Molecular Graph Contrastive Learning with Parameterized Explainable Augmentations

Yingheng Wang^{1,*}, Yaosen Min^{2,*}, Erzhuo Shao¹, and Ji Wu^{1,†}

¹Department of Electronic Engineering, Tsinghua University

²Institute of Interdisciplinary Information Sciences, Tsinghua University

¹ { wangyh20, shaoez20 } @mails.tsinghua.edu.cn, wuji_ee@tsinghua.edu.cn

²minys18@mails.tsinghua.edu.cn

ABSTRACT

Learning generalizable, transferable, and robust representations for molecule data has always been a challenge. The recent success of contrastive learning (CL) for self-supervised graph representation learning provides a novel perspective to learn molecule representations. The most prevailing graph CL framework is to maximize the agreement of representations in different augmented graph views. However, existing graph CL frameworks usually adopt stochastic augmentations or schemes according to pre-defined rules on the input graph to obtain different graph views in various scales (e.g. node, edge, and subgraph), which may destroy topological semantics and domain prior in molecule data, leading to sub-optimal performance. Therefore, designing parameterized, learnable, and explainable augmentation is quite necessary for molecular graph contrastive learning. A well-designed parameterized augmentation scheme can preserve chemically meaningful structural information and intrinsically essential attributes for molecule graphs, which helps to learn representations that are insensitive to perturbation on unimportant atoms and bonds. In this paper, we propose a novel Molecular Graph Contrastive Learning with Parameterized Explainable Augmentations, MolCLE for brevity, that self-adaptively incorporates chemically significant information from both topological and semantic aspects of molecular graphs. Specifically, we apply deep neural networks to parameterize the augmentation process for both the molecular graph topology and atom attributes, to highlight contributive molecular substructures and recognize underlying chemical semantics. Comprehensive experiments on a variety of real-world datasets demonstrate that our proposed method consistently outperforms compared baselines, which verifies the effectiveness of the proposed framework. Detailedly, our self-supervised MolCLE model surpasses many supervised counterparts, and meanwhile only uses hundreds of thousands of parameters to achieve comparative results against the state-of-the-art baseline, which has tens of millions of parameters. We also provide detailed case studies to validate the explainability of augmented graph views.

CCS CONCEPTS

• **Mathematics of computing** → **Graph algorithms**; • **Applied computing** → **Bioinformatics**; • **Computing methodologies** → **Neural networks**; **Unsupervised learning**.

*The first two authors contributed equally to this work.

†To whom correspondence should be addressed.

KEYWORDS

Molecule modeling, contrastive learning, graph representation learning, self-supervised learning, parameterized data augmentation

1 INTRODUCTION

The capability of obtaining informative molecular representations is an essential key and a crucial prerequisite in AI-driven drug design and discovery. Well-designed molecular representations will benefit various academic areas and industrial applications such as improving rational chemical design, reducing R&D cost, decreasing the failure rate in potential drug screening trials, as well as speeding the process of *de novo* drug design [65]. However, due to the large-scale vocabulary of possible stable chemical compounds, it is challenging to develop generalizable, transferable, and robust molecular representations covering the entire molecule space. Traditional molecular fingerprints like SMILES [58] and ECFP [46], provide a description of a specific part of the molecular structure [13]. However, they usually require intensive manual feature engineering and strong domain knowledge, and finally become highly task-dependent, with no generalizability for other new-come tasks.

Besides, these methods also suffer two main problems. Firstly, **most of them are task-specific**. For many areas like computer vision (CV) [20] and speech signal processing (SSP) [15], the number of labeled samples could easily reach several million or even more. However, it is not the same situation with molecule-related tasks, which often require an expensive and time-consuming process to get access to labeled molecule data. The scenario with unlimited unlabelled data while a tiny portion labeled ones are similar to natural language modeling. Thus, the state-of-the-art natural language pre-training techniques [6, 7] can be easily applied over string-based molecule fingerprints. But, these approaches will lead to the second problem: **string-based representations fail to encode important topology**. Molecules can be naturally represented by molecular graphs which preserve rich structural information. Over the past few years, graph representation learning has emerged as a powerful strategy for analyzing graph-structured data, which has also been extended to encode molecular graphs. Among them, the most prevailing graph neural networks (GNN) [8, 27, 29] receive considerable attention, which aims to transform graph components to low-dimensional dense embeddings that preserve attributive and structural features. Despite existing GNN models flourish in many tasks, such as molecular property prediction [56] and virtual screening [57], they are mostly established in a supervised manner [44], bringing back to the first problem.

Recently, Contrastive Learning (CL), as the revitalization of the classical information Maximization (InfoMax) principle [32], has performed great achievement in many fields, e.g., visual representation learning [4, 19] and natural language processing (NLP) [38]. The CL framework seeks to maximize the mutual information (MI) between the input (i.e. images) and its representations (i.e. image embeddings) by contrasting positive pairs with negative-sampled counterparts. Inspired by the previous success of the most famous Deep InfoMax (DIM) [2], Deep Graph InfoMax (DGI) [55] firstly introduces the InfoMax principle into GNN-based self-supervised learning. DGI simply applies random shuffling on node features to generate graph augmentations. Then, a contrastive objective is used to maximize the MI between node patch embeddings and graph summary representations. Following DGI, GRACE [72] generates corrupted views of graphs by applying random node feature masking and edge removing and measures MI from two views both in patch level. GraphCL [69] proposes four simple graph augmentations including node dropping, edge perturbation, attribute masking, and induced subgraph. To supplement more global information, MVGRL [18] firstly utilizes graph diffusion kernels [31] to augment graphs and constructs graph views by randomly sampling subgraphs. Then it contrasts patch and graph representations across views. Despite these methods achieve success in both node- and graph-level tasks, all of them only perform simple data augmentations but ignore exploring more abundant ones, which proved to be a critical component for visual representation learning [59]. GCA [73] studies the discrepancy in the impact of nodes and edges when performing data augmentation in graph CL. It identifies important edges and feature dimensions via centrality measures and then generates different graph views by masking the unimportant ones.

