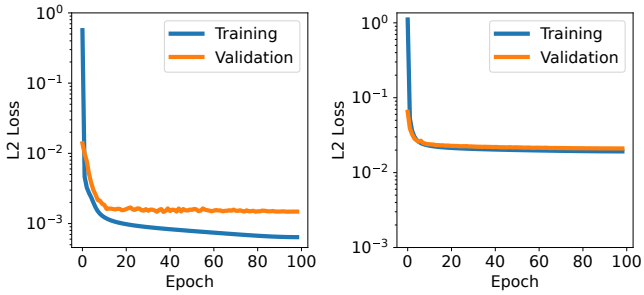
$$\tilde{\boldsymbol{R}}_i = \boldsymbol{R}_i + \sigma\boldsymbol{\epsilon}_i \text{ where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{I}_3) \tag{3}$$

with noise scale σ as hyperparameter. Then, we attach a prediction head $h^{\text{3D}} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times 3}$ to the f^{3D} such that the combined model outputs 3-dimensional vectors per atom.

$$h^{\text{3D}}(f^{\text{3D}}(\tilde{\mathcal{C}})) = (\hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_N) \tag{4}$$

Lastly, the model is trained to predict the noise that has been injected to create $\tilde{\mathcal{C}}$ from $\mathcal{C}$. The denoising loss minimized during training is as follows:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{p(\tilde{\mathcal{C}},\mathcal{C})}\left[\left\|h^{\text{3D}}(f^{\text{3D}}(\tilde{\mathcal{C}})) - (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)\right\|_2^2\right] \tag{5}$$



(a) D&D-GRAPH (b) D&D-NODE

Figure 3: Training and validation loss curves (in log-scale) during distillation of D&D-GRAPH and D&D-NODE on PCQM4Mv2. The 2D student is able to closely distill features from the 3D teacher with small generalization gap.

where $p(\tilde{\mathcal{C}}, \mathcal{C})$ denotes the probability distribution induced by the data distribution and the noise sampling procedure to create $\tilde{\mathcal{C}}$. Surprisingly, the denoising objective is equivalent to learning an approximation of the force field in the physical space derived by replacing the true distribution of conformers with a mixture of Gaussians (Zaidi et al. 2022). The Gaussian mixture potential corresponds to the classical harmonic oscillator potential in physics, a great approximation scheme for linearized equations such as denoising.

For our experiments, we use the TorchMD-NET (Thölke and De Fabritiis 2022) architecture for f^{3D} following (Zaidi et al. 2022) due to its equivariance to SE(3) roto-translations and high performance on quantum mechanical property prediction. Note that D&D is architecture-agnostic, and any other SE(3)-equivariant architecture can be used as well.

Step 2: Cross-modal distillation. After pretraining via denoising is done, we distill representations from the pretrained f^{3D} to a 2D graph encoder model f^{2D}. We consider two different variants of cross-modal KD, leading to two respective variants of our approach. For the first variant D&D-GRAPH, we minimize the difference between graph representations from 2D and 3D encoders:

$$\mathcal{L}_{\text{distill-graph}} = \left\|\text{pool}(f^{\text{2D}}(\mathcal{G})) - \text{pool}(f^{\text{3D}}(\mathcal{C}))\right\|_2^2 \tag{6}$$

During training, we freeze the teacher model f^{3D} and flow gradients only through the student model f^{2D}. This effectively trains the 2D encoder to leverage the bond features and graph topology to imitate representations from 3D conformers. To obtain graph representations, we average all node representations inferred by each encoder.

Inspired by structured KD (Liu et al. 2020), we propose another variant D&D-NODE that distills node-level representations without any pooling:

$$\mathcal{L}_{\text{distill-node}} = \left\|f^{\text{2D}}(\mathcal{G}) - f^{\text{3D}}(\mathcal{C})\right\|_2^2 \tag{7}$$

Unlike D&D-GRAPH, D&D-NODE makes full use of the one-to-one correspondence between atoms in the molecular graph and atoms in the conformer. Hence f^{2D} is trained to align towards representations from f^{3D} in a more fine-grained manner.

