Molecular Contrastive Pretraining with Collaborative Featurizations

Anonymous authors
Paper under double-blind review

Abstract

Molecular pretraining, which learns molecular representations over massive unlabeled data, has become a prominent paradigm to solve a variety of tasks in computational chemistry and drug discovery. Recently, prosperous progress has been made in molecular pretraining with different molecular featurizations, including 1D SMILES strings, 2D graphs, and 3D geometries. However, the role of molecular featurizations with their corresponding neural architectures in molecular pretraining remains largely unexamined. In this paper, through two case studies—chirality classification and aromatic ring counting—we first demonstrate that different featurization techniques convey chemical information differently. In light of this observation, we propose a simple and effective MOlecular pretraining framework with COllaborative featurizations (MOCO). MOCO comprehensively leverages multiple featurizations that complement each other and outperforms existing state-of-the-art models that solely relies on one or two featurizations on a wide range of molecular property prediction tasks.

1 Introduction

Molecular representation learning, which automates the process of feature learning for molecules, is fast driving the development of computational chemistry and drug discovery. It has been recognized as crucial for a variety of downstream tasks, spanning from molecular property prediction to molecule design (Yang et al., 2019; Du et al., 2022). Deep neural models, on the other hand, rely on a substantial amount of labeled data, which require expensive wet lab experiments in chemical domains. With insufficient annotated data, deep models easily overfit to such small training data and tend to learn spurious correlations (Sagawa et al., 2020).

In recent years, self-supervised pretraining has emerged as a promising strategy to alleviate the label scarcity problem and improve model robustness (Jing & Tian, 2021). A typical framework pretrains the encoder model with training objectives over large-scale unlabeled datasets and then fine-tunes the learned model on labeled downstream tasks. Motivated by its success, many molecular pretraining models have been developed (Wang et al., 2019; Chithrananda et al., 2020; Hu et al., 2020b; You et al., 2020a; Xu et al., 2021a; Fang et al., 2022; Stärk et al., 2021; Liu et al., 2022a). To capture chemical semantics of molecules, these models design several pretraining strategies based on different *molecular featurizations*, which translate chemical information into representations that can be recognized by machine learning algorithms. For example, early models (Wang et al., 2019; Chithrananda et al., 2020) propose to leverage masked language modeling (Bengio et al., 2003) to pretrain Simplified Molecular-Input Line-Entry System (SMILES) strings (Weininger, 1988), while others study contrastive learning on 2D graphs (Hu et al., 2020b; You et al., 2020a; Xu et al., 2021a) or 3D conformations (Fang et al., 2022). Some recent studies further propose to enrich 2D-topology-based pretraining with 3D geometry information (Stärk et al., 2021; Liu et al., 2022a).

Despite encouraging progress, prior studies tend to emphasize on pretraining on molecular graphs and overlook the impact of other molecular featurizations with their corresponding neural encoders, which represent chemical information in different ways. Consider SMILES strings as an example. It explicitly represents informative structures in special characters such as branches, rings, and chirality (Ross et al., 2022), which are difficult to learn in graph-based representations (Chen et al., 2020b). Moreover, the utility of different featurizations may vary across downstream tasks. Therefore, most previous models relying on

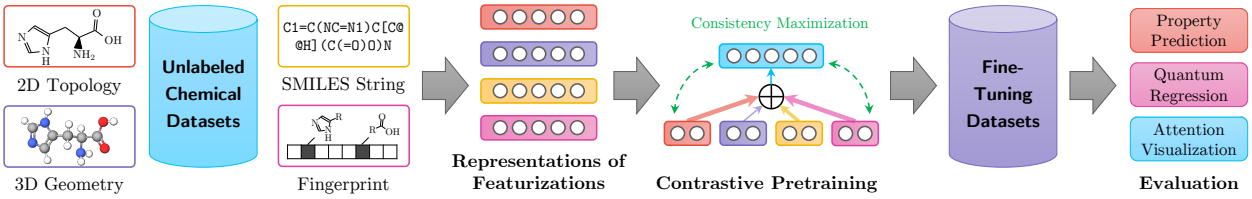


Figure 1: The proposed MOCO model. MOCO obtains four molecule featurizations with appropriate encoders. After that, an attention network is employed to aggregate each view embedding and compute a final embedding. The model is trained using a contrastive objective that maximizes the consistency between view embeddings and the final embedding.

only one or two featurizations might achieve sub-optimal performance across various downstream tasks. For example, 2D topology is important for many drug-related properties such as toxicity, while 3D geometry arguably determines properties related to quantum mechanics, such as single-point energy, atomic forces, or dipole moments (Zhang et al., 2018; Smith et al., 2017). Therefore, it is natural to ask whether we can enjoy the benefits from multiple molecular featurizations and take the relative utilities of different featurizations into consideration during fine-tuning on downstream tasks.

In this work, we first revisit four commonly used featurizations techniques: (a) 2D topology graphs, (b) 3D geometry graphs, (c) Morgan fingerprints, and (d) SMILES strings. We leverage four accompanying neural encoders with proper inductive bias and conduct two case studies, classifying tetrahedral chiral centers and counting aromatic rings, both of which are informative chemical descriptors, on representations obtained on different featurization techniques. The results show there is no one single featurization that dominates the others, indicating that different featurizations encode chemical semantics of molecules in different ways.

In light of this observation, we then propose a simple and effective MOlecular pretraining framework with COllaborative featurizations to comprehensively leverage every featurization during both pretraining and fine-tuning, which we term MOCO for brevity. Its graphical illustration is shown in Figure 1. The core idea of MOCO is to dynamically adjust the contribution of each featurization through an attention network, which *selectively* extracts information from each collaborative "view" of the raw molecular data. Besides, we design a novel multiview contrastive pretraining strategy, which trains the model by maximizing the consistency among different views in a self-supervised manner. Contrary to previous studies (Stärk et al., 2021; Liu et al., 2022a) that only consider 2D graph structures during fine-tuning, our MOCO utilizes multiple featurizations in *both* pretraining and fine-tuning stages and further allows interpretation analysis of different downstream tasks for domain scientists. Note that our proposed MOCO framework is generic, allowing for seamless integration of off-the-shelf neural architectures. To the best of our knowledge, this is the first work that studies how various featurization techniques should be utilized for molecular pretraining and downstream tasks.

We evaluate the effectiveness of our MOCO model on widely-used benchmark datasets including MoleculeNet (Wu et al., 2018) and QM9 (Ramakrishnan et al., 2014) that cover a wide range of molecular property prediction tasks. The results reveal that MOCO consistently improves non-pretraining baselines without negative transfer and outperforms existing state-of-the-art molecular pretraining models, achieving a 1.1% absolute improvement in terms of average ROC-AUC. Furthermore, the learned model weights of molecular featurizations for different end tasks are well aligned with prior chemical knowledge. We also suggest a series of guidelines on choosing effective featurization techniques for molecular representations.