Although graph CL frameworks with simple or adaptive graph augmentations have shown significant improvements on the social network, citation network, and other internet datasets, they can not be easily transferred to biomedical molecule datasets because such augmentations will **generate meaningless graph views at the molecule-level**. Uniformly dropping nodes or perturbing edges may drastically change the identity and even the validity of a compound, e.g., removing an aromatic bond or dropping an atom from a functional group. Besides, centrality-based adaptive graph augmentations largely borrow the hypothesis from network science that centrality reflects inter-node influences. Thus, the generated graph views are constructed by the most influential nodes, which will also deteriorate meaningful substructure in molecular graphs. Meanwhile, from the attribute perspective, the simple augmentations maybe **not sufficient to generate diverse contexts for nodes and improperly mask important feature dimensions**, e.g., masking feature dimensions like inter-atomic distances when predicting targets related to atomic orbital and energy gap. Therefore, such operations will cause pointless augmentations, leading to difficulty in optimizing the contrastive objective.

To this end, we propose a novel contrastive framework for self-supervised molecular graph representation learning, as shown in Figure 1, which we refer to as Molecular Graph Contrastive Learning with Parameterized Explainable Augmentations, MolCLE for brevity. (i) In MolCLE, we first design two parameterized graph data augmentations on both topology and attribute level. We apply

a generative probabilistic model to learn *intrinsic underlying molecular graph structures* as the topology-level augmentations, which are believed to make the most contribution to molecule representations readout from GNNs. Simultaneously, MolCLE also learns *feature selectors* that mask out unimportant atom features to generate attribute-level augmentations. (ii) Then, we utilize them to obtain correlated molecular graph views and train the model using a contrastive loss to maximize the agreement between augmented molecular graph representations in these two views. (iii) We pre-train MolCLE with millions of unlabelled molecule data, and then leverage the pre-trained GNN backbones to many downstream molecular property prediction datasets followed by task-specific fine-tuning. With only less than a million parameters, our approach achieves better performance than state-of-the-art methods with hundreds of million parameters. Extensive experiments show that MolCLE outperforms existing methods and our self-supervised method even surpasses its supervised counterparts on many downstream tasks.

Our contributions of this paper could be summarize as:

- Firstly, we propose a contrastive learning framework for self-supervised molecular graph representation learning with parameterized, explainable data augmentation. The proposed MolCLE framework jointly performs data augmentation on both topology and attribute levels that can automatically emphasize important parts of the graph structure and attributes, which encourages the model to extract essential features from both aspects.
- Secondly, we conduct comprehensive empirical studies using 6 downstream popular benchmark datasets on molecular property prediction compared with many current state-of-the-art results for each dataset. We observe a huge relative improvement of MolCLE over existing baselines that even surpasses some supervised counterparts.

2 RELATED WORK

2.1 Molecule Representation Learning

To represent molecules in the embedding space, traditional chemical molecule fingerprints, such as Morgan Identifier [39] and ECFP [46], try to encode the neighboring atoms of the central atom in the molecule into a fix-length vector. To improve the expressive power of chemical fingerprints, some studies [8] introduce learning-based methods like convolutional layers to learn the neural fingerprints of molecules. Then the learned molecule representations are applied to downstream tasks. Following these works, many other researchers take the SMILES representation [58] as input and use language modeling approaches to extract molecule fingerprints like Mol2vec [24] and Seq2seq [65]. Recently, considering molecule naturally as graph-structure data, many works [27, 56] explore the flourished GNNs to encode molecular graphs into low-dimensional representations. A couple of works [54, 71] extend attention mechanism to graph aggregation operation by learning the aggregation weights automatically. Moreover, Gilmer et al. [12] proposes a unified message passing framework to encode molecular graphs for quantum chemical property prediction. Following this work, some research works [34, 66] propose to model bond interactions over this framework. Besides, several works [34, 57] model molecular

Figure 1: Our proposed Molecular Graph Contrastive Learning with Parameterized Explainable Augmentations (MolCLE) model. (a) Molecular graph contrastive learning pipeline. We apply two parameterized graph augmentation components for each molecule, where they simultaneously learn topological and attributive augmentation while retaining domain priors, to generate two explainable augmented graph views. Finally, the augmented graph views from the same molecule are treated as positive pairs, pulling their representations closer in the following contrastive training process. And the negative samples are constructed from the anchor and augmented views of other graphs in the same minibatch. (b) Parameterized augmentation illustration. For each graph view, certain topological connectivity and attribute dimensions are selected, where atoms in the same view share a common attribute selector. Noted that some atoms' attributes will not be included if they are ignored by the topology selector.

graphs and cross-molecule interactions in a multi-level manner. More recently, several self-supervised GNN-based models [23, 47] tailored for molecule data design elaborate pre-training techniques and can obtain good molecule representations, which achieve great performance on downstream tasks.

2.2 Graph Neural Networks

Recent success of GNNs [8, 12, 29] have aroused many attention to use these methods for analyzing various graph-structured data. The general framework includes an iterative neighboring aggregation (namely message passing) scheme to update nodes' hidden states within information from their neighboring nodes and themselves. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote an undirected graph, $\mathbf{h}_n$ is the embedding of the vertex $v_n \in \mathcal{V}$. Considering a K -layer GNN $f_{GNN}(\cdot)$, the propagation of the k th layer is represented as:

$$\begin{aligned} \mathbf{m}_n^{(k)} &= \text{AGGREGATION}^{(k)}(\mathbf{h}_{n'}^{k-1} : n' \in \mathcal{N}(n)) \\ \mathbf{h}_n^{(k)} &= \text{COMBINE}^{(k)}(\mathbf{h}_n^{k-1}, \mathbf{m}_n^{(k)}) \end{aligned} \quad (1)$$

where $\mathcal{N}(n)$ is a set of neighboring vertices adjacent to v_n , and $\text{AGGREGATION}^{(k)}(\cdot)$ and $\text{COMBINE}^{(k)}(\cdot)$ are component functions of the GNN layer, which have multiple popular choices like mean, max pooling and attention mechanism [53]. After the K -layer propagation, the graph-level embedding for $\mathcal{G}$ is summarized on layer embeddings through the READOUT function:

$$\mathbf{h}_G = \text{READOUT}(\{\mathbf{h}_v^{(0,K_0)}, \dots, \mathbf{h}_v^{(L,K_L)} | v \in \mathcal{V}\}) \quad (2)$$

2.3 Contrastive Representation Learning

The main idea of contrastive learning is aiming to make representations agree with each other under proper transformations, raising a recent surge of interest in visual representation learning [4, 19]. For visual data, negative samples can be easily generated using a multi-stage augmentation pipeline [2, 10] including rotation, resizing, cropping, flipping, color jittering and distortion, etc. Existing works [52, 59] propose to store negative samples with a memory bank. In contrast, some other works [4, 67] explore in-batch negative samples. Considering an image patch as the anchor, the positive samples are often constructed by patches in neighboring views [21] or a global summary vector [22]. Then, the positives are contrasted with the negative counterparts like patches of other images [22].