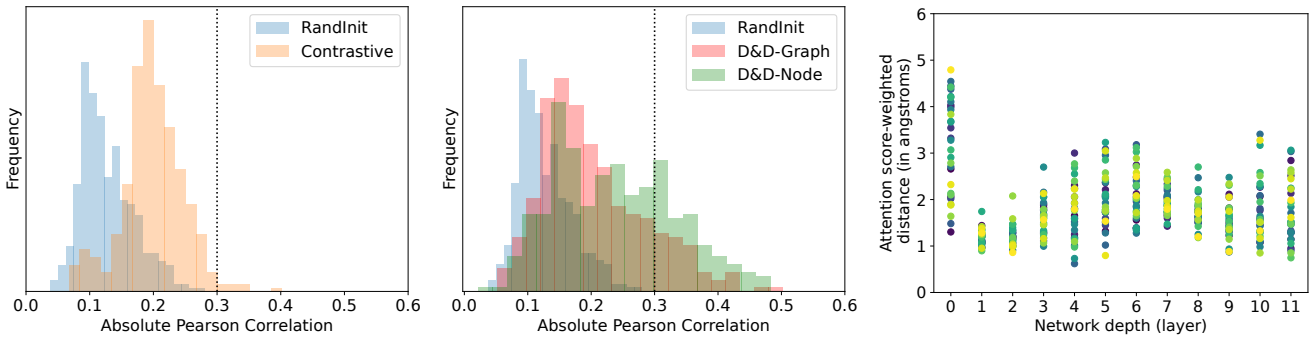


Figure 4: Histograms of Pearson correlation values between pre-softmax attention scores vs. 3D pairwise distance during inference on the PCQM4Mv2 validation set for (Left) CONTRASTIVE, (Middle) D&D-GRAPH and D&D-NODE. (Right) Average attention score-weighted 3D distances according to network depth from D&D-NODE. Each colored dot represents an attention head in the corresponding layer.

For f^{2D}, we mainly use the Tokenized Graph Transformer (TokenGT (Kim et al. 2022a)) architecture that theoretically enjoys maximal expressiveness across all possible permutation-equivariant operators on 2D graphs. Due to this flexibility, we expect f^{2D} to be trained to align representations from f^{3D} as closely as possible, by which we hope to see the effect of distillation to the fullest extent. Furthermore, using an attention-based architecture also allows analysis on the relationship between the attention scores of atom-pairs and their physical distance in the 3D space. Results from which are discussed later in the following section. Note that similarly with f^{3D}, however, any other permutation-equivariant graph neural network architecture can be adopted seamlessly.

Downstream finetuning. Assuming the downstream task does not provide accurate conformers as input, we discard f^{3D} after the distillation step and finetune f^{2D} with molecular graphs only. We use L1 loss and BCE loss for regression and binary-classification tasks, respectively, following previous work (Stärk et al. 2022).

Note that we finetune the entire f^{2D} model instead of just the newly attached prediction head on the downstream data. Given that the force fields induced by electron clouds provide knowledge that is generalizable to various molecular properties, we conjecture that D&D provides a good initial point in the parameter space from which finetuning f^{2D} entirely leads to a better local optima. This also aligns with previous observations in NLP that pretrained language models outperform models trained from scratch only when the entire model is finetuned (Rothermel et al. 2021).

Experiments

For empirical evaluation, we test our D&D pretraining pipeline on various molecular property prediction tasks using open-source benchmarks as well as four manually curated datasets. We also stress-test D&D under a downstream scenario where the number of labeled data points is extremely limited. All experiments are run in a remote GCP server equipped with 16 NVIDIA A100 Tensor Core GPUs.

Experimental Setup

Datasets. For pretraining, we use PCQM4Mv2 (Nakata and Shimazaki 2017), a large molecule dataset consisted of 3.7M molecules. Each molecule is paired with a single 3D conformer at the lowest-energy state computed via DFT. In case of D&D, we use the same PCQM4Mv2 dataset for both denoising and disillation steps. Note that even though PCQM4Mv2 provides the HOMO-LUMO energy gap of each molecule as labels, we do not use any supervision from such labels during training, and instead treat the dataset as a collection of unlabeled molecular graph-conformer pairs.

