Featurizations Matter:
A Multiview Contrastive Learning Approach to Molecular Pretraining

Anonymous Authors[1]

Abstract

Molecular representation learning, which aims to automate feature learning for molecules, is a vital task in computational chemistry and drug discovery. Despite rapid advances in molecular pretraining models with various types of featurizations, from SMILES strings, 2D graphs to 3D geometry, there is a paucity of research on how to utilize different molecular featurization techniques to obtain better representations. To bridge that gap, we present a novel multiview contrastive learning approach dubbed MEMO in this paper. Our pretraining framework, in particular, is capable of learning from four basic but nontrivial featurizations of molecules and adaptively learning to optimize the combinations of featurization techniques for different downstream tasks. Extensive experiments on a broad range of molecular property prediction benchmarks show that our MEMO outperforms state-of-the-art baselines and also yields reasonable an interpretation of molecular featurizations weights in accordance with chemical knowledge.

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 downstream tasks, spanning from molecular property prediction (Yang et al., 2019) to molecule design (Du et al., 2022). Deep learning models, on the other hand, often rely on a substantial amount of labeled data for proper training, 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).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

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 first pretrains the model by constructing training objectives from large-scale unlabeled datasets and then fine-tunes the learned model on labeled downstream datasets. Motivated by its success, many molecular pretraining models have been developed. To capture chemical structures of molecules, they design several pretraining strategies based on different *molecular featurizations*, which translate chemical information of molecules into numerical representations that can be recognized by machine learning algorithms. For example, some 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 propose contrastive objectives on 2D topology graphs (Hu et al., 2020b; You et al., 2020a; Xu et al., 2021b) or conformations (3D geometry) (Fang et al., 2022). Some recent studies also propose to enrich 2D-topology-based pretraining with 3D geometry information (Stärk et al., 2021; Liu et al., 2022a).

Although considerable progress has been made, these models overlook the impact of molecular featurizations in designing pretraining frameworks, where most present work focuses on only one featurization technique while ignoring the others that are probably essential to some tasks. Some recent studies (Liu et al., 2022a; Stärk et al., 2021) attempt to integrate 3D geometry with 2D topological information, but they assume 3D stereochemical structures are hard to obtain or not available for downstream datasets and still rely on one single featurization in fine-tuning. Moreover, we argue that the utility of different featurizations may vary across downstream tasks. For example, 2D topology information 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 also take downstream tasks into consideration when fine-tuning the model.

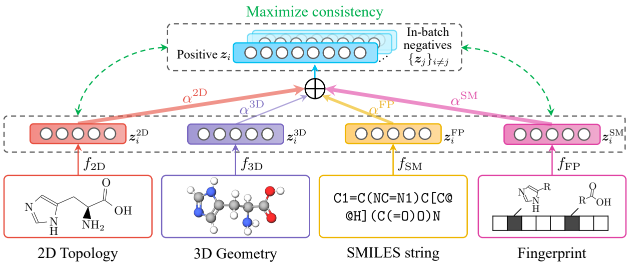

Figure 1. The proposed MEMO model. It begins by constructing four views with four molecule featurizations and then obtains four view-specific embeddings z^* with appropriate encoders. The final embedding z is then computed by taking a weighted average over the four view embeddings, with the coefficients α^* learned using an attention network. Finally, the model is trained using a contrastive objective that maximizes the consistency between view embeddings and the final embedding.

Towards this end, this paper presents a novel MolEcular pretraining framework with Multiview cOntrastive learning, which we term MEMO for brevity. Its graphical illustration is shown in Figure 1. Our proposed model considers four featurization techniques that are widely available or can be easily generated from the raw molecular data: (a) 2D topology graphs, (b) 3D geometry graphs, (c) Morgan fingerprints, and (d) SMILES strings. Then, we construct views based on these four featurizations and leverage different encoders with proper inductive bias to capture their intrinsic information. Following that, the MEMO model dynamically adjusts the contribution of each view through an attention network, which *selectively* extracts information from each view and further allows interpretation analysis of different downstream tasks for domain scientists. After that, we design a novel multiview contrastive pretraining strategy, which trains the model by maximizing the consistency among different views in a self-supervised manner.

We evaluate the effectiveness of our proposed MEMO model on widely-used benchmark datasets including a wide range of molecular property prediction tasks. The experimental results reveal that our work consistently improves non-pretraining baselines while avoiding negative transfer and outperforms existing state-of-the-art molecular pretraining models, achieving a 2.72% 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.

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. The main contributions of this work are three-fold:

- We comprehensively utilize different featurization spaces of molecules and design encoders with appropriate inductive bias corresponding to different representations.
- We propose a novel molecular contrastive pretraining framework that adaptive integrates information from multiple views and provides interpretability for downstream molecular property prediction tasks.
- Extensive experiments conducted on public benchmark datasets validate the effectiveness of our proposed model. MEMO is able to achieve the state-of-the-art across various downstream datasets without negative transfer.

2. Literature Review

This section briefly reviews the progress of molecular machine learning. We first recap four common featurization techniques, followed by molecular representation learning studies. A more broad literature review across the spectrum of self-supervised learning is presented in Appendix C.

2.1. A Brief Recapitulation of Featurization Techniques in Molecules

MEMO utilizes multiple of views to pre-train the model, where each view captures the information of the given data from one aspect. In this work, we consider the following four commonly-used molecular featurization techniques (Ramsundar et al., 2019) and leverage them to construct views for molecules:

- **2D topology graphs** model atoms and bonds as nodes and edges respectively. Featurizing molecules as 2D graphs is arguably a good 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. Related Work on Molecular Machine Learning

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., 2021a).

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 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 1. Compared to the above studies, our proposed MEMO is the only model that can *adaptively* leverage multiple featurizations for both pretraining and fine-tuning stages.

3. The Proposed MEMO Method

In this section, we first formulate the molecule pretraining problem. After that, we discuss the overall pretraining framework. Finally, we introduce the details of view-specific encoders, multiview representation fusion, and contrastive objectives.