On a parallel note, for graph data, traditional methods [30] trying to reconstruct the adjacency information of nodes [17], which could be seen as a contrast of locality. These methods emphasize the proximity information over the topological information [45], which ignores information from attributes. Also, these methods are known to be error-prone with inappropriate hyperparameter tuning [16, 43]. The most famous DGI [55] is the first work that combines GNN and CL, which applies an InfoMax objective and maximizes MI between global graph-level and local node-level embeddings. Following DGI, GMI [42] measures MI between input and representations of both nodes and edges without data augmentations by two discriminators. GRACE [72] focuses on maximizing the agreement of node embeddings across two corrupted views

Table 1: Summary of comparisons

Model	Augmentation			Contrastive objective	
	Topo.	Attr.	Type		
DGI [55]	✓	✗	Uniform	Node	Graph
GRACE [72]	✓	✓	Uniform	Node	Node
GraphCL [69]	✓	✓	Uniform	Graph	Graph
GMI [42]	✗	✗	None	Node	Node
InfoGraph [51]	✗	✗	None	Node	Graph
MVGRL [18]	✓	✗	Uniform	Node	Graph
GCA [73]	✓	✓	Adaptive	Node	Node
MolCLE	✓	✓	Self-adaptive	Graph	Graph

of the graph. GraphCL [69] proposes four graph data augmentations to generate different views and maximizing the agreement of these views. MVGRL [18] performs graph diffusion kernels to learn both node and graph representations and then contrast them. InfoGraph [51] applies a discriminator with a batch of graphs to determine which patches are from the same graph. GCA [73] introduces two adaptive graph data augmentations, aiming to discover the centrality-based influential substructures of the graph. However, these methods do not explicitly consider chemical and biological priors inside molecular graphs, leading to improper graph augmentations at both structural and attribute levels. Also, these augmented views will deteriorate optimization of the contrastive objective and reach a suboptimal performance.

Connection to related works. We summarize the comparisons between our proposed MolCLE method with other related state-of-the-art graph contrastive learning models, including DGI [55], GRACE [72], GraphCL [69], GMI [42], InfoGraph [51], MVGRL [18], and GCA [73] in Table 1. It is obvious that only our MolCLE model proposes parameterized explainable graph data augmentations on both topology and attribute levels for preserving domain priors better, which are especially crucial on molecular graph data.

3 METHODOLOGY

In the following section, we introduce our proposed MolCLE model in detail, starting with the parameterized graph data augmentations, followed by the overall graph contrastive learning framework.

3.1 Preliminaries

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent the graph with $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ denoting the node set and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ as the edge set. The numbers of nodes and edges are denoted by N and M , respectively. A graph can be described by an adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$, with $a_{ij} = 1$ if there is an edge connecting node i and j , and $a_{ij} = 0$ otherwise. Nodes in $\mathcal{V}$ are associated with the d -dimensional features, denoted by $\mathbf{X} \in \mathbb{R}^{N \times d}$. As stated in Section 2.2, the message and hidden state representation of node v_n in the k th layer denotes $\mathbf{m}_n^{(k)}$ and $\mathbf{h}_n^{(k)}$, respectively. And the hidden representation of the last GNN layer serves as the final node representation $\mathbf{z}_n = \mathbf{h}_n^{(L)}$, which can be then readout to a graph representation $\mathbf{h}_G$.

3.2 Parameterized Graph Augmentation

The core idea of CL is maximizing agreement between augmented views to obtain *perturbation-invariant* representations [62]. Existing graph data augmentation mainly lies in two categories: generative-based methods and simple random disturbance. The former often requires prohibitive additional computation cost [1] and the latter usually destroy domain priors that deteriorate the quality of the learned representations [69]. Thus, it is obvious that self-adaptive graph data augmentation preserving underlying essential graph structures still remains under-explored. In our MolCLE model, we propose to design parameterized explainable augmentation schemes that self-adaptively learn to focus on intrinsically important graph topology and node attributes while perturbing relatively unimportant edges and features. Specifically, MolCLE learns *graph topology masks* and *feature selectors* using generative probabilistic models which select the important subgraph and mask out unimportant node features of input molecular graph as topology- and attribute-level augmentations, respectively. We argue that augmented graph views in molecule modeling scenarios should preserve chemically meaningful topological information instead of randomly dropping edges. Also, target-influential attributes should be emphasized to preserve semantic patterns rather than randomly masking out them.

3.2.1 Topological augmentation. Previous topological augmentation methods usually apply uniform node dropping, edge perturbation [69] or graph corruption guided by rules in specific domains [73]. However, these approaches can not be adapted to molecular graph augmentation. They may generate meaningless augmentation with no chemical significance. Thus, we consider a self-adaptive approach that automatically learning intrinsic underlying topology for input molecular graphs. Inspired by previous works on explaining GNNs [35, 68, 70], we utilize a generative probabilistic model to obtain topological augmentations. Different from these works, we do not have training labels in contrastive learning. Therefore, we try to find a graph view G_A by maximizing the MI between graph embeddings obtained from original graph topology G_O and the underlying graph topology G_A :

$$\max_{G_A} MI(\mathbf{Z}_O, G_A) = \max_{G_A} [H(\mathbf{Z}_O) - H(\mathbf{Z}_O|G = G_A)] \quad (3)$$

where $\mathbf{Z}_O$ is the graph embeddings output from the GNN model with G_O as the input. The MI measures the proximity of the graph embedding $\mathbf{Z}_A$ to $\mathbf{Z}_O$ when the input graph to the GNN model is limited to the augmented graph view. Maximizing MI can help to find the underlying graph view that makes the most contributions to generate $\mathbf{Z}_A$ closest to $\mathbf{Z}_O$. The intuition behind this is that the remaining topology of the graph view G_A should preserve the original graph embedding $\mathbf{Z}_O$ as much as possible, e.g., if removing an edge dramatically disturbs $\mathbf{Z}_O$, thus, the edge is essential and should be selected in G_A . Otherwise, it can be removed for its irrelevance.