The main contributions of this work are three-fold:

- We explore the featurization spaces of molecules with appropriate neural encoders and highlight the importance of incorporating different featurizations for molecular pretraining.
- We propose a novel molecular contrastive pretraining framework that adaptively integrates information from multiple collaborative featurizations during both pretraining and fine-tuning stages and provides interpretability for downstream molecular property prediction tasks.

- Extensive experiments conducted on public benchmark datasets validate the effectiveness of our proposed model. MOCO is able to achieve the state-of-the-art across various downstream datasets without negative transfer.

2 Preliminaries

2.1 A Brief Recapitulation of Molecular Featurization Techniques

Molecular featurizations translate chemical information of molecules into representations that can be understood by machine learning algorithms. Concretely, we consider the following molecular featurizations covering string-, graph-, scalar-, and vector-based representations for 1D/2D molecules and 3D structures, which are popular in literature (Ramsundar et al., 2019; Atz et al., 2021):

- **2D topology graphs** model atoms and bonds as nodes and edges respectively. It is arguably a common technique, especially for capturing substructure information by means of graph topology.
- **3D geometry graphs** incorporate atomic coordinates (conformations) in their representations and are able to depict how atoms are positioned relative to each other in the 3D space. We consider conformers in an equilibrium state, corresponding to the minima in a potential energy surface.
- **Morgan fingerprints** (Morgan, 1965; Glem et al., 2006) encode molecules in fixed-length binary strings, with bits indicating presence or absence of specific substructures. They represent each atom according to a set of atomic invariants and iteratively update these features among neighboring atoms using a hash function.
- **SMILES strings** are a concise technique that represents chemical structures in a linear notation using ASCII characters, with explicitly depicting information about atoms, bonds, rings, connectivity, aromaticity, and stereochemistry.

2.2 Learning Representations with Different Featurizations

Next, we introduce four encoders with different inductive bias to capture the intrinsic information with each featurization. Here we only discuss the high-level design of each encoder; please refer to Appendix A for detailed implementations of each encoder.

Notations. Each molecule can be represented as an undirected graph, where nodes are atoms and edges describe inter-atomic bonds. Formally, each graph is denoted as $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{X}, \mathbf{E})$, where $\boldsymbol{A} \in \{0,1\}^{N\times N}$ is the adjacency matrix of N nodes, $\boldsymbol{R} \in \mathbb{R}^{N\times 3}$ is the 3D position matrix, $\boldsymbol{X} \in \mathbb{R}^{N\times K}$ is the matrix of atom attributes of K dimension, and $\mathbf{E} \in \mathbb{R}^{N\times N\times E}$ is the tensor for bond attributes of E dimension. Additionally, each molecule is attached with a binary fingerprint vector $\boldsymbol{f} \in \{0,1\}^F$ of length F and a SMILES string $\mathbf{S} = [s_j]_{j=1}^S$ of length S. In what follows, the subscript i is used to index the i-th molecule.

Embedding 2D graphs. To capture the 2D topological information, we employ a widely-used Graph Isomorphism Network (GIN) model (Xu et al., 2019) denoted by f_{2D}, which receives as input the graph adjacency matrix and attributes of atoms and bonds, and produces the embedding vector $\boldsymbol{z}_i^{\text{2D}} \in \mathbb{R}^D$:

$$\boldsymbol{z}_i^{\text{2D}} = f_{\text{2D}}(\boldsymbol{X}_i, \mathbf{E}_i, \boldsymbol{A}_i). \tag{1}$$

Embedding 3D graphs. To model additional spatial coordinates associated with atoms, we leverage SchNet (Schütt et al., 2017) as the backbone, which models message passing as continuous-filter convolutions and is able to preserve rotational invariance for energy predictions. We denote its encoding function as f_{3D} which takes atom features and positions as input and produces the 3D embedding $\boldsymbol{z}_i^{\text{3D}} \in \mathbb{R}^D$:

$$\boldsymbol{z}_i^{\text{3D}} = f_{\text{3D}}(\boldsymbol{X}_i, \boldsymbol{R}_i). \tag{2}$$

Table 1: Results of two case studies with different featurizations: chirality classification and aromatic ring count regression.

Target	2D	3D	SM	FP
Chirality (AP, ↑)	0.4952	0.4959	**0.5505**	0.5246
#Rings (MAE, ↓)	**0.1949**	0.2021	0.3077	0.2590

Embedding molecular fingerprints. Since there is a lack of proper neural encoders for fingerprints, we propose an attention-based network to model interactions of feature fields in fingerprint vectors, which considers the discrete and extremely sparse nature of fingerprints. Specifically, we first transform all F feature fields into a dense embedding matrix $\boldsymbol{F}_i \in \mathbb{R}^{F \times D_{\mathrm{F}}}$ via embedding lookup. Then, we use a multihead self-attention network f_{FP} (Vaswani et al., 2017) to model the interaction among those feature fields, resulting in an embedding matrix $\widehat{\boldsymbol{Z}}_i^{\mathrm{FP}} \in \mathbb{R}^{F \times D_{\mathrm{F}}}$. Following that, we perform sum pooling and use a linear model f_{LIN} to obtain the final fingerprint embedding $z_i^{\mathrm{FP}} \in \mathbb{R}^D$:

$$\widehat{\boldsymbol{Z}}_i^{\mathrm{FP}} = f_{\mathrm{FP}}(\boldsymbol{F}_i), \qquad \boldsymbol{z}_i^{\mathrm{FP}} = f_{\mathrm{LIN}}\left(\sum_{d=1}^{D_{\mathrm{F}}} \widehat{\boldsymbol{Z}}_{i,d}^{\mathrm{FP}}\right). \tag{3}$$

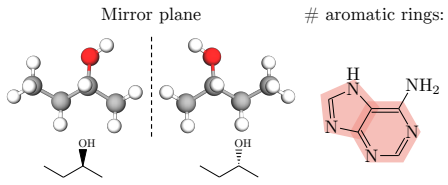
Embedding SMILES strings. To encode SMILES strings, we use a pretrained RoBERTa (Liu et al., 2019b) as the backbone model. As SMILES strings do not possess consecutive relationships, the RoBERTa model is pretrained using the masked language model as the only objective, unlike conventional natural language models (Devlin et al., 2019). After that, in order to reduce the computational burden, we freeze the RoBERTa encoder (denoted by f_{SM}) in our model and employ an additional learnable MultiLayer Perceptron (MLP) on the representation $\boldsymbol{s}_i \in \mathbb{R}^{D_{\mathrm{S}}}$ to get the final embedding $\boldsymbol{z}_i^{\mathrm{SM}} \in \mathbb{R}^D$:

$$\boldsymbol{s}_i = f_{\mathrm{SM}}(\mathbf{S}_i), \qquad \boldsymbol{z}_i^{\mathrm{SM}} = f_{\mathrm{MLP}}(\boldsymbol{s}_i). \tag{4}$$