3.1. Problem Formulation

Notations. We represent each molecule as an undirected graph, where nodes are atoms and edges describe interatomic bonds. Formally, each graph is denoted as $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{R}, \boldsymbol{X}, \mathsf{E})$, 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 $\mathsf{E} \in \mathbb{R}^{N\times N\times E}$ 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.

Problem statement. As with generic SSL pipelines, the whole framework is divided into two stages, pretraining and fine-tuning. At the first stage, given an unlabeled dataset, we train an encoding function that learns representations with the four featurization techniques. Subsequently, we are provided with several datasets containing molecules with annotations of particular properties. During this fine-tuning phase, we take the weights of the encoders from the pretrained model and then tune the model on specific downstream tasks in a supervised fashion.

3.2. Molecule Pretraining via Multiview Contrastive Learning

We next introduce the MEMO pretraining framework. We first use four view-specific encoders to independently extract information from the four views, each of which is constructed from one of the four featurization strategies. Then, we integrate these four view-specific embeddings to compute a final representation for each molecule through an attention network. Finally, we pretrain the whole model using a multiview contrastive objective.

3.2.1. View-Specific Representation Learning

We leverage four encoders with different inductive bias to capture the intrinsic information in each view. In what follows, the subscript i is used to index the i-th molecule. Due to page limitations, we only discuss the high-level design of each encoder; please refer to Appendix A in the supplementary material for detailed implementations of each view encoder.

Table 1. Comparing MEMO with representative SSL methods for molecular representation learning.

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., 2021b)	✓				✓			
GROVER (Rong et al., 2020)	✓				✓			
GEM (Fang et al., 2022)		✓				✓		
3D Infomax (Stärk et al., 2021)	✓	✓			✓			
GraphMVP (Liu et al., 2022a)	✓	✓			✓			
MEMO (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

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

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

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 $\boldsymbol{z}_i^{\text{3D}} \in \mathbb{R}^D$:

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

Embedding molecular fingerprints. Due to the discrete and extremely sparse nature of fingerprint vectors, we first transform all F feature fields into a dense embedding matrix $\boldsymbol{F}_i \in \mathbb{R}^{F \times D_{\text{F}}}$ 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 $\widehat{\boldsymbol{Z}}_i^{\text{FP}} \in \mathbb{R}^{F \times D_{\text{F}}}$. Following that, we perform sum pooling and use a linear model f_{LIN} to obtain the final fingerprint embedding $z_i^{\text{FP}} \in \mathbb{R}^D$:

$$\widehat{\boldsymbol{Z}}_i^{\text{FP}} = f_{\text{FP}}(\boldsymbol{F}_i), \tag{3}$$

$$\boldsymbol{z}_i^{\text{FP}} = f_{\text{LIN}}\left(\sum_{d=1}^{D_{\text{F}}} \widehat{\boldsymbol{Z}}_{i,d}^{\text{FP}}\right). \tag{4}$$

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 $\boldsymbol{s}_i \in \mathbb{R}^{D_{\text{S}}}$ to get the final embedding $\boldsymbol{z}_i^{\text{SM}} \in \mathbb{R}^D$:

$$\boldsymbol{s}_i = f_{\text{SM}}(\mathbf{S}_i), \tag{5}$$

$$\boldsymbol{z}_i^{\text{SM}} = f_{\text{MLP}}(\boldsymbol{s}_i). \tag{6}$$

3.2.2. Multiview Representation Fusion

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

$$\boldsymbol{z}_i = \sum_{m \in \mathcal{M}} \alpha^m \boldsymbol{z}_i^m, \tag{7}$$

where $\mathcal{M} = \{\text{2D}, \text{3D}, \text{FP}, \text{SM}\}$ 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:

$$\alpha^m = \frac{\exp(w^m)}{\sum_{m' \in \mathcal{M}} \exp(w^{m'})}, \tag{8}$$

$$w^m = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \boldsymbol{q}^\top \cdot \tanh\left(\boldsymbol{W} \frac{\boldsymbol{z}_i^m}{\|\boldsymbol{z}_i^m\|_2} + \boldsymbol{b}\right), \tag{9}$$

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 view embeddings to regularize the scale across different views when computing the intermediate attention scores.

3.2.3. 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., 2020; He et al., 2020; Bachman et al., 2019; Zhu et al., 2020; You et al., 2020a; Zhu et al., 2021a), we leverage the InfoNCE objective, which can be formally written as:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} - \log \frac{e^{\theta(\boldsymbol{z}_i^m, \boldsymbol{z}_i)/\tau}}{\sum_{j \in \mathcal{B}} e^{\theta(\boldsymbol{z}_i^m, \boldsymbol{z}_j)/\tau}} \right], \tag{10}$$

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

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{g(\boldsymbol{x})^\top g(\boldsymbol{y})}{\|g(\boldsymbol{x})\|_2 \|g(\boldsymbol{y})\|_2}. \tag{11}$$

After pretraining the model with the self-supervised objective function $\mathcal{L}$, we fine-tune the model weights of view encoders and the attention multiview fusion module with 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 MEMO able to improve non-pretraining baselines and outperform state-of-the-arts on molecular property prediction tasks?
- **RQ2 (Intrepretation).** Are the learned attention weights of molecular featurizations on different downstream tasks consistent with chemical knowledge?
- **RQ3 (Ablation studies).** How do the multiview fusion 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. These downstream tasks include eight binary classification and four 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%. Basic dataset statistics is summarized in Table 2; for detailed description, we refer readers of interest to Appendix B.

Table 2. Statistics of datasets used in experiments. The first section describes the datasets with 3D information which is used for pretraining; the later two sections describe datasets for fine-tuning.

	Dataset	#Molecules	Avg. #atoms	Avg. #bonds	#Tasks	Avg. degree
	GEOM-Drug	304,466	44.40	46.40	—	2.09
Classification	BBBP	2,039	24.06	25.95	1	2.16
	Tox21	7,831	18.57	19.29	12	2.08
	ToxCast	8,576	18.78	19.26	617	2.05
	SIDER	1,427	33.64	35.36	27	2.10
	ClinTox	1,477	26.16	27.88	2	2.13
	MUV	93,087	24.23	26.28	17	2.17
	HIV	41,127	25.51	27.47	1	2.15
	BACE	1,513	34.09	36.86	1	2.16
Regression	ESOL	1,128	13.30	13.69	1	2.06
	Lipophilicity	4,200	27.04	29.50	1	2.18
	Malaria	9,999	30.36	33.20	1	2.19
	CEP	29,978	27.66	33.39	1	2.41

Implementation details. In the GEOM-Drugs dataset, 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.