Due to the intractable number of potential subgraphs hindering the model from optimizing the objective directly [35], we follow Janson et al. [26] to consider a direct way that factorizing the probability of a specific augmented graph view:

$$P(G_A) = \prod P\{(i, j) \in \tilde{\mathcal{E}}\} \quad (4)$$

where $\tilde{\mathcal{E}}$ denotes the edge set of the graph view, and $P\{(i, j) \in \tilde{\mathcal{E}}\}$ is drawn from a Bernoulli distribution $\mathcal{B}(p_{ij}^e)$. Considering $p_{ij}^e = \phi_{ij}$ and G_A is drawn from the distribution $d(\Phi)$ of the ϕ -parameterized graph view, we reformulate Equation (3) as:

$$\max_{G_A} MI(Z_O, G_A) \approx \min_{\Phi} [\mathbb{E}_{G_A \sim d(\Phi)} H(Z_O | G = G_A) - H(Z_O)] \quad (5)$$

Besides, for efficiently optimizing Equation (5) with gradient descent [68], we also adopt the reparameterization trick [25] to relax edge sampling weight variables $\tilde{E}$ from binary to continuous between 0 to 1. Formally, the entry of $\tilde{E}$ can be instantiated as a binary concrete distribution [36]:

$$\tilde{e}_{uv} = \sigma((\log \delta - \log(1 - \delta) + \psi_{uv})/\lambda) \quad (6)$$

where δ is drawn from a Uniform distribution $\mathcal{U}(0, 1)$ and $\sigma(\cdot)$ denotes the Sigmoid function. We parameterize $\psi_{uv} \in \mathbb{R}$ by a fully-connected neural network $FCNN_{\Gamma}([z_u; z_v])$, where $[\cdot]$ represents a concatenation operation.

Although Maddison et al. [36] have demonstrated that using binary concrete distribution to approximate the Bernoulli distribution is reasonable, the conditional entropy is still remained difficult to be optimized. Inspired by Ying et al. [68], we modify it with a soft cross-entropy $H(Z_O, \tilde{Z}_A)$ and adopt Monte Carlo [37] to complete approximate optimization for Equation (5):

$$\begin{aligned} \max_{G_A} MI(Z_O, G_A) &\approx \min_{\Psi} [\mathbb{E}_{\Psi} H(Z_O, \tilde{Z}_A) - H(Z_O)] \\ &= \min_{\Psi} \left[\log Z_O - \frac{1}{K} \sum_{k=1}^K (Z_O \cdot \log \tilde{Z}_A^{(k)}) \right] \end{aligned} \quad (7)$$

where $\tilde{Z}_A^{(k)}$ denotes the k th sampled graph view with Equation (6) and K is the total number. In practice, an m -dimensional Softmax layer is applied upon Z_O and $\tilde{Z}_A^{(k)}$ for stable optimization. Thus, Equation (7) can be rewritten as follows:

$$\min_{\Psi} \left\{ \log S(Z_O) - \frac{1}{K} \sum_{k=1}^K [\text{Softmax}(Z_O) \cdot \log S(\tilde{Z}_A^{(k)})] \right\} \quad (8)$$

where $\log S(\cdot)$ combines the logarithm with the $\text{Softmax}(\cdot)$ function. The intuition behind is obvious that the graph embedding with a well-augmented graph view as GNN's input should be closest to the original one.

Through optimizing Equation (8), the proposed model can easily learn appropriate topological augmentations for molecular graph instances, which include meaningful semantic information over molecular structures. We summarize the topological augmentation pipeline in Figure 2. However, the obtained graph views maybe not concerning constraints avoiding chemical meaningless. Therefore, we propose two regularization terms to preserve desired chemical properties.

Ratio constraint. Ying et al. [68] introduces a *size constraint* on the obtained explanatory graphs by adding $\|\Phi\|_1$, the l_1 norm on latent variables Φ , as a regularization term. But in our case, we need to preserve a reasonable size for augmented graph views instead of only penalizing the sum of all elements of Φ . Previous work [35] imposes an alternative solution, namely *budget constraint*, which applies a hard threshold B to limit the size of explanatory graphs.

Figure 2: Topological augmentation.

However, it can not be adapted to molecular graphs having various sizes. Thus, we propose a *ratio constraint* that self-adaptively restricts the size of each molecular graph, i.e., ensuring the topological augmentation ratio. The ratio constraint can be formulated as follows:

$$\mathcal{L}_R = \text{ReLU} \left(\sum_{(u,v) \in \mathcal{E}} \tilde{e}_{uv} - \eta \cdot |G_A| \right) \quad (9)$$

where $\text{ReLU}(\cdot)$ is the ReLU activation and η denotes the default topological augmentation ratio. The ratio constraint $\mathcal{L}_R = 0$ when the ratio of the size of the augmented graph view to the size of the original molecular graph is lower than the default ratio η .

Continuity constraint. In the molecular modeling scenario, the augmentation graph views are usually expected to be connected. Therefore, adding a constraint on generating continuous topological graphs is necessary. The significance of this claim has been empirically proved and explicitly introduced [35, 68]. Thus, we design a more specific regularization term on the topological connectivity of augmented graph views:

$$\hat{n} = \underset{\substack{m \in \mathcal{V} \\ n \in \mathcal{N}(m)}}{\text{argmax}} \tilde{e}_{mn} \quad (10)$$

$$\mathcal{L}_C = \begin{cases} \sum_{m \in \mathcal{V}} \sum_{n \in \mathcal{N}(m)} \left(\tilde{e}_{m\hat{n}} \cdot \log \frac{\tilde{e}_{m\hat{n}}}{\tilde{e}_{mn}} \right), & \tilde{e}_{mn} \geq \text{threshold}, \\ \sum_{m \in \mathcal{V}} \sum_{n \in \mathcal{N}(m)} \left(\tilde{e}_{m\hat{n}} \cdot \log \frac{\tilde{e}_{m\hat{n}}}{(1 - \tilde{e}_{mn})} \right), & \tilde{e}_{mn} < \text{threshold}. \end{cases} \quad (11)$$

where $\text{threshold} := \frac{1}{|\mathcal{N}(n)|} \sum_{n \in \mathcal{N}(m)} \tilde{e}_{mn}$. In practice, we add a uniform noise over $\tilde{e}_{mn}$ to avoid local optimum. Obviously, Equation (11) encourages the augmented graph views to include small connected subgraphs rather than many discretized edges.