2.3 Case Studies

In this section, we present two case studies—chirality classification and aromatic ring counting—to demonstrate that the representation ability of each featurization with the corresponding neural encoder is different. For chirality classification, we randomly select 10K molecules with one chirality center from GEOM-Drugs (Axelrod & Gómez-Bombarelli, 2022) and test whether the representations obtained using the four featurizations can classify tetrahedral chiral centers as R/S. For aromatic ring counting, we randomly draw another 10K molecules and test whether these models can recognize the number of aromatic rings of each molecule. Note that both chirality properties and ring counts are informative chemical descriptors (Ritchie & Macdonald, 2009) and can be easily computed with existing implementations such as RDKit (Landrum et al., 2022).

Figure 2: (a) Chirality: even if two graphs are isomorphic, they can have two distinct stereochemistry structures. (b) The aromatic ring is an important functional group.



(a) (*R*)-(−)-2-Butanol, (*S*)-(+)-2-Butanol (b) Adenine

We report classification and regression performance in Average Precision (AP) and Mean Absolute Error (MAE) respectively. The results are summarized in Table 1. It is seen from the table that no single featurization performs the best on all targets and four representations contain collaborative information to each other, suggesting us to leverage multiple featurizations for molecular pretraining.

3 Molecular Pretraining with Collaborative Featurizations

As with generic self-supervised learning pipelines, the MOCO framework is divided into two stages, pretraining and fine-tuning. In the first stage, given an unlabeled dataset, we train an encoding function that learns

representations with the four featurization techniques. In the subsequent fine-tuning phase, we take the weights of the encoders from the pretrained model and tune the model on molecules with annotations of particular properties in a supervised fashion.

We next introduce the MOCO pretraining framework in detail. We first use obtain four "view" representations based on the aforementioned four featurizations. Then, we integrate these four embeddings to compute a final representation for each molecule through an attention network. Finally, we pretrain the whole model using a contrastive objective.

3.1 Representation Aggregation from Multiple Featurizations

Since each featurization technique reflects the molecule from one certain aspect, we take weighted average of every view embedding to obtain a comprehensive final representation:

$$\boldsymbol{z}_i = \sum_{m \in \mathcal{M}} \alpha^m \boldsymbol{z}_i^m, \tag{5}$$

where $\mathcal{M} = \{2\text{D}, 3\text{D}, \text{FP}, \text{SM}\}$ is the set of all views. We leverage an attention network (Bahdanau et al., 2015) that learns to adjust the contribution of each view. Formally, the attention coefficient α^m denoting the contribution of the m-th view is computed by:

$$\alpha^m = \frac{\exp(w^m)}{\sum_{m' \in \mathcal{M}} \exp(w^{m'})}, \qquad w^m = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \boldsymbol{q}^\top \cdot \tanh\left(\boldsymbol{W} \frac{\boldsymbol{z}_i^m}{\|\boldsymbol{z}_i^m\|_2} + \boldsymbol{b}\right), \tag{6}$$

where $\boldsymbol{q}, \boldsymbol{b} \in \mathbb{R}^D$, $\boldsymbol{W} \in \mathbb{R}^{D \times D}$ are trainable parameters in the attention network, and $\mathcal{B}$ denotes the set of molecules in the current training batch. Note that we perform ℓ_2 normalization on all embeddings to regularize the scale across different views when computing the attention scores.

3.2 Contrastive Objectives for Pretraining

Finally, we train the model using a contrastive objective by aligning the aggregated embedding with all view-specific embeddings. Particularly, for one molecule i, we designate its four view embeddings $\boldsymbol{z}_i^m$ as the anchors and the aggregated embeddings $\boldsymbol{z}_i$ as the positive instance. Other aggregated embeddings $\{\boldsymbol{z}_j\}_{i \neq j}$ in the same batch are then chosen as the negative samples. Following prior studies (Chen et al., 2020a; He et al., 2020; Bachman et al., 2019; Zhu et al., 2020; You et al., 2020a; Zhu et al., 2021a), we leverage the Information Noice Contrastive Estimation (InfoNCE) objective, which can be formally written as:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} -\log \frac{\exp(\theta(\boldsymbol{z}_i^m, \boldsymbol{z}_i)/\tau)}{\sum_{j \in \mathcal{B}} \exp(\theta(\boldsymbol{z}_i^m, \boldsymbol{z}_j)/\tau)} \right], \tag{7}$$

where the critic function θ computes the likelihood scores of contrastive pairs and the hyperparameter τ adjusts the dynamic range of the likelihood scores of contrastive pairs. Specifically, the critic function θ performs non-linear transformation via an MLP function g (Chen et al., 2020a) and then measures their cosine similarity:

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{g(\boldsymbol{x})^\top g(\boldsymbol{y})}{\|g(\boldsymbol{x})\|_2 \|g(\boldsymbol{y})\|_2}. \tag{8}$$

After pretraining the model with the self-supervised objective function $\mathcal{L}$, we fine-tune the model weights of view encoders along with the attentive representation aggregation module with the supervision of downstream tasks at a smaller learning rate.

4 Experiments

In this section, we present empirical evaluation of our proposed work. Specifically, the experiments aim to investigate the following three key questions.

- **RQ1 (Overall performance).** Is the proposed MOCO able to improve non-pretraining baselines and outperform state-of-the-arts on molecular property prediction tasks?
- **RQ2 (Interpretation).** Are the learned attention weights of molecular featurizations on different downstream tasks consistent with chemical knowledge?
- **RQ3 (Ablation studies).** How do the representation aggregation module and the fine-tuning strategy affect the model performance?

In the following, we first summarize experimental setup and proceed to results and analysis.

4.1 Experimental Configurations

Datasets. We closely follow the experimental setup of GraphMVP (Liu et al., 2022a) for fair comparison. Specifically, we pretrain the model using the GEOM-Drugs dataset (Axelrod & Gómez-Bombarelli, 2022) containing both 2D and 3D information. For fine-tuning, we choose a variety datasets extracted from MoleculeNet (Wu et al., 2018), ChEMBL (Gaulton et al., 2011), and CEP (Hachmann et al., 2011), that cover a wide range of applications, including physiological, biological, and pharmaceutical tasks, and QM9 (Ramakrishnan et al., 2014) that focuses on quantum property prediction. These downstream tasks include 8 binary classification and 12 regression tasks. For those datasets for fine-tuning, we follow OGB (Hu et al., 2020a) that uses scaffolds to split training/test/validation subsets with a split ratio of 80%/10%/10%. For detailed description, we refer readers of interest to Appendix B.