For finetuning, we use 10 datasets from MoleculeNet (Wu et al. 2018), three of which are regression tasks and the rest are binary classification tasks. We use scaffold splitting to obtain the train-test splits for each task. As shown in the appendix, the MoleculeNet datasets exhibit different molecule distributions from PCQM4Mv2: some tasks involve atom types that the encoder has never observed during pretraining. As most datasets in MoleculeNet involve classification, we also curate four molecular property regression datasets from open-source databases: CCS measures the collisional cross sections of molecules (Kim et al. 2022b), CML and CMQ measures the emissive chromophore life times and fluorescence quantum yields, respectively (Joung et al. 2020). SLE measures the solid-liquid phase change entropy (I et al. 2011). The property values of these datasets are normalized by mean and standard deviation before training and evaluation. For curated datasets, we randomly split the dataset 8:1:1 for training, validation, and testing. Further details on the datasets can be found in the appendix.

Baselines. For MoleculeNet experiments, we compare our method against baselines that pretrain representations using only 2D graph topology: ATTRMASK and CONTEXTPRED (Hu et al. 2019) masks out and predicts key node and edge attributes or surrounding graph substructures, GRAPHCL (You et al. 2020) performs contrastive learning with graph augmentations, JOAO (You et al. 2021) and JOAOv2 (You et al. 2021) also perform contrastive pretraining, but adaptively chooses which graph augmentation to use during training. We also compare against base-

Pretraining	Bace(↑)	BBBP(↑)	ClinTox(↑)	HIV(↑)	Sider(↑)	Tox21(↑)	ToxCast(↑)	Esol(↓)	Freesolv(↓)	Lipo(↓)
GRAPH ISOMORPHISM NETWORK (GIN)										
NO PRETRAIN	78.88±1.79	68.68±1.07	71.97±8.73	77.01±0.55	55.95±0.64	75.06±0.40	61.46±0.21	3.64±0.05	3.76±0.18	0.66±0.02
ATTRMASK	79.10±1.39	68.21±0.55	74.06±3.43	74.03±1.29	55.86±0.77	73.48±0.64	61.02±0.23	2.83±0.30	4.46±0.37	0.68±0.01
CONTEXTPRED	72.91±4.59	65.21±1.14	54.72±2.18	71.89±1.28	55.90±0.84	70.84±0.53	59.39±0.47	3.72±0.08	5.04±0.08	0.81±0.07
GRAPHCL	74.92±3.18	64.00±1.18	81.66±2.27	76.92±0.76	56.64±0.90	75.05±0.39	62.44±0.50	3.18±0.03	4.86±0.33	0.66±0.01
JOAO	74.24±2.50	65.76±1.19	84.35±2.33	77.03±1.05	58.19±0.67	73.71±0.37	62.50±0.33	2.97±0.06	4.74±0.15	0.66±0.01
JOAOV2	76.72±3.64	65.81±1.01	**84.95±3.06**	77.44±0.54	57.23±0.24	73.33±0.26	62.60±0.33	3.34±0.09	4.59±0.41	0.65±0.01
GRAPHMVP	**81.30±1.46**	67.92±0.85	66.30±1.32	75.69±0.77	59.46±0.50	73.24±0.16	62.80±0.35	3.27±0.07	4.33±0.16	0.627±0.01
MOLECULESDE	78.15±4.10	67.37±1.23	77.16±1.58	76.47±0.79	**60.22±0.55**	75.38±0.59	63.22±0.35	2.08±0.19	**3.64±0.17**	**0.610±0.00**
D&D (OURS)	79.44±0.22	**69.92±1.18**	84.81±5.40	**78.02±0.58**	59.28±0.14	**75.81±0.31**	**63.57±0.25**	**1.24±0.16**	4.18±0.23	0.66±0.01
TOKENIZED GRAPH TRANSFORMER (TOKENGT)										
NO PRETRAIN	77.53±1.97	65.94±0.77	85.44±2.17	70.44±1.86	57.65±2.13	72.34±0.48	60.63±0.62	0.811±0.05	1.624±0.14	0.71±0.01
GRAPHMVP	69.12±0.84	66.17±1.47	81.45±2.50	68.17±0.86	58.13±0.56	74.23±0.42	62.79±0.54	0.84±0.02	1.62±0.08	0.66±0.02
MOLECULESDE	74.33±3.14	**68.17±0.99**	**87.65±1.50**	72.21±0.42	**63.50±1.37**	74.55±0.48	63.79±0.42	**0.67±0.03**	1.53±0.08	0.61±0.02
D&D (OURS)	**81.11±2.65**	66.40±1.71	81.51±2.63	**77.31±1.03**	63.00±1.11	**76.61±1.04**	**65.21±0.71**	0.70±0.02	**1.38±0.12**	**0.52±0.01**