Edge-entropy constraint. Although Equation (11) ensures the topological connectivity of augmented graph views, it will make the topological masks smoother, which means that different edge weight variables gradually become indistinguishable. Thus, we also apply the element-wise entropy [35, 68] constraint to increase the

gap of edge variables $\tilde{e}$ with large values and those with small values:

$$\mathcal{L}_E = \frac{1}{|\mathcal{E}|} \sum_{(u,v) \in \mathcal{E}} [-\tilde{e}_{uv} \cdot \log \tilde{e}_{uv} - (1 - \tilde{e}_{uv}) \cdot \log(1 - \tilde{e}_{uv})] \quad (12)$$

3.2.2 Attributive augmentation. Existing attributive augmentation methods mainly concentrate on uniform or rules-guided attribute masking [69, 72, 73], ignoring to explore more strategies on self-adaptively selecting influential feature dimensions. Here we design learnable feature selectors to add noise on node attributes via automatically masking a fraction of dimensions with zeros in node features. Formally, MolCLE learns a binary feature selector $\tilde{s} \in \{0, 1\}^F$ where each dimension $\tilde{s}_i$ is drawn from a Bernoulli distribution $\mathcal{B}(p_i^s)$. With $\mathbf{x}_A$ representing all dimensions of node features in G_A , the selected node features can be considered as the element-wise multiplication of the feature selector and $\mathbf{x}_A$, which is formulated as $\tilde{\mathbf{x}}_A = \mathbf{x}_A \odot \tilde{s}$. Thus, we rewrite Equation (3) to jointly optimize topological and attributive augmentation via maximizing the MI objective:

$$\max_{G_A, \tilde{s}} MI(\mathbf{Z}_O, (G_A, \tilde{s})) = \max_{G_A, \tilde{s}} [H(\mathbf{Z}_O) - H(\mathbf{Z}_O | G = G_A, X = \tilde{\mathbf{x}}_A)] \quad (13)$$

where $\tilde{\mathbf{x}}_A := \{\tilde{\mathbf{x}}_{A,j} | v_j \in G_A\}$ is obviously related to $\tilde{\mathbf{Z}}_A$ in Equation (7) and (8) but do not change the formulation of these two equations. Therefore, we still serve Equation (8) as the optimization objective. Here we provide some intuition behind this objective. If a specific feature is important, the feature selector will learn to select it, which means that the corresponding weights in $\tilde{s}$ will be close to one, and vice versa.

However, similar to topological augmentation, for backpropagating gradients in Equation (13) to the feature selector $\tilde{s}$ smoothly, we have to apply a reparameterization trick [25] on $\mathbf{x}_A$:

$$\tilde{\mathbf{x}}_A = \mathbf{r} + (\mathbf{x}_A - \mathbf{r}) \odot \tilde{s} \quad (14)$$

where $\mathbf{r}$ denotes a random variable drawn from the empirical distribution. Meanwhile, we also give some additional constraints for attributive augmentation, aiming to generate more meaningful graph views.

Size constraint. Similar to topological augmentation, we also need to keep a reasonable number of selected features in attributive augmentation. Thus, we propose a *size constraint* upon the number of selected feature dimensions like the setting in *ratio constraint*:

$$\mathcal{L}_S = \text{ReLU} \left(\sum_{\tilde{s}_i \in \tilde{s}} \tilde{s}_i - \eta \cdot d_s \right) \quad (15)$$

where $\text{ReLU}(\cdot)$ is the ReLU activation and η denotes the default attributive augmentation ratio. The ratio constraint $\mathcal{L}_S = 0$ when the ratio of the number of selected feature dimensions to the number of original feature dimensions is lower than the default ratio η .

Feature-entropy constraint. As we have done in Section 3.2.1, we also apply element-wise entropy to encourage the feature selector $\tilde{s}$ to be discrete:

$$\mathcal{L}_F = \frac{1}{d_s} \sum_{\tilde{s}_i \in \tilde{s}} [-\tilde{s}_i \cdot \log \tilde{s}_i - (1 - \tilde{s}_i) \cdot \log(1 - \tilde{s}_i)] \quad (16)$$

Finally, by jointly performing topological and attributive augmentation stated above, we can generate two augmented molecular graph views $\tilde{\mathcal{G}}_1$ and $\tilde{\mathcal{G}}_2$. Considering these two augmentations as an overall augmentation pipeline, we perform it independently when generating each view but optimize the objectives in an end-to-end style with the contrastive objective in Section 3.3. Due to different initialization, proposed constraints, and the involved random variables, two generated views will display slightly different but chemically meaningful. We provide some visualized examples in Section 4.4 for more direct observation.

3.3 Graph Contrastive Learning Framework

We follow the general and widely-used graph contrastive learning framework [4, 22, 51, 69, 72, 73] which applies a contrastive objective to maximize the agreement of representations between different views. However, due to our graph-level learning scenario [51, 69], we employ the best-matched pipeline that enforces the encoded embeddings of each molecular graph in two different augmentation views to agree with each other and can be discriminated from embeddings of the others.

In our MolCLE model, at each iteration, we construct two augmentation pipelines $\mathcal{A}_1$ and $\mathcal{A}_2$ and generate two correlated molecular graph views, denoted as $\tilde{\mathcal{G}}_1 = \mathcal{A}_1(\mathcal{G})$ and $\tilde{\mathcal{G}}_2 = \mathcal{A}_2(\mathcal{G})$, as a positive pair. Then, a GNN encoder $f_{GNN}(\cdot)$ defined in Section 2.2 will extract graph-level embeddings $\mathbf{h}_{G_1}, \mathbf{h}_{G_2}$ for augmented graphs $\tilde{\mathcal{G}}_1$ and $\tilde{\mathcal{G}}_2$. Following previous works [4, 69, 72, 73], we apply a projection head to map augmented graph embeddings to another latent space to calculate contrastive loss. The projection head is usually defined as a non-linear transformation, which can enhance the expressive power of the contrastive discrimination. Thus, in our study, we use a two-layer fully-connected neural networks to obtain representations $\mathbf{z}_{G_1}, \mathbf{z}_{G_2}$ of $\tilde{\mathcal{G}}_1$ and $\tilde{\mathcal{G}}_2$ in the other latent space.