Baselines. For comprehensive comparison, we select the following two groups of SSL methods as primary baselines in our experiments.

- Generic graph SSL models: GraphSAGE (Hamilton et al., 2017), InfoGraph (Sun et al., 2020a), GPT-GNN (Hu et al., 2020c), AttrMask, ContextPred (Hu et al., 2020b), GraphLoG (Xu et al., 2021a), GraphCL (You et al., 2020a), JOAO (You et al., 2021), and GraphMAE (Hou et al., 2022).
- Molecular SSL models: GROVER-Contextual (GROVER-C), GROVER-Motif (GROVER-M) (Rong et al., 2020), and GraphMVP[1] (Liu et al., 2022a).

In the pretraining stage, all the above SSL approaches are trained on the same dataset based on GEOM-Drugs. We also report performance with a randomly initialized model as the non-pretraining baseline. To ensure the performance is comparable with existing work, we report all baseline performance from previously published results (Liu et al., 2022a; Hou et al., 2022).

Implementation details. In the GEOM-Drugs dataset, since the original full set is too large (containing 317K molecules with over 9M conformations), we randomly select 50K molecules as the pretraining dataset. For each molecule, we select to use its top-5 conformers of the lowest energy in virtue of their sufficient geometry information. Since molecules in the fine-tuning datasets do not have 3D information available, we use ETKDG (Riniker & Landrum, 2015) in RDkit (Landrum et al., 2022) to compute molecular conformations. For both pretraining and fine-tuning datasets, we use RDkit to generate 1024-bit molecular fingerprints with radius $R = 2$, which is roughly equivalent to the ECFP4 scheme (Rogers & Hahn, 2010). We would like to emphasis that all dataset preprocessing and graph encoder architectures are kept in line with GraphMVP (Liu et al., 2022a) to ensure fair comparison. Readers of interest may refer to Appendix C for implementation details regarding software/hardware platforms, model training, and hyperparameter specifications.

Evaluation protocols. For classification tasks, we report the performance in terms of the Area Under the ROC-Curve (ROC-AUC), where higher values indicate better performance. For quantum property and other non-quantum regression tasks, we measure the performance in Mean Absolute Error (MAE) and Root Mean

[1]In our experiments, we do not include its two variants GraphMVP-G and GraphMVP-C since they are essentially two ensemble models that combine AttrMask and ContextPred (Hu et al., 2020b) respectively.

Table 2: Results for eight molecule property prediction tasks in terms of ROC-AUC (%, ↑). We highlight the best- and the second-best performing results in **boldface** and underlined, respectively.

Pretraining	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	Avg.
—	71.0±0.5	75.9±0.3	64.7±2.3	57.7±3.1	71.5±5.3	77.7±1.0	75.9±0.7	71.5±2.7	70.63
GraphSAGE	64.5±3.1	74.5±0.4	60.8±0.5	56.7±0.1	55.8±6.2	73.3±1.6	75.1±0.8	64.6±4.7	65.64
AttrMask	70.2±0.5	74.2±0.8	62.5±0.4	60.4±0.6	68.6±9.6	73.9±1.3	74.3±1.3	77.2±1.4	70.16
GPT-GNN	64.5±1.1	75.3±0.5	62.2±0.1	57.5±4.2	57.8±3.1	76.1±2.3	75.1±0.2	77.6±0.5	68.27
InfoGraph	69.2±0.8	73.0±0.7	62.0±0.3	59.2±0.2	75.1±5.0	74.0±1.5	74.5±1.8	73.9±2.5	70.10
ContextPred	71.2±0.9	73.3±0.5	62.8±0.3	59.3±1.4	73.7±4.0	72.5±2.2	75.8±1.1	78.6±1.4	70.89
GraphLoG	67.8±1.7	73.0±0.3	62.2±0.4	57.4±2.3	62.0±1.8	73.1±1.7	73.4±0.6	78.8±0.7	68.47
GROVER-C	70.3±1.6	75.2±0.3	62.6±0.3	58.4±0.6	59.9±8.2	72.3±0.9	75.9±0.9	79.2±0.3	69.21
GROVER-M	66.4±3.4	73.2±0.8	62.6±0.5	60.6±1.1	77.8±2.0	73.3±2.0	73.8±1.4	73.4±4.0	70.14
GraphCL	67.5±3.3	75.0±0.3	62.8±0.2	60.1±1.3	78.9±4.2	77.1±1.0	75.0±0.4	68.7±7.8	70.64
JOAO	66.0±0.6	74.4±0.7	62.7±0.6	60.7±1.0	66.3±3.9	77.0±2.2	76.6±0.5	72.9±2.0	69.57
GraphMVP	68.5±0.2	74.5±0.4	62.7±0.1	**62.3±1.6**	79.0±2.5	75.0±1.4	74.8±1.4	76.8±1.1	71.69
GraphMAE	70.9±0.9	75.0±0.4	64.1±0.1	59.9±0.5	81.5±2.8	76.9±2.6	76.7±0.9	81.4±1.4	73.31
MOCO	**71.6±1.0**	**76.7±0.4**	**64.9±0.8**	61.2±0.6	**81.6±3.7**	**78.5±1.4**	**78.3±0.4**	**82.6±0.3**	**74.41**

Squared Error (RMSE) respectively, where lower values are better. We repeat every experiment on three seeds with scaffold splitting and report the averaged performance with standard deviation, following previous work (Liu et al., 2022a).

4.2 Main Results on Molecular Property Prediction

The performance of molecular property prediction tasks is summarized in Table 2. It can be found that our MOCO shows strong empirical performance across all eight low-data downstream datasets, delivering seven out of eight state-of-the-art results and acquiring a 1.1% absolute improvement on average. The outstanding results validate the superiority of our proposed model.

We make other observations as follows. Firstly, MOCO obtains more accurate and stabler predictions compared to the randomly initialized baseline, indicating that our pretraining framework can transfer the knowledge from large, unannotated datasets to smaller downstream datasets without negative transfer. Secondly, previous work has already achieved pretty high performance. For example, the current state-of-the-art GraphMVP only obtains a 0.8% absolute improvement over its best baseline ContextPred in terms of average ROC-AUC. Our work pushes that boundary without extensive hyperparameter tuning, with an absolute improvement of up to 3.4% over GraphMVP in terms of average ROC-AUC. Lastly, it is worth mentioning that, the non-pretraining baseline even achieves better performance than some graph-based pretraining models. On some challenging datasets (e.g., Tox21, MUV, and ToxCast), it even achieves the second to best performance. This once more demonstrates the effectiveness of leveraging multiple featurization techniques.