Table 1: MoleculeNet Results averaged across 5 random seeds with one standard deviation. The modality of each method indicates whether the method uses 2D graphs only (*2D only*) or 3D conformers as well (*2D&3D*) during pretraining. The first 7 tasks show ROC-AUC (higher is better) while the remaining three show results in MAE (lower is better). We also report average rank across all targets for each GNN architecture. Best results for each task within each architecture are marked **bold**.

lines that use both 2D and 3D modalities, GRAPHMVP and MOLECULESDE, discussed above as related work. As most previous work have used the graph isomorphism network (GIN (Xu et al. 2018)) as the 2D encoder, we also test D&D using GIN in addition to TOKENGT. For the remaining tasks, we mainly compare D&D against MoleculeSDE using the TokenGT backbone, as other methods have not shown significant performance gains compared to randomly initialized models without pretraining.

During finetuning, we compute graph representations via mean-pooling of all node representations within each molecule, and feed it to a linear adapter. We test both node-wise and graph-wise distillation, and report the better of the two results. For consistency, we follow the same featurization step provided by the OGB library (Hu et al. 2020) across all tasks, which produces a 9 and 3-dimensional feature vector for each atom and bond, respectively. Further details such as hyperparameterization can be found in the appendix.

Pretraining Results

Prior to downstream evaluation, we discuss interesting findings from pretraining with D&D.

The 2D graph encoder can closely mimic representations from 3D conformers using only the molecular graph. Figure 3 shows the training and validation loss curves of the distillation step of D&D. When pretraining with D&D-NODE, the distillation loss converges to slightly over 10^{-2} with a very small generalization gap between validation and training. This shows that the 2D molecular graph contains enough information to closely imitate representations from the 3D teacher f^{3D}. The small gap between training and validation also reflects that the guidance provided via D&D-NODE can well-generalize towards unseen molecules. When pretraining with D&D-GRAPH, we find that the training loss converges to a much lower optima of 10^{-3}, but with a much larger generalization gap of approximately 10^{-3}. This implies that while the task of distilling mean-pooled representations is easier than distilling node-wise representations, it leads to less generalizable knowledge due to not considering the graph topological structure.

The intermediate encoding procedure of the 2D encoder trained via D&D aligns with 3D geometry. As we use an attention-based architecture for f^{2D} encoding, we qualitatively assess how well 2D atom-wise interactions modeled via attention resembles its interactions in 3D geometry without actual conformers (e.g. do atoms nearby in the 3D space tend to attend to each other?). Specifically, we compute the absolute Pearson correlation between the 3D pairwise distances of atoms and the inner product of their features prior to the softmax layer in each attention head, averaged across all molecules in the PCQM4Mv2 validation set. Note that a larger inner product implies a relatively larger exchange of information between the two atoms. The first two figures in Figure 4 show histograms depicting distributions of averaged absolute Pearson correlation values from all attention heads for each layer in the 2D encoder after pretraining by each method. We find that using a contrastive objective only leads to a slight increase in correlation compared to random initialization: most correlation values are distributed under 0.3. When pretrained with our D&D-variants, however, many attention heads show correlations that exceed 0.3, a value that is never reached with randomly initialized weights. This implies that our approach provides guidance to the 2D graph encoder towards processing molecular graphs while respecting its 3D geometry. For further investigation, we also measure the average pairwise distances weighted by the attention scores from D&D-NODE with results shown in the rightmost plot in Figure 4. A higher value indicates that the attention head tends to exchange information across atoms that are far apart. Interestingly, the first layer exhibits a diverse range of distances, but the layer that immediately follows uses attention mostly to exchange information across atoms that are geometrically nearby each other, similar to a SE(3)-convolutional layer. Considering