Then, we utilize a contrastive objective $\mathcal{L}(\cdot, \cdot)$ to enforce maximizing the consistency between positive pairs $\mathbf{z}_{G_1}, \mathbf{z}_{G_2}$ compared with negative pairs. For the k th graph G_k in a minibatch of N graphs, its augmented graph-level representation in latent space can be annotated as $\mathbf{z}_{G_{k,1}}, \mathbf{z}_{G_{k,2}}$ to form the positive sample. We serve the embedding in one view, i.e. $\mathbf{z}_{G_{k,1}}$, as the anchor. Naturally, the negative samples can be constructed from the anchor and augmented views of other graphs in the same minibatch [4, 5]. In our MolCLE model, we apply the wide-spread NT-Xent [40, 49, 61] loss as the contrastive objective and define the pairwise formulation over each positive sample:

$$\mathcal{L}(\mathbf{z}_{G_{k,1}}, \mathbf{z}_{G_{k,2}}) = -\log \left[\frac{\exp \left(\frac{1}{\tau} \cdot \frac{\mathbf{z}_{G_{k,1}}^T \cdot \mathbf{z}_{G_{k,2}}}{\|\mathbf{z}_{G_{k,1}}\| \cdot \|\mathbf{z}_{G_{k,2}}\|} \right)}{\sum_{l=1, l \neq k}^N \exp \left(\frac{1}{\tau} \cdot \frac{\mathbf{z}_{G_{k,1}}^T \cdot \mathbf{z}_{G_{l,2}}}{\|\mathbf{z}_{G_{k,1}}\| \cdot \|\mathbf{z}_{G_{l,2}}\|} \right)} \right] \quad (17)$$

where τ is the temperature parameter. It should be noted that we compute the contrastive loss across all positive samples in a minibatch. The contrastive objective is jointly optimized with the molecular graph augmentation process. Equation (17) can be served as maximizing the MI between the representations of two augmented

Table 2: Dataset summary

Category	Dataset	#Tasks	#Molecules	Metrics
Biophysics	BBBP	1	2039	ROC-AUC
Physiology	BACE	1	1513	ROC-AUC
	Tox21	12	7831	
	ToxCast	617	8575	
	ClinTox	2	1478	
	SIDER	27	1427	

molecular graph views h_{G_1} and h_{G_2} , satisfying the classical Info-Max principle [32]. We follow Chen et al. [4] and You et al. [69] to allocate the same GNN encoder for two different views while there are many specific alternative CL frameworks [19, 22, 52].

4 EXPERIMENTS

In this section, we present the details of the experiments and the corresponding results. To illustrate the effectiveness of our proposed model, we compare it with some strong baselines on molecular property prediction tasks. We start with four research questions (RQs) to lead the experiments and the following discussions.

- **(RQ1)** Does our proposed MolCLE model outperforms existing baseline methods in molecular property prediction tasks over various datasets?
- **(RQ2)** Do all proposed parameterized graph augmentation schemes benefit the contrastive learning process? How does each one affect the model performance?
- **(RQ3)** Can the proposed parameterized graph augmentation schemes generate meaningful augmented graph views?
- **(RQ4)** Are the proposed constraints make explainable augmentations? How do they influence specific molecules?

4.1 Experimental setups

4.1.1 Datasets.

Contrastive learning dataset. To train our MolCLE model, a large number of unlabeled molecule data is necessary to perform contrastive learning. Thus, we follow Hu et al. [23] to use 2 million unlabeled molecules¹ sampled from the ZINC15 database [50] for graph contrastive learning.

Fine-tuning tasks and datasets. To prove the effectiveness of our MolCLE model, we select six molecule property prediction tasks from the MoleculeNet benchmark² [60]. We summarize the used benchmark datasets in Table 2.

Fine-tuning dataset splitting. Similar to implementation in Rong et al. [47], we also use scaffold-splitting [3] to avoid obtaining overly optimistic results, which clusters compounds by scaffold (molecular graph substructure) and reassemble the clusters by placing the most common scaffolds in the training datasets, generating validation and test datasets which include structurally different compounds. Previous research works have demonstrated that this special scaffold splitting is more approximate to the ideal chronological split, which can stimulate the real-world application scenarios better. The

model evaluated in the setting of scaffold splitting is convinced to give more reliable predictions. Therefore, in our research work, we use the special scaffold splitting to split the downstream benchmark datasets. The split ratio for train, validation and test datasets are typically 80%, 10%, and 10% respectively.

4.1.2 Implementation details. In our experiments, we employ GIN with the default setting in Xu et al. [64] as the GNN encoder [23, 69]. Experiments are performed 10 times with mean and standard deviation of ROC-AUC scores (%) reported as Hu et al. [23]. We fine-tuned our MolCLE model on the training datasets of downstream molecular property prediction tasks. We use a batch size of 32 and dropout 0.5. We use Adam optimizer [28] for both contrastive learning and fine-tuning. All variables are initialized with Xavier [14]. We use the validation loss to select the best model. All models are implemented using PyTorch Geometric [11] and PyTorch [41]. All fine-tuning datasets are available in PyTorch Geometric libraries. The initial learning rates of contrastive learning and fine-tuning are 0.0001 and 0.001, respectively. All experiments are implemented on 1 NVIDIA Titan RTX GPU.

4.1.3 Baselines. We comprehensively evaluate MolCLE against 8 popular supervised counterparts from MoleculeNet [60] and 3 state-of-the-arts self-supervised baselines. Among them, TF_Robust [44] is a DNN-based multitask framework taking the molecular fingerprints as the input. GraphConv [29], Weave [27] and SchNet [48] are three graph convolutional models. MPNN [12] and its variants DMPNN [66] and MGCN [34] are models considering the edge features during message passing. AttentiveFP [63] is an extension of the graph attention network. Specifically, to demonstrate the superiority of our proposed contrastive learning framework and parameterized graph augmentation schemes, we also compare MolCLE with three self-supervised models: Pre-GIN [23], GraphCL [69] and GROVER³ [47].