4.3 Interpretation and Analysis

In order to analyze the correlation between tasks and featurization techniques, we visualize the attention weights $\boldsymbol{\alpha}$ learned on different downstream tasks in Figure 3. Note that most of the datasets in MoleculeNet (Wu et al., 2018) are ADMET property prediction tasks: chemical Absorption (A), Distribution (D), Metabolism (M), Excretion (E), and Toxicity (T), and we thus group the eight end tasks according to their prediction targets in the following analysis.

In general, we can interpret from the visualization that *2D-based features are more significant than 3D-based features in the studied tasks*, which is well aligned with chemical knowledge. We provide detailed analysis as follows:

- In Tox21, ClinTox, SIDER, and ToxCast, we find that 2D graphs play the most important role. These four datasets are related to toxicity (or side effects). Although it is a very complex biological issue to explain, such properties can still be partially deduced from certain functional groups patterns contained in 2D graphs. Actually, medicinal chemists have developed such a database to provide them with necessary alerts of potential side effects in drug design (Baell & Holloway, 2010).
- BBBP, which measures blood-brain barrier permeability, is mostly dominated by the following properties: liposolubility/water-solubility, molecular weight, and interaction between molecules and transporter proteins. Similarly, these properties can also be inferred from 2D topology, such as molecules with too many hydrogen bond acceptors/donors are unlikely to break the blood-brain barrier due to poor liposolubility (Suckling et al., 1986).
- On BACE and MUV we see 2D graphs and SMILES strings contribute most. These two datasets are about predicting protein-ligand binding activities, which are theoretically relevant to 3D conformations. However, it is still an open question that whether the conformation sampling methods can produce conformations that resemble bioactive conformations, which provide the key information for protein-ligand binding. Nevertheless, in each of these tasks, the target protein is fixed so that bioactivity can be partially deduced from 2D structures, which is supported by the success of fragment-based Quantitive Structure-Activity Relationship (QSAR) models (Manoharan et al., 2010).
- Due to the complicated pathogenetic mechanisms, it is hard to draw an explanation to why attention weights of fingerprints outweigh the other three features in the HIV task. Given that the HIV dataset is the largest one (over 40,000 molecules per task), one possible explanation of this phenomenon is that we use a high-dimensional fingerprint representations (1024 bits).

Concerning the difference between three 2D-based features (namely 2D topological graphs, fingerprints, and SMILES strings), we make the following findings, which we hope could serve as guidelines for future research on molecular representation learning:

- 2D graph representations can encode local information explicitly by resembling chemical structures. Besides, graph-based neural networks can capture long-range local chemical environment through message passing. For example, with molecular graphs, it is more convenient to identify which part of the molecule serves as a scaffold.
- In principle, SMILES strings contain all 2D information of certain molecules, but with atoms and bonds represented in ASCII characters, neural networks may have difficulty in distilling semantic meanings of chemical structures in a numerical way.
- Fingerprint representations are based on local structures and thus such features may be less effective in circumstances where long-range effects induced by topologically distant functional groups predominate, which accounts for relatively small attention weights of fingerprints in Figure 3.

4.4 More Experiments on Molecular Property Regression

To demonstrate that the conformations generated by RDKit are helpful, we further conduct an experiment on quantum property regression on the QM9 dataset (Ramakrishnan et al., 2014), where 3D conformations generated by RDKit are used for the fine-tuning datasets. This task is known to be closely related to 3D structures. Table 3 presents the performance comparison of MEMO with two non-pretraining (supervised) baselines SchNet and MOCO (denoted by SchNet-NP and MOCO-NP) and two state-of-the-art pretraining baselines GraphMVP (Liu et al., 2022a) and 3D Infomax (Stärk et al., 2021).

It is seen that our MOCO model achieves the best performance on all datasets. GraphMVP that consider only 2D structures

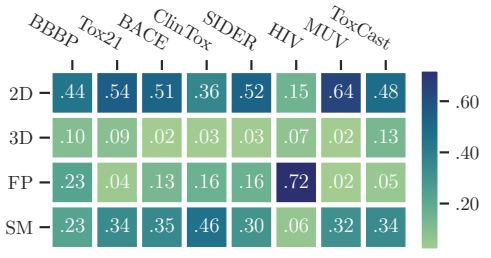


Figure 3: Visualizing the learned attention weights on eight molecular property prediction datasets.

Table 3: Results for eight molecule quantum property regression tasks in terms of Mean Absolute Error (MAE, ↓). The highest performance is highlighted in **bold**.

Target Unit	μ D	α Bohr^3	ϵ_{HOMO} meV	ϵ_{LUMO} meV	ϵ_{gap} meV	U_0 meV	U meV	$\langle R^2 \rangle$ Bohr^3
SchNet-NP	0.4604	0.3251	95.9740	78.5870	136.4720	98.1240	100.1650	24.3277
MOCO-NP	0.3767	0.2439	73.0625	69.8780	102.2332	77.4708	92.8562	17.5842
GraphMVP	0.3726	0.4390	75.3750	72.3820	104.8370	278.8900	325.8021	22.6433
3D Infomax	0.3644	0.4190	72.0558	67.6203	99.4032	207.2148	219.5415	20.3934
MOCO	**0.3618**	**0.2236**	**71.5120**	**58.5890**	**97.7440**	**64.3550**	**66.3958**	**15.5571**

during fine-tuning even result in negative transfer on some datasets. Our MOCO, on the contrary, achieves better performance than the supervised baseline, underscoring the value of leveraging 3D structures (as well as other sources of 2D information) during fine-tuning.

We also perform experiments on non-quantum property regression tasks. Our proposed MOCO also obtains promising improvements compared to the current state-of-the-art baselines. Please refer to Appendix D.1 for performance comparison and analysis.

4.5 Ablation Studies

Finally, we conduct ablation studies on the representation aggregation module and the fine-tuning strategy. We consider the following model variants for further inspection. Except the modifications in specific modules, other implementations remain the same as previously described.

- **MOCO–Max** removes the attention network in the representation aggregation module in Equation (5) and simply uses max pooling to combine view embeddings.
- **MOCO–Mean** modifies representation aggregation by taking average over view embeddings.
- **MOCO–Freeze** does not fine-tune the representation aggregation module but instead uses the frozen weights of the pretrained model.