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 regression tasks, we measure the performance in Root Mean Squared Error (RMSE), 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).

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, Con-

textPred (Hu et al., 2020b), GraphLoG (Xu et al., 2021b), GraphCL (You et al., 2020a), and JOAO (You et al., 2021).
- Molecular SSL models: GROVER-Contextual (G-Contextual), GROVER-Motif (G-Motif) (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 GIN model (Xu et al., 2019) 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).

4.2. Main Results on Molecular Property Prediction

The performance of eight low-data molecular property prediction tasks is summarized in Table 3. Generally, it can be found from the table that our MEMO shows strong empirical performance across all eight downstream datasets, delivering seven out of eight state-of-the-art results and acquiring a 2.72% absolute improvement on average. The outstanding results validate the superiority of our proposed model.

We make other observations as follows. Firstly, MEMO 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, our model also achieves much better performance than other state-of-the-art baselines on average, with an absolute improvement of up to 3.4%. It is worth mentioning that, on some challenging datasets (e.g., Tox21, HIV, and ToxCast), while other models *barely* improve the non-pretraining baselines, our model nevertheless attains promising performance increments, which demonstrates the effectiveness of leveraging multiple featurization techniques.

4.3. More Experiments on Molecular Property Regression

We further conduct experiments on four additional regression tasks for molecular property prediction, where the results are presented in Table 4. It can be clearly seen from the table that our MEMO considerably improves the performance of baselines on three datasets and achieves similar performance to the baseline approaches on the Lipophilicity dataset, which once again verifies the effectiveness of our framework and demonstrates the importance of integrating different molecular featurization techniques.

[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.

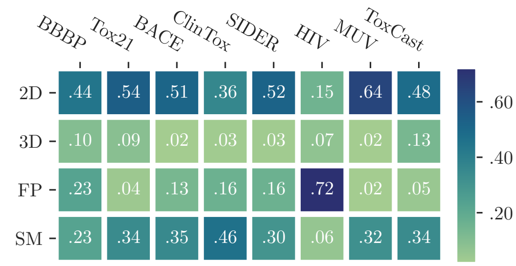

Figure 2. Visualizing the learned attention weights on eight molecular property prediction datasets.

4.4. Interpretation and Analysis

In order to analyze the correlation between tasks and featurization techniques, we visualize the attention weights α learned on different downstream tasks in Figure 2. 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 confor-

Table 3. 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.
—	65.4±2.4	74.9±0.8	61.6±1.2	58.0±2.4	58.8±5.5	71.0±2.5	75.3±0.5	72.6±4.9	67.21
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
G-Contextual	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
G-Motif	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
MEMO	**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**

Table 4. Additional results on four molecular property regression tasks in terms of Root-Mean-Square Error (RMSE). The lowest prediction error is highlighted in **boldface**.

Pretraining	ESOL	Lipophilicity	Malaria	CEP	Avg.
—	1.178±0.044	0.744±0.007	1.127±0.003	1.254±0.030	1.07559
AttrMask	1.112±0.048	0.730±0.004	1.119±0.014	1.256±0.000	1.05419
ContextPred	1.196±0.037	**0.702±0.020**	1.101±0.015	1.243±0.025	1.06059
JOAO	1.120±0.019	0.708±0.007	1.145±0.010	1.293±0.003	1.06631
GraphMVP	1.091±0.021	0.718±0.016	1.114±0.013	1.236±0.023	1.03968
MEMO	**0.984±0.034**	0.707±0.001	**1.093±0.009**	**1.101±0.007**	**0.97125**

mation 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 have 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.
- Fingerprints are representations 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. This can account for relatively smaller attention weights of fingerprints in Figure 2.

4.5. Ablation Studies

Finally, we conduct ablation studies on the multiview fusion 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.

- **MEMO–Max** removes the attention network in the multiview fusion module in Equation (7) and simply uses max pooling to combine view embeddings.

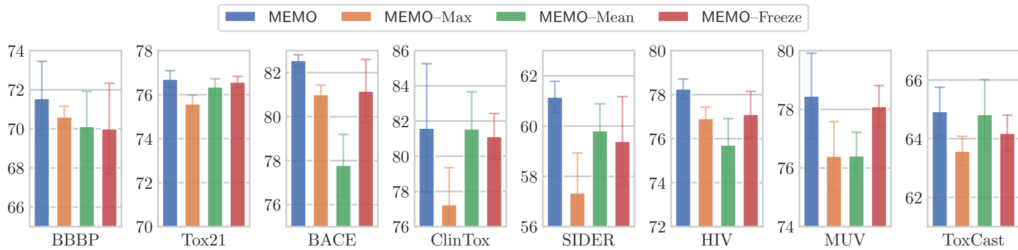

Figure 3. Ablation studies on multiview fusion and the fine-tuning strategy.

- **MEMO–Mean** modifies the multiview fusion module by taking the average over view embeddings.
- **MEMO–Freeze** does not fine-tune the multiview fusion module but instead uses the frozen weights of the pretrained model.

We report the performance of the three model variants in Figure 3. It is seen that all model variants achieve downgraded performance, which empirically rationalizes the design choice of our multiview contrastive pretraining framework. Specifically, the performance of two variants, MEMO–Max and MEMO–Mean, without attention fusion mechanisms of multiview representations is inferior to that of MEMO, demonstrating the necessity of adaptively combining information from multiple views. In addition, MEMO–Freeze occasionally obtains better performance than the two other variants, which indicates that our proposed attention network is able to select information from different views. It does not, however, fine-tune the contribution of different featurizations with downstream datasets, where the optimal combination might differ, resulting in performance deterioration.