4.2 (RQ1) Performance results compared to baselines on multiple datasets

The experimental results are summarized in Table 3. Overall, from the table, we can see that our proposed model achieves strong performance across all six datasets. Moreover, our MolCLE outperforms all the supervised counterparts by considerable margins on molecular property prediction tasks. Besides, compared with Pre-GIN [23], our MolCLE pushes the performance more forward, demonstrating the superiority of the proposed contrastive learning framework. Furthermore, our MolCLE model achieves higher results across all datasets compared to GraphCL [69], which can be attributed to our proposed parameterized graph augmentation schemes, verifying their significance on molecular graph contrastive learning. We will analyze these schemes in detail when answering the following research questions. Finally, we can notice that our MolCLE model shows a comparative performance to GROVER [47], which has the largest GNNs architecture tailored for molecular graph data with tens of millions of parameters. However, our MolCLE, less than one million parameters, outperforms GROVER over three datasets and achieves similar results across the others, performing a notable success on extending contrastive learning to molecular graph data.

¹The contrastive learning dataset is here: <https://snap.stanford.edu/gnn-pretrain>

²All fine-tuning datasets can be found here: <http://moleculenet.ai/datasets-16>

³We pre-train this model using the same dataset as our contrastive training stage.

Table 3: The performance results compared to baselines (ROC-AUC(%)).

Learning paradigm	Dataset #Molecules	BBBP 2039	BACE 1513	Tox21 7831	ToxCast 8575	ClinTox 1478	Sider 1427
Supervised	TF_Robust[44]	86.01 $\pm$ 0.090	82.38 $\pm$ 0.022	69.78 $\pm$ 0.011	56.39 $\pm$ 0.056	76.51 $\pm$ 0.089	60.68 $\pm$ 0.031
	GraphConv[29]	87.67 $\pm$ 0.044	85.42 $\pm$ 0.013	77.21 $\pm$ 0.034	66.67 $\pm$ 0.022	84.45 $\pm$ 0.035	54.28 $\pm$ 0.023
	Weave[27]	83.71 $\pm$ 0.067	79.12 $\pm$ 0.011	74.11 $\pm$ 0.034	65.45 $\pm$ 0.012	82.33 $\pm$ 0.022	54.31 $\pm$ 0.023
	SchNet[48]	84.67 $\pm$ 0.023	75.01 $\pm$ 0.023	76.67 $\pm$ 0.245	67.12 $\pm$ 0.013	71.73 $\pm$ 0.024	59.54 $\pm$ 0.023
	MPNN[12]	91.30 $\pm$ 0.044	81.52 $\pm$ 0.045	80.81 $\pm$ 0.033	66.50 $\pm$ 0.025	87.89 $\pm$ 0.051	59.48 $\pm$ 0.095
	DMPNN[66]	91.89 $\pm$ 0.031	85.18 $\pm$ 0.055	81.62 $\pm$ 0.025	67.79 $\pm$ 0.013	89.72 $\pm$ 0.048	63.21 $\pm$ 0.023
	MGCN[34]	85.04 $\pm$ 0.064	73.45 $\pm$ 0.032	70.71 $\pm$ 0.024	65.81 $\pm$ 0.017	63.42 $\pm$ 0.049	55.22 $\pm$ 0.018
Self-supervised	AttentiveFP[63]	90.81 $\pm$ 0.050	86.34 $\pm$ 0.002	80.78 $\pm$ 0.020	57.04 $\pm$ 0.011	93.36 $\pm$ 0.020	60.54 $\pm$ 0.060
	Pre-GIN[23]	91.45 $\pm$ 0.041	79.91 $\pm$ 0.037	79.72 $\pm$ 0.025	64.21 $\pm$ 0.027	76.78 $\pm$ 0.061	61.02 $\pm$ 0.058
	GraphCL[69]	91.87 $\pm$ 0.012	75.45 $\pm$ 0.041	78.94 $\pm$ 0.032	62.41 $\pm$ 0.028	85.82 $\pm$ 0.037	60.45 $\pm$ 0.045
	GROVER[47]	93.53 $\pm$ 0.031	88.63 $\pm$ 0.067	81.88 $\pm$ 0.023	68.89 $\pm$ 0.011	90.21 $\pm$ 0.046	63.01 $\pm$ 0.033
	MolCLE	92.66 $\pm$ 0.037	91.71 $\pm$ 0.045	80.57 $\pm$ 0.091	69.01 $\pm$ 0.016	92.51 $\pm$ 0.021	62.67 $\pm$ 0.011

Table 4: Ablation studies on parameterized graph augmentation schemes (ROC-AUC(%)).

Model variants	Topo.	Attr.	BBBP	BACE	Tox21	ToxCast	ClinTox	Sider
MolCLE-TA-AA	rand.	rand.	91.25 $\pm$ 0.051	75.83 $\pm$ 0.035	78.81 $\pm$ 0.039	62.34 $\pm$ 0.026	85.26 $\pm$ 0.057	60.97 $\pm$ 0.044
MolCLE-TA	rand.	Param.	91.85 $\pm$ 0.012	76.48 $\pm$ 0.039	79.17 $\pm$ 0.044	64.53 $\pm$ 0.031	86.69 $\pm$ 0.024	61.39 $\pm$ 0.023
MolCLE-AA	Param.	rand.	92.43 $\pm$ 0.052	90.96 $\pm$ 0.048	79.89 $\pm$ 0.018	68.78 $\pm$ 0.020	91.95 $\pm$ 0.034	62.14 $\pm$ 0.031
MolCLE	Param.	Param.	92.66 $\pm$ 0.037	91.71 $\pm$ 0.045	80.57 $\pm$ 0.091	69.01 $\pm$ 0.016	92.51 $\pm$ 0.021	62.67 $\pm$ 0.011

4.3 (RQ2) Ablation studies on parameterized graph augmentation schemes

To answer this research question, we conduct ablation studies regarding our proposed parameterized graph augmentation schemes, where we substitute the parameterized topological and attributive augmentation with random edge perturbation and attribute masking. The experimental results are shown in Table 4. Obviously, our two parameterized augmentation components regarding topology and attribute all achieve performance improvements. Comparing *MolCLE-TA* with *MolCLE-AA*, we can conclude that the parameterized topological augmentation plays a more important role than the attributive one. The model variant without both these two parameterized schemes obtains the poorest performance which indicates the superiority of our proposed two augmentation schemes.