We report the performance of model variants in Figure 4. It is seen that all three variants achieve downgraded performance, which empirically rationalizes the design choice of our molecular pretraining framework with collaborative featurizations. Specifically, the performance of MOCO–Max and MOCO–Mean without attention aggregation mechanisms of multiple featurizations is inferior to that of MOCO, demonstrating the necessity of adaptively combining information from multiple featurizations. In addition, MOCO–Freeze occasionally obtains better performance than the two other variants, which indicates that our proposed attention network is

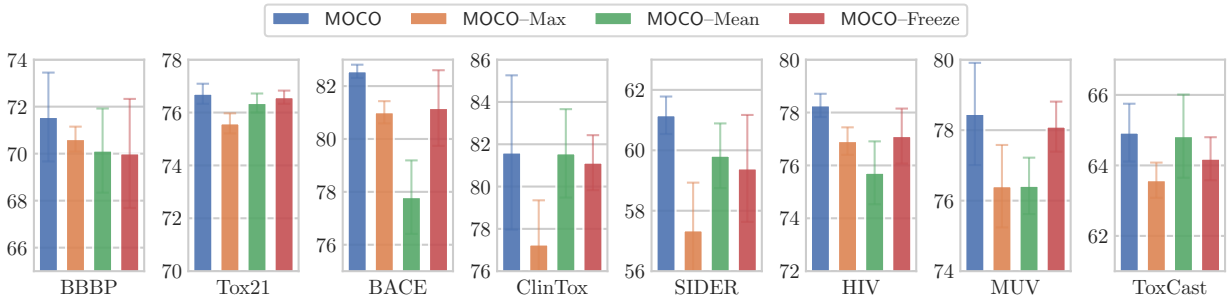


Figure 4: Ablation studies on representation aggregation and the fine-tuning strategy.

able to select information from different views. It does not, however, fine-tune the contribution of featurizations with downstream datasets, where the optimal combination might differ, resulting in performance deterioration.

Moreover, we conduct ablation studies on models that include only three view representations, where the results can be found in Appendix D.2. Results demonstrate the necessity of comprehensively leveraging four views in the proposed MOCO model.

5 Related Work

Traditional methods (Carhart et al., 1985; Nilakantan et al., 1987; Rogers & Hahn, 2010) represent molecular structures with fingerprints. Some prior studies (Svetnik et al., 2004; Meyer et al., 2019; Wu et al., 2018) employ tree-based machine leaning models such as random forests (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016) on fingerprints to predict the properties of molecules. With the development of deep learning, neural approaches have been dominating the field given their strong representation ability. One line of work (Wang et al., 2019; Chithrananda et al., 2020) leverages language modeling techniques such as BERT (Devlin et al., 2019) to learn molecular representations based on SMILES strings (Weininger, 1988). However, some argue that sequence-based representations cannot fully capture substructure information and propose to leverage Graph Neural Networks (GNNs), which model molecules as graphs with atoms as nodes and bonds as edges (Gilmer et al., 2017; Liu et al., 2019a; Ying et al., 2021). Despite the prosperous progress, they only model 2D topological structures of molecules, without considering the 3D coordinates of atoms that are known to determine certain chemical and physical functionalities of molecules. To address this deficiency, recent work further explicitly considers such 3D geometry and designs equivariant networks to obtain the representations (Schütt et al., 2017; Klicpera et al., 2020; Satorras et al., 2021; Fuchs et al., 2020; Schütt et al., 2021; Du et al., 2021; Liu et al., 2021; Gasteiger et al., 2021; Batzner et al., 2021; Brandstetter et al., 2022; Xu et al., 2021b).

Even though molecular representation learning techniques have been extensively investigated, there are very few labeled datasets available for studying the molecular properties of interest (e.g., drug-likeness or quantum properties). On the other hand, there are abundant unannotated molecules available, which motivates researchers to study pretraining techniques that learn the model weights in a self-supervised manner and transfer the knowledge to downstream datasets with limited annotations via fine-tuning. A series of pretraining frameworks on 2D molecular graph representations have been developed so far (Rong et al., 2020; Hu et al., 2020b; Zhang et al., 2021; Wang et al., 2022; Li et al., 2020). Recent work GEM (Fang et al., 2022) studies large-scale pretraining for 3D geometry representations. Additionally, researchers also study to supplement 2D-graph-based pretraining with 3D conformation information (Yang et al., 2021; Liu et al., 2022a; Stärk et al., 2021).

A succinct comparison of our work with other representative methods is provided in Table 4. Compared to the above studies, our proposed MOCO is the only model that can *adaptively* leverage multiple featurizations for both pretraining and fine-tuning stages.

6 Conclusions and Discussions

This paper examines different featurizations for molecular data and highlights the importance of incorporating multiple featurizations during both pretraining and fine-tuning. Then, we develop a novel pretraining framework MOCO with collaborative featurizations for molecular data, which is able to adaptively distill information from each featurization and allows interpretability from the learned model weights. Extensive experiments on a wide range of property prediction benchmarks show that MOCO consistently outperforms existing baselines without negative transfer.

The study of featurization techniques for molecular machine learning in general remains widely open. We would like to acknowledge that the relative utility of various featurizations for different molecular predictive tasks could be usefully explored in further work. Moreover, more future research should be undertaken to specifically analyze the relationship between several featurizations, the representation ability of corresponding neural architectures, as well as the task-featurization correlation.

Table 4: Comparing MOCO with representative self-supervised methods on molecular pretraining.

Method	Pretraining				Fine-tuning			
	2D	3D	Fingerprint	SMILES	2D	3D	Fingerprint	SMILES
SMILES-BERT (Wang et al., 2019)				✓				✓
ChemBERTa (Chithrananda et al., 2020)				✓				✓
AttrMask, ContexPred (Hu et al., 2020b)	✓				✓			
GraphCL (You et al., 2020a)	✓				✓			
GraphLoG (Xu et al., 2021a)	✓				✓			
GROVER (Rong et al., 2020)	✓				✓			
GEM (Fang et al., 2022)		✓				✓		
3D Infomax (Stärk et al., 2021)	✓	✓			✓			
GraphMVP (Liu et al., 2022a)	✓	✓			✓			
MOCO (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

References

AIDS Antiviral Screen Data. URL https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data. 22

Tox21 Data Challenge 2014, 2014. URL https://tripod.nih.gov/tox21/challenge/. 23

Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric Deep Learning on Molecular Representations. *Nat. Mach. Intell.*, 3(12):1023–1032, 2021. 3

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation. *Sci. Data*, 9(1):185, 2022. 4, 6, 22

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. In *NeurIPS*, pp. 15509–15519, 2019. 5, 26