5. Conclusion

This paper has developed a novel pretraining framework MEMO with multiview contrastive learning for molecular data. Our proposed model constructs multiple views with different molecular featurizations, leverages proper encoders to capture intrinsic information within each view, and designs a multiview contrastive objective to adaptively distill information from each view. Extensive experiments conducted on public molecular property prediction benchmarks show that our pretraining framework is able to transfer knowledge from large unlabeled datasets to a wide range of low-data downstream datasets. The interpretation of the learned model weights is also in line with prior chemical knowledge.

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 as well as the task-featurization correlation.

References

AIDS Antiviral Screen Data. URL https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data. 14

Tox21 Data Challenge 2014, 2014. URL https://tripod.nih.gov/tox21/challenge/. 14

Axelrod, S. and Gómez-Bombarelli, R. GEOM, Energy-Annotated Molecular Conformations for Property Prediction and Molecular Generation. *Sci. Data*, 9(1):185, 2022. 5, 14

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views. In *NeurIPS*, pp. 15509–15519, 2019. 5, 15

Baell, J. B. and Holloway, G. A. 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. 6

Bahdanau, D., Cho, K., and Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2015. 4

Batzner, S. L., Smidt, T. E., Sun, L., Mailoa, J. P., Kornbluth, M., Molinari, N., and Kozinsky, B. SE(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *arXiv.org*, 2021. 3

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, 2003. 1, 15

Brandstetter, J., Hesselink, R., van der Pol, E., Bekkers, E. J., and Welling, M. Geometric and Physical Quantities Improve E(3) Equivariant Message Passing. In *ICLR*, 2022. 3

Breiman, L. Random Forests. *Mach. Learn.*, 45(1):5–32, 2001. 3

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *NeurIPS*, pp. 1877–1901, 2020. 15

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.*, 25(2):64–73, 1985. 3

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, pp. 139–156, 2018. 15

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *NeurIPS*, pp. 9912–9924, 2020. 15

Chen, R. T. Q. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *KDD*, pp. 785–794, 2016. 3

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, pp. 1597–1607, 2020. 5, 15

Chen, X. and He, K. Exploring Simple Siamese Representation Learning. In *CVPR*, pp. 15745–15753, 2021. 15

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv.org*, 2020. 1, 3, 4

Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.*, 44(3): 1000–1005, 2004. 14

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pp. 4171–4186, 2019. 3, 4, 15

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 15

Du, W., Zhang, H., Du, Y., Meng, Q., Chen, W., Shao, B., and Liu, T.-Y. Equivariant Vector Field Network for Many-Body System Modeling. *arXiv.org*, 2021. 3

Du, Y., Fu, T., Sun, J., and Liu, S. MolGenSurvey: A Systematic Survey in Machine Learning Models for Molecule Design. *arXiv.org*, 2022. 1

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.*, 4:127–134, 2022. 1, 3, 4

Fuchs, F., Worrall, D. E., Fischer, V., and Welling, M. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. In *NeurIPS*, pp. 1970–1981, 2020. 3

Gage, P. A New Algorithm for Data Compression. *C Users J.*, 12 (2):23–38, 1994. 13

Gamo, F.-J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.-L., Vanderwall, D. E., Green, D. V. S., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R., and Garcia-Bustos, J. F. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature*, 465(7296):305–310, 2010. 14

Gao, T., Yao, X., and Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, pp. 6894–6910, 2021. 15

Gasteiger, J., Becker, F., and Günnemann, S. GemNet: Universal Directional Graph Neural Networks for Molecules. In *NeurIPS*, pp. 6790–6802, 2021. 3

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.*, 40(D1):D1100–D1107, 2011. 5, 14

Gayvert, K. M., Madhukar, N. S., and Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem. Biol.*, 23(10):1294–1301, 2016. 14

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*, 2018. 15

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *ICML*, pp. 1263–1272, 2017. 3

Glem, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs*, 9(3):199–204, 2006. 2, 14

Grimme, S. 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. 14

Hachmann, J., Olivares-Amaya, R., Atahan-Evrenk, S., Amador-Bedolla, C., Sánchez-Carrera, R. S., Gold-Parker, A., Vogt, L., Brockway, A. M., and Aspuru-Guzik, A. 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. 5, 14

Hamilton, W. L., Ying, Z., and Leskovec, J. Inductive Representation Learning on Large Graphs. In *NIPS*, pp. 1024–1034, 2017. 5, 15

Hassani, K. and Khasahmadi, A. H. Contrastive Multi-View Representation Learning on Graphs. In *ICML*, pp. 4116–4126, 2020. 15

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, pp. 9726–9735, 2020. 5, 15

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2022. 15

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *NeurIPS*, pp. 22118–22133, 2020a. 5, 14

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for Pre-training Graph Neural Networks. In *ICLR*, 2020b. 1, 3, 4, 6, 13, 15

Hu, Z., Dong, Y., Wang, K., Chang, K.-W., and Sun, Y. GPT-GNN: Generative Pre-Training of Graph Neural Networks. In *KDD*, pp. 1857–1867, 2020c. 5, 15

Jing, L. and Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):4037–4058, 2021. 1, 15

Kipf, T. N. and Welling, M. Variational Graph Auto-Encoders. In *BDL@NIPS*, 2016. 15

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-Thought Vectors. In *NIPS*, pp. 3294–3302, 2015. 15

Klicpera, J., Groß, J., and Günnemann, S. Directional Message Passing for Molecular Graphs. In *ICLR*, 2020. 3

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.*, 44 (D1):D1075–D1079, 2016. 14

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*, 2020. 15

Landrum, G., Tosco, P., Kelley, B., Ric, sriniker, gedeck, Vianello, R., NadineSchneider, Kawashima, E., Dalke, A., N, D., Cosgrove, D., Cole, B., Swain, M., Turk, S., AlexanderSavelyev, Jones, G., Vaucher, A., Wójcikowski, M., Take, I., Probst, D., Ujihara, K., Scalfani, V. F., guillaume godin, Pahl, A., Berenger, F., JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_03_2 (q1 2022) release, 2022. 5, 14

Larsson, G., Maire, M., and Shakhnarovich, G. Colorization as a Proxy Task for Visual Understanding. In *CVPR*, pp. 840–849, 2017. 15