that a carbon-carbon single bond has an average length of 1.5 angstroms, this result indicates that f^{2D} pretrained with D&D-NODE can reason about 3D geometry to exchange information across atoms that are nearby in the 3D conformer, even though they may be far apart in the 2D graph.

Finetuning Results

Here we provide empirical observations from various downstream molecular property prediction datasets.

D&D transfers knowledge that is generalizable across diverse tasks. Table 1 shows finetuning results on MoleculeNet, in which each experiment is averaged across 5 randomly seeded runs. Comparing results from D&D and NO PRETRAIN, D&D shows superior performance in 9 out of 10 tasks, with an average performance improvement of 4.6% and 18.6% across classification and regression tasks, respectively. In particular, we find that the tasks on which D&D shows the largest performance gain against NO PRETRAIN coincide with properties known to align well with 3D geometric properties (65.9% for ESOL with GIN, 27.0% for LIPO with TokenGT). Both ESOL and LIPO tasks are tightly associated with the overall polarity of electron clouds in the molecule, which is highly associated with spatial atom-wise positions. This implies that denoising followed by distillation effectively transfers spatial 3D knowledge. With respect to 2D graph-based self-supervised pretraining baselines, D&D outperforms all methods on 8 out of 10 task using GIN as the backbone, which again demonstrates the effectiveness of using 3D ground-state conformers in addition to 2D graph topologies for pretraining. We also find that 2D-only baselines generally suffer from negative knowledge transfer, which aligns with previous findings (Sun 2022).

When comparing D&D against other methods that use cross-modal objectives, D&D shows 9.7% and 3.2% better performances overall than GRAPHMVP and MOLECULESDE, respectively. This suggests that focusing on transferring molecular-specific knowledge of force fields without any additional tasks is more effective downstream than having a contrastive objective as a proxy task; attracting and repelling molecular representations across the two modalities fail to fully capture generalizable similarities and discrepancies in the chemical space. Another limitation of the contrastive approach is that the gain in performance becomes limited when only a single conformer is provided per molecule (Liu et al. 2021; Stärk et al. 2022). This aligns well with our intuition that each conformer can be seen as a distribution of 3D configurations, and that learning a single local optima within the distribution does not provide much information. Meanwhile, D&D can learn and transfer knowledge of the overall distribution with denoising and distilling, relieving practitioners from the need to obtain multiple low-energy conformers per molecule for pretraining.

D&D enables label-efficient finetuning. To evaluate D&D under downstream tasks with limited labels, we perform finetuning on four curated datasets using the full downstream data set as well as a smaller randomly sampled subset of the original training data for each task. Note that label-efficiency is crucial for molecular property prediction espe-

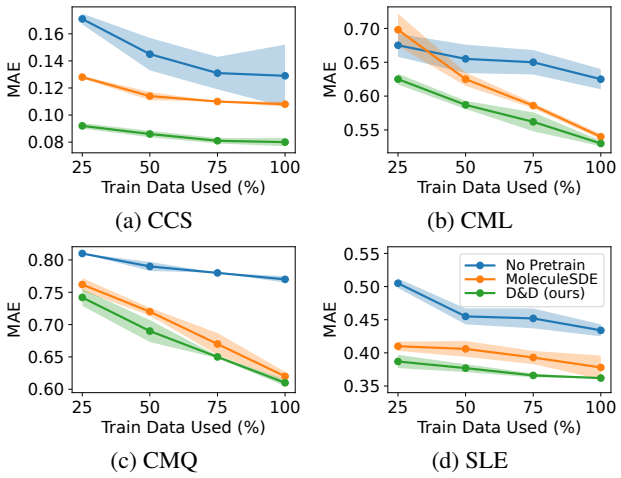


Figure 5: Results on the four curated datasets, averaged across 3 seeds with one standard deviation shown in the shaded area. The X-axis indicates the percentage of finetuning data used, and the Y-axis shows test MAE performances.