Jonathan B. Baell and Georgina A. Holloway. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, 53(7): 2719–2740, 2010. 8

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2015. 5

Simon L. Batzner, Tess E. Smidt, Lixin Sun, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. SE(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *arXiv.org*, 2021. 10

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003. 1, 27

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J. Bekkers, and Max Welling. Geometric and Physical Quantities Improve E(3) Equivariant Message Passing. In *ICLR*, 2022. 10

Leo Breiman. Random Forests. *Mach. Learn.*, 45(1):5–32, 2001. 10

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, pp. 1877–1901, 2020. 27

Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.*, 25(2):64–73, 1985. 10

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, pp. 139–156, 2018. 26, 27

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, pp. 9912–9924, 2020. 26

Ricky T. Q. Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *KDD*, pp. 785–794, 2016. 10

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, pp. 1597–1607, 2020a. 5, 26, 27

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *CVPR*, pp. 15745–15753, 2021. 26

Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can Graph Neural Networks Count Substructures? In *NeurIPS*, pp. 10383–10395, 2020b. 1

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv.org*, 2020. 1, 10, 11

John S. Delaney. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.*, 44(3):1000–1005, 2004. 22

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pp. 4171–4186, 2019. 4, 10, 27

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 27

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Bin Shao, and Tie-Yan Liu. Equivariant Vector Field Network for Many-Body System Modeling. *arXiv.org*, 2021. 10

Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. *arXiv.org*, 2022. 1

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. *Nat. Mach. Intell.*, 4:127–134, 2022. 1, 10, 11

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *RLGM@ICLR*, 2019. 24

Fabian Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. In *NeurIPS*, pp. 1970–1981, 2020. 10

Philip Gage. A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38, 1994. 21

Francisco-Javier Gamo, Laura M. Sanz, Jaume Vidal, Cristina de Cozar, Emilio Alvarez, Jose-Luis Lavandera, Dana E. Vanderwall, Darren V. S. Green, Vinod Kumar, Samiul Hasan, James R. Brown, Catherine E. Peishoff, Lon R. Cardon, and Jose F. Garcia-Bustos. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature*, 465(7296):305–310, 2010. 22

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, pp. 6894–6910, 2021. 26

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules. In *NeurIPS*, pp. 6790–6802, 2021. 10

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, 2011. 6, 22

Kaitlyn M. Gayvert, Neel S. Madhukar, and Olivier Elemento. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem. Biol.*, 23(10):1294–1301, 2016. 23

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*, 2018. 26

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *ICML*, pp. 1263–1272, 2017. 10

Robert C. Glem, Andreas Bender, Catrin H. Arnby, Lars Carlsson, Scott Boyer, and James Smith. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs*, 9(3):199–204, 2006. 3, 23

Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *AISTATS*, pp. 249–256, 2010. 24

Stefan Grimme. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.*, 15(5): 2847–2862, 2019. 22

Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.*, 2(17):2241–2251, 2011. 6, 22

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *NIPS*, pp. 1024–1034, 2017. 6, 27

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive Multi-View Representation Learning on Graphs. In *ICML*, pp. 4116–4126, 2020. 27

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, pp. 9726–9735, 2020. 5, 26, 27

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2022. 27

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-Supervised Masked Graph Autoencoders. In *KDD*, pp. 594–604, 2022. 6

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *NeurIPS*, pp. 22118–22133, 2020a. 6, 23, 24

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for Pre-training Graph Neural Networks. In *ICLR*, 2020b. 1, 6, 10, 11, 20, 24, 27

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. GPT-GNN: Generative Pre-Training of Graph Neural Networks. In *KDD*, pp. 1857–1867, 2020c. 6, 27

Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):4037–4058, 2021. 1, 26

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 24

Thomas N. Kipf and Max Welling. Variational Graph Auto-Encoders. In *BDL@NIPS*, 2016. 27

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In *NIPS*, pp. 3294–3302, 2015. 26

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional Message Passing for Molecular Graphs. In *ICLR*, 2020. 10

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.*, 44(D1):D1075–D1079, 2016. 23

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*, 2020. 27

Greg Landrum, Paolo Tosco, Brian Kelley, Ric, sriniker, gedeck, Riccardo Vianello, NadineSchneider, Eisuke Kawashima, Andrew Dalke, Dan N, David Cosgrove, Brian Cole, Matt Swain, Samo Turk, AlexanderSavelyev, Gareth Jones, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Daniel Probst, Kazuya Ujihara, Vincent F. Scalfani, guillaume godin, Axel Pahl, Francois Berenger, JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_03_2 (q1 2022) release, 2022. 4, 6, 23, 24

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a Proxy Task for Visual Understanding. In *CVPR*, pp. 840–849, 2017. 26

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, pp. 7871–7880, 2020. 26

Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. Learn Molecular Representations From Large-Scale Unlabeled Molecules for Drug Discovery. *arXiv.org*, 2020. 10

Shuai Lin, Pan Zhou, Zi-Yuan Hu, Shuojia Wang, Ruihui Zhao, Yefeng Zheng, Liang Lin, Eric P. Xing, and Xiaodan Liang. Prototypical Graph Contrastive Learning. *arXiv.org*, 2021. 27

Shengchao Liu, Mehmet Furkan Demirel, and Yingyu Liang. N-Gram Graph: Simple Unsupervised Representation for Graphs, with Applications to Molecules. In *NeurIPS*, pp. 8464–8476, 2019a. 10

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training Molecular Graph Representation with 3D Geometry. In *ICLR*, 2022a. 1, 2, 6, 7, 8, 10, 11, 20, 22, 24

Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical Message Passing for 3D Graph Networks. *arXiv.org*, 2021. 10

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv.org*, 2019b. 4, 21, 27

Yixin Liu, Shirui Pan, Ming Jin, Chuan Zhou, Yu Zheng, Feng Xia, and Philip S. Yu. Graph Self-Supervised Learning: A Survey. *IEEE Trans. Knowl. Data Eng.*, 2022b. 27

Prabu Manoharan, R.S.K. Vijayan, and Nanda Ghoshal. Rationalizing Fragment Based Drug Discovery for BACE1: Insights From FB-QSAR, FB-QSSR, Multi Objective (MO-QSPR) and MIF Studies. *J. Comput. Aided Mol. Des.*, 24(10):843–864, 2010. 8

Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.*, 52(6):1686–1697, 2012. 23

Jesse G. Meyer, Shengchao Liu, Ian J. Miller, Joshua J. Coon, and Anthony Gitter. Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests. *J. Chem. Inf. Model.*, 59(10):4438–4449, 2019. 10

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pp. 3111–3119, 2013. 26

H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures — A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.*, 5(2):107–113, 1965. 3, 23

Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.*, 27(2):82–85, 1987. 10

Mehdi Noroozi and Paolo Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV*, pp. 69–84, 2016. 26

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, pp. 8024–8035, 2019. 24

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In *CVPR*, pp. 2536–2544, 2016. 26

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *NAACL-HLT*, pp. 2227–2237, 2018. 27

Senthil Purushwalkam and Abhinav Gupta. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases. In *NeurIPS*, pp. 3407–3418, 2020. 27

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI Blog, 2018. 27

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI Blog, 2019. 27

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum Chemistry Structures and Properties of 134 kilo Molecules. *Sci. Data*, 1(1):1–7, 2014. 2, 6, 8, 23

Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media, 2019. 3

Ann M. Richard, Richard S. Judson, Keith A. Houck, Christopher M. Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T. Martin, John F. Wambaugh, Thomas B. Knudsen, Jayaram Kancherla, Kamel Mansouri, Grace Patlewicz, Antony J. Williams, Stephen B. Little, Kevin M. Crofton, and Russell S. Thomas. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.*, 29(8):1225–1251, 2016. 23

Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015. 6, 23

Timothy J. Ritchie and Simon J.F. Macdonald. The Impact of Aromatic Ring Count on Compound Developability – Are Too Many Aromatic Rings a Liability in Drug Design? *Drug Discov. Today*, 14(21): 1011–1020, 2009. 4

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010. 6, 10, 23

Sebastian G. Rohrer and Knut Baumann. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.*, 49(2):169–184, 2009. 22

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *NeurIPS*, pp. 12559–12571, 2020. 6, 10, 11

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Molformer: Large Scale Chemical Language Representations Capture Molecular Structure and Properties. *Research Square*, 2022. 1

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *ICML*, pp. 8346–8356, 2020. 1

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks. In *ICML*, pp. 9323–9332, 2021. 10

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. In *NIPS*, pp. 991–1001, 2017. 3, 10, 20

Kristof Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra. In *ICML*, pp. 9377–9388, 2021. 10

J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.*, 8:3192–3203, 2017. 2

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. 24

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D Infomax improves GNNs for Molecular Property Prediction. *arXiv.org*, 2021. 1, 2, 8, 10, 11, 22

Anthony J. Suckling, M.G. Rumsby, and Michael William Blackburn Bradbury. *Blood-Brain Barrier in Health and Disease*. Ellis Horwood Health Science Series. Ellis Horwood, 1986. 8

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*, 2020a. 6, 27

Ke Sun, Zhouchen Lin, and Zhanxing Zhu. Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes. In *AAAI*, pp. 5892–5899, 2020b. 27

Vladimir Svetnik, Andy Liaw, Christopher Tong, and Ting Wang. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In *MCS*, pp. 334–343, 2004. 10

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? In *NeurIPS*, pp. 6827–6839, 2020. 27

Puja Trivedi, Ekdeep Singh Lubana, Yujun Yan, Yaoqing Yang, and Danai Koutra. Augmentations in Graph Contrastive Learning: Current Methodological Flaws & Towards Better Practices. In *WWW*, pp. 1538–1549, 2022. 27

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv.org*, 2018. 26

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Uszkoreit Kaiser, and Illia Polosukhin. Attention is All You Need. In *NIPS*, pp. 5998–6008, 2017. 4, 21

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep Graph Infomax. In *ICLR*, 2019. 27

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeruIPS*, pp. 16451–16467, 2021. 27

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *BCB*, pp. 429–436, 2019. 1, 10, 11

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.*, 4:279–287, 2022. 10

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *arXiv.org*, 2021. 27

David Weininger. SMILES, A Chemical Language and Information System. *J. Chem. Inf. Comput. Sci.*, 28 (1):31–36, 1988. 1, 10

Boris Weisfeiler and Andrei Leman. A Reduction of a Graph to a Canonical Form and an Algebra Arising During This Reduction. *Nauchno-Technicheskaya Informatsia*, 2(9):12–16, 1968. 20

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. In *EMNLP (Demo)*, pp. 38–45, 2020. 24

Lirong Wu, Haitao Lin, Zhangyang Gao, Cheng Tan, and Stan Z. Li. Self-supervised on Graphs: Contrastive, Generative, or Predictive. *IEEE Trans. Knowl. Data Eng.*, 2022. 27

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. *Chem. Sci.*, 9: 513–530, 2018. 2, 6, 7, 10, 22

Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z. Li. SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation. In *WWW*, pp. 1070–1079, 2022. 27

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What Should Not Be Contrastive in Contrastive Learning. In *ICLR*, 2021. 27

Yaochen Xie, Zhao Xu, Zhengyang Wang, and Shuiwang Ji. Self-Supervised Learning of Graph Neural Networks: A Unified Review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 27

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *ICLR*, 2019. 3, 20

Minghao Xu, Hang Wang, Bingbing Ni, Hongyu Guo, and Jian Tang. Self-supervised Graph-level Representation Learning with Local and Global Structure. In *ICML*, pp. 11548–11558, 2021a. 1, 6, 11, 27

Minkai Xu, Shitong Luo, Yoshua Bengio, Jian Peng, and Jian Tang. Learning Neural Generative Dynamics for Molecular Conformation Generation. In *ICLR*, 2021b. 10

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, 2019. 1

Shuwen Yang, Ziyao Li, Guojie Song, and Lingsheng Cai. Deep Molecular Representation Learning via Fusing Physical and Chemical Information. In *NeurIPS*, pp. 16346–16357, 2021. 10

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do Transformers Really Perform Badly for Graph Representation? In *NeurIPS*, pp. 28877–28888, 2021. 10

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph Contrastive Learning with Augmentations. In *NeurIPS*, pp. 5812–5823, 2020a. 1, 5, 6, 11, 20, 27

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When Does Self-Supervision Help Graph Convolutional Networks? In *ICML*, pp. 10871–10880, 2020b. 27

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph Contrastive Learning Automated. In *ICML*, pp. 12121–12132, 2021. 6, 20, 27

Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.*, 120:143001, 2018. 2

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful Image Colorization. In *ECCV*, pp. 649–666, 2016. 26

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. In *NeurIPS*, pp. 15870–15882, 2021. 10

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep Graph Contrastive Representation Learning. In *GRL+@ICML*, 2020. 5, 27

Yanqiao Zhu, Yichen Xu, Qiang Liu, and Shu Wu. An Empirical Study of Graph Contrastive Learning. In *NeurIPS Datasets and Benchmarks*, 2021a. 5, 27

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph Contrastive Learning with Adaptive Augmentation. In *WWW*, pp. 2069–2080, 2021b. 27