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, pp. 7871–7880, 2020. 15

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., and Song, S. Learn Molecular Representations From Large-Scale Unlabeled Molecules for Drug Discovery. *arXiv.org*, 2020. 3

Lin, S., Zhou, P., Hu, Z.-Y., Wang, S., Zhao, R., Zheng, Y., Lin, L., Xing, E. P., and Liang, X. Prototypical Graph Contrastive Learning. *arXiv.org*, 2021. 16

Liu, S., Demirel, M. F., and Liang, Y. N-Gram Graph: Simple Unsupervised Representation for Graphs, with Applications to Molecules. In *NeurIPS*, pp. 8464–8476, 2019a. 3

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pretraining Molecular Graph Representation with 3D Geometry. In *ICLR*, 2022a. 1, 3, 4, 5, 6, 13, 14

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv.org*, 2019b. 4, 14, 15

Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B., and Ji, S. Spherical Message Passing for 3D Graph Networks. *arXiv.org*, 2021. 3

Liu, Y., Pan, S., Jin, M., Zhou, C., Zheng, Y., Xia, F., and Yu, P. S. Graph Self-Supervised Learning: A Survey. *IEEE Trans. Knowl. Data Eng.*, 2022b. 15

Manoharan, P., Vijayan, R., and Ghoshal, N. 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. 7

Martins, I. F., Teixeira, A. L., Pinheiro, L., and Falcao, A. O. A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.*, 52(6):1686–1697, 2012. 14

Meyer, J. G., Liu, S., Miller, I. J., Coon, J. J., and Gitter, A. Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests. *J. Chem. Inf. Model.*, 59(10):4438–4449, 2019. 3

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, pp. 3111–3119, 2013. 15

Morgan, H. L. 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. 2, 14

Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.*, 27(2):82–85, 1987. 3

Noroozi, M. and Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV*, pp. 69–84, 2016. 15

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. Context Encoders: Feature Learning by Inpainting. In *CVPR*, pp. 2536–2544, 2016. 15

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep Contextualized Word Representations. In *NAACL-HLT*, pp. 2227–2237, 2018. 15

Purushwalkam, S. and Gupta, A. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases. In *NeurIPS*, pp. 3407–3418, 2020. 15

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI Blog, 2018. 15

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI Blog, 2019. 15

Ramsundar, B., Eastman, P., Walters, P., and Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media, 2019. 2

Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M. T., Wambaugh, J. F., Knudsen, T. B., Kancherla, J., Mansouri, K., Patlewicz, G., Williams, A. J., Little, S. B., Crofton, K. M., and Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.*, 29 (8):1225–1251, 2016. 14

Riniker, S. and Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015. 5, 14

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010. 3, 5, 14, 15

Rohrer, S. G. and Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.*, 49(2):169–184, 2009. 14

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *NeurIPS*, pp. 12559–12571, 2020. 3, 4, 6

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *ICML*, pp. 8346–8356, 2020. 1

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) Equivariant Graph Neural Networks. In *ICML*, pp. 9323–9332, 2021. 3

Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. In *NIPS*, pp. 991–1001, 2017. 3, 4, 13

Schütt, K., Unke, O. T., and Gastegger, M. Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra. In *ICML*, pp. 9377–9388, 2021. 3

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

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. 3D Infomax improves GNNs for Molecular Property Prediction. *arXiv.org*, 2021. 1, 3, 4, 14

Suckling, A. J., Rumsby, M., and Bradbury, M. W. B. *Blood-Brain Barrier in Health and Disease*. Ellis Horwood Health Science Series. Ellis Horwood, 1986. 6

Sun, F.-Y., Hoffmann, J., Verma, V., and Tang, J. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*, 2020a. 5, 16

Sun, K., Lin, Z., and Zhu, Z. Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes. In *AAAI*, pp. 5892–5899, 2020b. 15

Svetnik, V., Liaw, A., Tong, C., and Wang, T. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In *MCS*, pp. 334–343, 2004. 3

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What Makes for Good Views for Contrastive Learning? In *NeurIPS*, pp. 6827–6839, 2020. 15

Trivedi, P., Lubana, E. S., Yan, Y., Yang, Y., and Koutra, D. Augmentations in Graph Contrastive Learning: Current Methodological Flaws & Towards Better Practices. In *WWW*, pp. 1538–1549, 2022. 16

van den Oord, A., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv.org*, 2018. 15

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, U., and Polosukhin, I. Attention is All You Need. In *NIPS*, pp. 5998–6008, 2017. 4, 13

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. Deep Graph Infomax. In *ICLR*, 2019. 15

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeruIPS*, pp. 16451–16467, 2021. 15

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *BCB*, pp. 429–436, 2019. 1, 3, 4

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.*, 4:279–287, 2022. 3

Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and Feichtenhofer, C. Masked Feature Prediction for Self-Supervised Visual Pre-Training. *arXiv.org*, 2021. 15

Weininger, D. SMILES, A Chemical Language and Information System. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988. 1, 3

Weisfeiler, B. and Leman, A. A Reduction of a Graph to a Canonical Form and an Algebra Arising During This Reduction. *Nauchno-Technicheskaya Informatsia*, 2(9):12–16, 1968. 13

Wu, L., Lin, H., Gao, Z., Tan, C., and Li, S. Z. Self-supervised on Graphs: Contrastive, Generative, or Predictive. *IEEE Trans. Knowl. Data Eng.*, 2022. 15

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.*, 9: 513–530, 2018. 3, 5, 6, 14

Xia, J., Wu, L., Chen, J., Hu, B., and Li, S. Z. SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation. In *WWW*, pp. 1070–1079, 2022. 16

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What Should Not Be Contrastive in Contrastive Learning. In *ICLR*, 2021. 15

Xie, Y., Xu, Z., Wang, Z., and Ji, S. Self-Supervised Learning of Graph Neural Networks: A Unified Review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 15

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How Powerful are Graph Neural Networks? In *ICLR*, 2019. 4, 6, 13

Xu, M., Luo, S., Bengio, Y., Peng, J., and Tang, J. Learning Neural Generative Dynamics for Molecular Conformation Generation. In *ICLR*, 2021a. 3