cially since gathering refined labels often require extensive validation through costly wet-lab experiments. For this experiment, we mainly use TOKENGT as our backbone architecture and compare against MOLECULESDE to focus on the downstream effect of knowledge transfer solely from denoising to the fullest extent.

Figure 5 shows finetuning results on four curated regression datasets. In all four tasks, D&D trained using only 25% of training data outperforms the model without pretraining, which shows that knowledge of conformer denoising serves as a great initialization for property prediction that prevents overfitting to small training data and induces high generalizability to unseen molecules. D&D also consistently outperforms MOLECULESDE across all targets and train set sizes, lowering the MAE by at most 26.6% on predicting collisional cross sections (CCS). This is particularly interesting as CCS measures the probability of two or more particles colliding to each other, a quantity that requires accurate 3D structural information of molecules to predict accurately. Based on this insight, we hope to explore other molecular properties that are highly related to the 3D structure of ground-state conformers and apply D&D as future work.

Concluding Remarks

We propose D&D, a self-supervised molecular representation learning method that distills features from a 3D conformer denoiser to a 2D graph encoder. Learning knowledge of force fields that provide chemically generalizable information, D&D does not require 3D conformers downstream and is effective on diverse molecular property prediction tasks. D&D also enjoys high label-efficiency, achieving high performance with limited downstream labeled data. As future work, we hope to extend D&D to multitask setups (Liu et al. 2020) and molecular diffusion models (Xu et al. 2022; Jing et al. 2022) for property prediction.

References

Ba, J.; and Caruana, R. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27.

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv preprint arXiv:2204.01678*.

Bronstein, M. M.; Bruna, J.; Cohen, T.; and Veličković, P. 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral Networks and Locally Connected Networks on Graphs.

Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; and Jensen, K. F. 2017. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling*, 57(8): 1757–1772. PMID: 28696688.

Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

Fuchs, F.; Worrall, D.; Fischer, V.; and Welling, M. 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33: 1970–1981.

Ganea, O.; Pattanaik, L.; Coley, C.; Barzilay, R.; Jensen, K.; Green, W.; and Jaakkola, T. 2021. Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems*, 34: 13757–13769.

Gasteiger, J.; Becker, F.; and Günnemann, S. 2021. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

Guvench, O. 2016. Computational functional group mapping for drug discovery. *Drug Discovery Today*, 21(12): 1928–1931.

Hamilton, W. 2020. *Graph Representation Learning*. Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool Publishers. ISBN 9781681739632.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, 4116–4126. PMLR.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

I, S.; PANDEY, A.; Novotarskyi, S.; Koerner, R.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; VV, P.; Tanchuk, V.; Todeschini, R.; Varnek, A.; Marcou, G.; P, E.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Baskin, I.; VA, P.; and Tetko, I. 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of Cheminformatics*, 3.

Jin, W.; Yang, K.; Barzilay, R.; and Jaakkola, T. 2018. Learning Multimodal Graph-to-Graph Translation for Molecular Optimization.

Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; and Jaakkola, T. 2022. Torsional Diffusion for Molecular Conformer Generation. *arXiv preprint arXiv:2206.01729*.

Joung, J. F.; Han, M.; Jeong, M.; and Park, S. 2020. DB for chromophore.

Kanakaveti, V.; Sakthivel, R.; Rayala, S. K.; and Gromiha, M. M. 2017. Importance of functional groups in predicting the activity of small molecule inhibitors for Bcl-2 and Bcl-xL. *Chemical Biology & Drug Design*, 90(2): 308–316.