Xu, M., Wang, H., Ni, B., Guo, H., and Tang, J. Self-supervised Graph-level Representation Learning with Local and Global Structure. In *ICML*, pp. 11548–11558, 2021b. 1, 4, 6, 16

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, 2019. 1

Yang, S., Li, Z., Song, G., and Cai, L. Deep Molecular Representation Learning via Fusing Physical and Chemical Information. In *NeurIPS*, pp. 16346–16357, 2021. 3

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do Transformers Really Perform Badly for Graph Representation? In *NeurIPS*, pp. 28877–28888, 2021. 3

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph Contrastive Learning with Augmentations. In *NeurIPS*, pp. 5812–5823, 2020a. 1, 4, 5, 6, 13, 16

You, Y., Chen, T., Wang, Z., and Shen, Y. When Does Self-Supervision Help Graph Convolutional Networks? In *ICML*, pp. 10871–10880, 2020b. 15

You, Y., Chen, T., Shen, Y., and Wang, Z. Graph Contrastive Learning Automated. In *ICML*, pp. 12121–12132, 2021. 6, 13, 16

Zhang, L., Han, J., Wang, H., Car, R., and E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.*, 120:143001, 2018. 1

Zhang, R., Isola, P., and Efros, A. A. Colorful Image Colorization. In *ECCV*, pp. 649–666, 2016. 15

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. In *NeurIPS*, pp. 15870–15882, 2021. 3

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. Deep Graph Contrastive Representation Learning. In *GRL+@ICML*, 2020. 5, 16

Zhu, Y., Xu, Y., Liu, Q., and Wu, S. An Empirical Study of Graph Contrastive Learning. In *NeurIPS Datasets and Benchmarks*, 2021a. 5, 15

Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., and Wang, L. Graph Contrastive Learning with Adaptive Augmentation. In *WWW*, pp. 2069–2080, 2021b. 16

A. Implementation of View Encoders

In this section, we introduce the detailed implementation of the four view encoders. We denote the representation for node (atom) v_i as $\boldsymbol{h}_i$ and the representation at the graph (molecule) level as $\boldsymbol{z}$. As each encoder is independent to each other, we omit the superscript representing the specific view $m \in \{2\text{D}, 3\text{D}, \text{FP}, \text{SM}\}$ is clear for notation simplicity. Also, for clarity, when the context is clear, we omit the subscript j that indexes the molecule.

Embedding 2D graphs. Graph Isomorphism Network (GIN) (Xu et al., 2019) is a simple and effective model to learn discriminative graph representations, which is proved to have the same representational power as the Weisfeiler-Lehman test (Weisfeiler & Leman, 1968). Since GIN has been widely adopted for 2D graph representation learning (Hu et al., 2020b; You et al., 2021; 2020a), we leverage a GIN model to obtain the representations for the 2D molecular graphs. Recall that each molecule is represented as $\mathcal{G} = (\boldsymbol{A}, \boldsymbol{X}, \mathsf{E})$, where $\boldsymbol{A}$ is the adjacency matrix, $\boldsymbol{X}$ and E are features for atoms and bonds respectively. The layer-wise propagation rule of GIN can be written as:

$$\boldsymbol{h}_i^{(k+1)} = f_{\text{atom}}^{(k+1)}\left(\boldsymbol{h}_i^{(k)} + \sum_{j\in\mathcal{N}(i)}\left(\boldsymbol{h}_j^{(k)} + f_{\text{bond}}^{(k+1)}(E_{ij})\right)\right), \tag{12}$$

where the input features $\boldsymbol{h}_i^{(0)} = \boldsymbol{x}_i$, $\mathcal{N}(i)$ is the neighborhood set of atom v_i, and $f_{\text{atom}}, f_{\text{bond}}$ are two MultiLayer Perceptron (MLP) layers for transforming atoms and bonds features, respectively. By stacking K layers, we can incorporate K-hop neighborhood information into each center atom in the molecular graph. Then, we take the output of the last layer as the atom representations and further use the mean pooling to get the graph-level molecular representation:

$$\boldsymbol{z}^{2\text{D}} = \frac{1}{N}\sum_{i\in\mathcal{V}} \boldsymbol{h}_i^{(K)}. \tag{13}$$

Embedding 3D graphs. Following GraphMVP (Liu et al., 2022a), we use the SchNet (Schütt et al., 2017) as the encoder for the 3D geometry graphs. SchNet models message passing in the 3D space as continuous-filter convolutions, which is composed of a series of hidden layers, given as follows:

$$\boldsymbol{h}_i^{(k+1)} = f_{\text{MLP}}\left(\sum_{j=1}^{N} f_{\text{FG}}(\boldsymbol{h}_j^{(t)}, \boldsymbol{r}_i, \boldsymbol{r}_j)\right) + \boldsymbol{h}_i^{(t)}, \tag{14}$$

where the input $\boldsymbol{h}_i^{(0)} = \boldsymbol{a}_i$ is an embedding dependent on the type of atom v_i, $f_{\text{FG}}(\cdot)$ denotes the filter-generating network. To ensure rotational invariance of a predicted property, the message passing function is restricted to depend only on rotationally invariant inputs such as distances, which satisfying the energy properties of rotational equivariance by construction. Moreover, SchNet adopts radial basis functions to avoid highly correlated filters. The filter-generating network is defined as follow:

$$\begin{aligned} f_{\text{FG}}(\boldsymbol{x}_j, \boldsymbol{r}_i, \boldsymbol{r}_j) &= \boldsymbol{x}_j \cdot e_k(\boldsymbol{r}_i - \boldsymbol{r}_j) \\ &= \boldsymbol{x}_j \cdot \exp(-\gamma \| \|\boldsymbol{r}_i - \boldsymbol{r}_j\|_2 - \mu\|_2^2). \end{aligned} \tag{15}$$

Similarly, for non-quantum properties prediction concerned in this work, we take the average of the node representations as the 3D molecular embedding:

$$\boldsymbol{z}^{3\text{D}} = \frac{1}{N}\sum_{i\in\mathcal{V}} \boldsymbol{h}_i^{(K)}, \tag{16}$$

where K is the number of hidden layers.

Embedding fingerprints. Due to the discrete and extremely sparse nature of fingerprint vectors, we first transform all F feature fields into a dense embedding matrix $\boldsymbol{F} \in \mathbb{R}^{F\times D_{\text{F}}}$ via embedding lookup. Then, we use a multi-head self-attention network (Vaswani et al., 2017) to model the interaction among those feature fields. Specifically, we first transform each feature into a new embedding space as:

$$\boldsymbol{Q}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\text{Q}}^{(h)}, \tag{17}$$

$$\boldsymbol{K}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\text{K}}^{(h)}, \tag{18}$$

$$\boldsymbol{V}^{(h)} = \boldsymbol{F}\boldsymbol{W}_{\text{V}}^{(h)}, \tag{19}$$

where the three linear transformation matrices $\boldsymbol{W}_{\text{Q}}^{(h)}, \boldsymbol{W}_{\text{K}}^{(h)}, \boldsymbol{W}_{\text{V}}^{(h)} \in \mathbb{R}^{D_{\text{F}}\times D/H}$ parameterize the query, key, and value transformations for the h-th attention head, respectively. Following that, we compute the attention scores among all feature pairs and then linearly combine the value matrix from all H attention heads:

$$\boldsymbol{W}_{\text{A}}^{(h)} = \text{softmax}\left(\frac{\boldsymbol{Q}^{(h)}(\boldsymbol{K}^{(h)})^\top}{\sqrt{D_{\text{H}}}}\right), \tag{20}$$

$$\widehat{\boldsymbol{Z}} = \left[\boldsymbol{W}_{\text{A}}^{(1)}\boldsymbol{V}^{(1)}; \boldsymbol{W}_{\text{A}}^{(2)}\boldsymbol{V}^{(2)}; \ldots; \boldsymbol{W}_{\text{A}}^{(H)}\boldsymbol{V}^{(H)}\right], \tag{21}$$

Finally, we perform sum pooling on the resulting embedding matrix $\widehat{\boldsymbol{Z}} \in \mathbb{R}^{F\times D_{\text{F}}}$ and use a linear model f_{LIN} to obtain the final fingerprint embedding $\boldsymbol{z}^{\text{FP}} \in \mathbb{R}^{D}$:

$$\boldsymbol{z}^{\text{FP}} = f_{\text{LIN}}\left(\sum_{d=1}^{D_{\text{F}}} \widehat{\boldsymbol{Z}}_d\right). \tag{22}$$

Embedding SMILES strings. Given ASCII-encoded SMILES strings, we first tokenize them with the Byte-Pair Encoder (BPE) tokenizer (Gage, 1994), which strikes a balance among character- and word-level representations and allows to handle large vocabularies in molecular corpora. Specifically, BPE finds the best word segmentation by iteratively and greedily merging frequent pairs of characters. In our implementation, we use a max vocabulary size of 52K tokens for both pretraining and downstream datasets.

After tokenization, we first pretrain a RoBERTa (Liu et al., 2019b) model on the pretraining dataset with the masking language model as the sole training objective, as SMILES strings do not possess sequential relationships. To be specific, 15% tokens in a SMILES string are randomly selected and replaced with the special token [MASK]. We also insert a special token [CLS] to each string to represent the whole string. Then, the training objective function is to independently predict the original tokens given the output on masked tokens. Finally, the representation of the [CLS] token is regarded as the molecular embedding.

After pretraining the RoBERTa backbone, we freeze its parameters and leverage an additional MLP layer on top of each molecular embedding to obtain the final representation for each SMILES string. This strategy improves memory efficiency and thus enables larger batch sizes for contrasive pretraining.

B. Dataset Description

In this section, we briefly introduce the datasets used for pretraining and fine-tuning, as well as details of dataset prepossessing.

B.1. Pretraining Datasets

We choose GEOM-Drugs[2] (Axelrod & Gómez-Bombarelli, 2022) as the pre-training dataset, which contains high-quality conformers for 304,466 mid-sized organic molecules with experimental data. The conformer information in GEOM-Drugs is generated using the CREST (Grimme, 2019) program, which provides reliable and accurate structure generation. Note that atoms usually have multiple conformations resulting in potentially different chemical properties. In this work, we focus on the conformations of the lowest energy, since they are more likely to occur naturally (Stärk et al., 2021; Liu et al., 2022a). Specifically, following GraphMVP (Liu et al., 2022a), we utilize the top five conformers for each molecule in pretraining.

B.2. Fine-tuning Datasets

For fine-tuning, we use twelve datasets collected from MoleculeNet[3] (Wu et al., 2018), which target on different properties and distinct tasks. These properties can be divided into three main categories: physical chemistry, biophysics, and physiology.

[2] https://github.com/learningmatter-mit/geom
[3] https://github.com/deepchem/deepchem

Physical chemistry. ESOL (Delaney, 2004) consists of water solubility data recording whether molecules are water-soluble. The Lipophilicity dataset is a subset of ChEMBL (Gaulton et al., 2011) measuring the molecule octanol/water distribution coefficient. The CEP dataset is a subset of the Havard Clean Energy Project (CEP) (Hachmann et al., 2011), which estimates the organic photovoltaic efficiency.

Biophysics. The HIV dataset (AID) is introduced by Drug Therapeutics Program (DTP) AIDS Antiviral Screen, which tests the molecular ability to inhibit HIV replication. The Maximum Unbiased Validation (MUV) group (Rohrer & Baumann, 2009) is another benchmark dataset selected from PubChem BioAssay by applying a refined nearest neighbor analysis. The BACE dataset provides qualitative binding results for a set of inhibitors of human β-secretase 1 (BACE-1). The Malaria dataset (Gamo et al., 2010) assesses the drug efficacy in inhibiting parasites that cause malaria.

Physiology. The Blood–brain barrier penetration (BBBP) dataset (Martins et al., 2012) models the barrier permeability of molecules targeting central nervous system. Tox21 (Tox, 2014), ToxCast (Richard et al., 2016), and ClinTox (Gayvert et al., 2016) are all related to the toxicity of molecular compounds. The Side Effect Resource (SIDER) (Kuhn et al., 2016) is a dataset measuring the adverse drug reactions of 27 system organ classes of marketed drugs.