Kim, J.; Nguyen, T. D.; Min, S.; Cho, S.; Lee, M.; Lee, H.; and Hong, S. 2022a. Pure transformers are powerful graph learners. *arXiv preprint arXiv:2207.02505*.

Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; and Bolton, E. E. 2022b. PubChem 2023 update. *Nucleic Acids Research*, 51(D1): D1373–D1380.

Landrum, G. 2016. RDKit: Open-Source Cheminformatics Software.

Liu, S.; Du, W.; Ma, Z.-M.; Guo, H.; and Tang, J. 2023. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *International Conference on Machine Learning*, 21497–21526. PMLR.

Liu, S.; Guo, H.; and Tang, J. 2022. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*.

Liu, S.; Qu, M.; Zhang, Z.; Cai, H.; and Tang, J. 2022. Structured multi-task learning for molecular property prediction. In *International Conference on Artificial Intelligence and Statistics*, 8906–8920. PMLR.

Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; and Tang, J. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Liu, Y.; Shu, C.; Wang, J.; and Shen, C. 2020. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*.

Mezey, P. G. 2001. Distributions and averages of molecular conformations. *Computers & Chemistry*, 25(1): 69–75.

Nakata, M.; and Shimazaki, T. 2017. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling*, 57(6): 1300–1308. PMID: 28481528.

Parr, R. G.; and Weitao, Y. 1995. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press. ISBN 9780195092769.

Pyzer-Knapp, E. O.; Li, K.; and Aspuru-Guzik, A. 2015. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Advanced Functional Materials*, 25(41): 6495–6502.

Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; and Curioni, A. 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Mathematics*, 8: 84.

Riniker, S.; and Landrum, G. A. 2015. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12): 2562–2574. PMID: 26575315.

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33: 12559–12571.

Rothermel, D.; Li, M.; Rocktäschel, T.; and Foerster, J. 2021. Don't Sweep your Learning Rate under the Rug: A Closer Look at Cross-modal Transfer of Pretrained Transformers. *arXiv preprint arXiv:2107.12460*.

Satorras, V. G.; Hoogeboom, E.; and Welling, M. 2021. E(n) equivariant graph neural networks. In *International conference on machine learning*, 9323–9332. PMLR.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1): 61–80.

Schmidt, J.; Marques, M. R. G.; Botti, S.; and Marques, M. A. L. 2019. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Mathematics*, 5: 83.

Stärk, H.; Beaini, D.; Corso, G.; Tossou, P.; Dallago, C.; Günnemann, S.; and Liò, P. 2022. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, 20479–20502. PMLR.

Suh, C.; Fare, C.; Warren, J. A.; and Pyzer-Knapp, E. O. 2020. Evolving the Materials Genome: How Machine Learning Is Fueling the Next Generation of Materials Discovery. *Annual Review of Materials Research*, 50(1): 1–25.

Sun, R. 2022. Does GNN Pretraining Help Molecular Representation? *arXiv preprint arXiv:2207.06010*.

Thölke, P.; and De Fabritiis, G. 2022. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. *arXiv preprint arXiv:2202.02541*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Trivedi, P.; Lubana, E. S.; Heimann, M.; Koutra, D.; and Thiagarajan, J. J. 2022. Analyzing data-centric properties for contrastive learning on graphs. *arXiv preprint arXiv:2208.02810*.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, 9: 513–530.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; and Tang, J. 2022. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*.

You, Y.; Chen, T.; Shen, Y.; and Wang, Z. 2021. Graph contrastive learning automated. In *International Conference on Machine Learning*, 12121–12132. PMLR.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33: 5812–5823.

Zaidi, S.; Schaarschmidt, M.; Martens, J.; Kim, H.; Teh, Y. W.; Sanchez-Gonzalez, A.; Battaglia, P.; Pascanu, R.; and Godwin, J. 2022. Pre-training via Denoising for Molecular Property Prediction. *arXiv preprint arXiv:2206.00133*.