B.3. Dataset Preprocessing

For classification tasks, we leverage atom types and chirality tags as atom attributes, while the type and direction of the bond are corresponding bond attributes. Both the atom and bond attributes are expressed in the form of discrete indices without further embedding. For regression tasks, we first transform discrete atom and bond attributes through learnable embedding lookup layers following OGB (Hu et al., 2020a).

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 molecular fingerprints, which is roughly equivalent to the ECFP4 scheme (Rogers & Hahn, 2010).

Constructing fingerprints. Morgan fingerprints (Morgan, 1965; Glem et al., 2006) encode molecules in fixed-length binary strings, with bits indicating presence or absence of specific substructures. The algorithm assigns an initial identifier to each non-hydrogen atom according to a set of atomic invariants, iteratively updates the identifiers among neighborhood atoms within certain hops, and encodes the identifiers using a hash function. After hashing all of these identifiers into a fixed-length binary string, the representation provides information on topological characteristics of the molecule.

In our implementation, we set the diameter of neighborhood to 2, the length of fingerprints to 1024, and follow the default configuration of ECFP4 (Rogers & Hahn, 2010), which uses the following connectivity invariants:

- The atomic number
- The number of heavy (non-hydrogen) neighbor atoms
- The number of attached hydrogens
- The formal charge
- Atom isotopes
- Whether the atom is part of at least one ring

C. More Related Work

The following section provides a more broad literature review across the spectrum of self-supervised representation learning.

C.1. Self-Supervised Representation Learning on Visual and Natural Language Data

A SSL model trains itself by learning a part of the input from another through pretext tasks. Depending on the pretext task, the existing SSL studies can be divided into three main categories.

Early SSL work studies *predictive training* on pseudo-labels directly computed from the raw data. In Computer Vision (CV) domains, typical pretext tasks include image inpainting (Pathak et al., 2016), rearranging shuffled image patches (Noroozi & Favaro, 2016), colorizing grayscale images (Zhang et al., 2016; Larsson et al., 2017), recognizing geometric transformations (Gidaris et al., 2018), and predicting cluster assignments (Caron et al., 2018). In Natural Language Processing (NLP), word2vec (Mikolov et al., 2013) popularizes this paradigm by proposing Continuous Bag-Of-Words (CBOW) and skip-gram models for predicting center and neighboring words, respectively. Other exemplary work includes Kiros et al. (2015) that predicts neighborhood sentences and BART (Lewis et al., 2020) that recovers sentence permutation.

The second group of SSL is *contrastive* learning, which seeks to maximize the agreement of embeddings in the latent space under stochastic data augmentations by contrasting positive and negative samples (Jing & Tian, 2021). It has revolutionized unsupervised representation learning in recent years (van den Oord et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020; Caron et al., 2020; Chen & He, 2021; Gao et al., 2021) and has been witnessed to perform on par with its supervised counterparts (He et al., 2020; Chen et al., 2020). A key success to contrastive models is to leverage strong data augmentations that induce invariance irrelevant to properties of the end tasks (Xiao et al., 2021; Tian et al., 2020; Purushwalkam & Gupta, 2020; von Kügelgen et al., 2021).

The third line of development focuses on *generative modeling* of input data. Its core idea is to randomly remove a portion of data and train the model to recover the removed content. This so-called masked language modeling and its autoregressive counterparts are first pioneered in the NLP community (Bengio et al., 2003; Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2020; Radford et al., 2018; 2019; Brown et al., 2020) and have since gained increasing popularity in the CV domain (Dosovitskiy et al., 2021; He et al., 2022; Wei et al., 2021). Unlike contrastive learning, generative approaches do not rely on curated data augmentations. It has been reported that they scale well and generalize to different downstream tasks (Devlin et al., 2019; Brown et al., 2020; He et al., 2022).

C.2. Graph Self-Supervised Representation Learning

Analogous to the above studies on visual and natural language data, SSL approaches in the graph domain can also be organized into the same three categories. Due to the rapid development of graph SSL, we only review the most representative studies in each group. Readers may refer to recent surveys (Wu et al., 2022; Xie et al., 2022; Liu et al., 2022b) for comprehensive reviews and Zhu et al. (2021a) for a benchmarking study.

Firstly, the pioneering *predictive* model Hu et al. (2020b) explores four strategies at both node and graph levels, including masked attribute prediction, context prediction, supervised attribute prediction, and structural similarity prediction. You et al. (2020b) study three SSL tasks through a multi-task framework to enable predictive training of graph-structured data. M3S (Sun et al., 2020b) explores the use of cluster assignments (Caron et al., 2018) as pseudo-labels and proposes a self-training framework that incrementally adds high-confident nodes to the labeled dataset.

The second group of work studies *generative* training. GraphSAGE (Hamilton et al., 2017) performs the link prediction task to reconstruct the graph structure in a once-for-all manner, similar to graph autoencoders (Kipf & Welling, 2016). GPT-GNN (Hu et al., 2020c) proposes to perform node and edge reconstruction iteratively.

Lastly, along the line of graph *contrastive* learning, some investigate contrasting modes for graph data, typical work of which includes cross-scale contrasting (Veličković et al., 2019; Hassani & Khasahmadi, 2020), same-scale contrasting (Sun et al., 2020a; Zhu et al., 2020; You et al., 2020a), and hierarchical contrasting (Xu et al., 2021b; Lin et al., 2021). Another line of work investigates data augmentations. GraphCL (You et al., 2020a) proposes four heuristic augmentation schemes including edge dropping, node dropping, attribute masking, and subgraph cropping; its follow-up JOYO (You et al., 2021) proposes to learn the augmentation priors via bi-level optimization. GCA (Zhu et al., 2021b) proposes adaptive augmentation that better preserves important semantics and structures of the underlying graph. SimGRACE (Xia et al., 2022) eschews the need of explicit augmentation; Trivedi et al. (2022) propose content-aware augmentation to avoid corrupting task-relevant information.