4.4 (RQ3) Meaningful augmented graph views

Here we provide many molecule examples to illustrate what augmented graph views our MolCLE model generates. Figure 3 shows two augmented views for each of the three molecular graphs in a column. As shown in this figure, we can find that, although the two augmented views are very similar, they still remain slightly different, which demonstrates our proposed parameterized graph augmentation can generate effective positive pairs. Besides, the selected fragments for each molecular graph are obviously preserving chemical meanings. As shown in Figure 3, effective functional groups like benzene ring, carbonyl, hydroxy, carboxyl and halogen are all detected. On the opposite, futile fragments like carbon chains are ignored. Finally, we can notice that all of the augmented graph views keep a certain ratio with the original topological size, instead of select only a few bonds or all bonds.

4.5 (RQ4) Explaining chemical intuitions from constrained augmentations

We demonstrate that MolCLE generates explainable augmentations with chemical meanings by case studies on the *BBBP* dataset. Various visualization examples verify the effectiveness of the proposed constraints to preserve chemical priors in generated augmentations automatically, as summarized in Figure 4.

We first concentrate on the most important aromatic rings in organic chemistry. On the left column of Figure 4, MolCLE learns to generate graph views preserving aromaticity, which allows contrastive training to learn the correct semantic information of the original molecule. Secondly, we demonstrate that, compared to generated graph views without constraints, MolCLE learns to cover the most meaningful functional groups in the molecule. As shown in the middle column in Figure 4, MolCLE could focus on both aromatic rings and the other amine group on the other side of the molecule. This result is in agreement with existing work that the property of the original molecule could be considered of the summation of the properties of its functional groups [9]. In addition, we find that molecule augmentations learned from MolCLE are decent molecular backbone fragments, similar to what the rule-based molecular fragmentation approach obtained [33]. This phenomenon reveals that our proposed parameterized graph augmentation scheme could decompose molecules from the connecting atoms, usually alkyl groups, to obtain a set of functional group fragments. As a matter of fact, molecular virtual fragmentation has proved to be an efficient way for *in silico* screening the chemical space in drug discovery, indirectly proving the effectiveness of our proposed parameterized graph augmentation scheme.

Figure 3: Generated views from two topological augmentation components in MolCLE. The two graph views from the same molecule serve as positive pairs during contrastive learning.

Figure 4: A comparison of topological augmentations learned from MolCLE without (blue) and with (red) the proposed constraints. With constraints, MolCLE (a) tends to generate augmentations with closed aromatic rings, such as phenyl, imidazolyl, and pyridineyl groups; (b) learns to cover all functional groups rather than the linkage atoms, e.g., selecting amine groups and aromatic rings on both sides of the molecule but not the alkyl groups between them; (c) decomposes a whole molecule into non-overlapping but property-preserving functional groups by bonds even without linkage atoms between functional groups.

In conclusion, MolCLE could learn explainable molecule augmentations with chemical intuitions, providing a novel method to learn task-agnostic molecule embeddings from unlabeled data.

5 CONCLUSION

In this paper, we have developed a novel molecular graph contrastive learning framework with parameterized explainable augmentations. We pre-train our MolCLE model by maximizing the agreement between two augmented graph views from the same molecule. The proposed parameterized graph augmentation scheme self-adaptively learns topological and attributive augmentations for each molecule, which preserves the domain priors to make all augmented graph views chemically meaningful. The augmentation scheme can capture implicit graph connectivity patterns with

chemical intuitions, and meanwhile, emphasize the underlying semantic information by selecting important feature dimensions. Our model outperforms many supervised counterparts and achieves a considerable performance compared to other state-of-the-art baselines. We also analyze the explainability of our generated graph views by case studies. Comprehensive experiments across multiple real-world datasets demonstrate the superiority of our model. In our future works, we will explore more contrastive learning frameworks and tailor a best-matched one for molecular graph data. Besides, we will try more downstream tasks like quantum chemical properties of molecules for better validate the generalizability and transferability of our proposed MolCLE model.

REFERENCES

- [1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340* (2017).
- [2] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910* (2019).
- [3] Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* 39, 15 (1996), 2887–2893.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [5] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 767–776.
- [6] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [8] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292* (2015).
- [9] Peter Ertl, Bernhard Rohde, and Paul Selzer. 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of medicinal chemistry* 43, 20 (2000), 3714–3717.
- [10] William Falcon and Kyunghyun Cho. 2020. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104* (2020).
- [11] Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428* (2019).
- [12] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*. PMLR, 1263–1272.
- [13] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* 9, 3 (2006), 199.
- [14] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.
- [15] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 6645–6649.
- [16] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.
- [17] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216* (2017).
- [18] Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*. PMLR, 4116–4126.
- [19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *European conference on computer vision*. Springer, 630–645.
- [21] Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*. PMLR, 4182–4192.
- [22] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).
- [23] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2019. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).
- [24] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* 58, 1 (2018), 27–35.
- [25] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
- [26] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. 2011. *Random graphs*. Vol. 45. John Wiley & Sons.
- [27] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30, 8 (2016), 595–608.
- [28] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [29] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [30] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [31] Johannes Klicpera, Stefan Weyenberger, and Stephan Günnemann. 2019. Diffusion improves graph learning. *arXiv preprint arXiv:1911.05485* (2019).
- [32] Ralph Linsker. 1988. Self-organization in a perceptual network. *Computer* 21, 3 (1988), 105–117.
- [33] Tairan Liu, Misagh Naderi, Chris Alvin, Supratik Mukhopadhyay, and Michal Brylinski. 2017. Break down in order to build up: decomposing small molecules for fragment-based drug design with e molfrag. *Journal of chemical information and modeling* 57, 4 (2017), 627–631.
- [34] Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1052–1060.
- [35] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573* (2020).
- [36] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).
- [37] Nicholas Metropolis and Stanislaw Ulam. 1949. The monte carlo method. *Journal of the American statistical association* 44, 247 (1949), 335–341.
- [38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546* (2013).
- [39] Harry L Morgan. 1965. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation* 5, 2 (1965), 107–113.
- [40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
- [42] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020. Graph representation learning via graphical mutual information maximization. In *Proceedings of The Web Conference 2020*. 259–270.
- [43] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.
- [44] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* (2015).
- [45] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 385–394.
- [46] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.
- [47] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems* 33 (2020).
- [48] Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566* (2017).
- [49] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 1857–1865.
- [50] Teague Sterling and John J Irwin. 2015. ZINC 15—ligand discovery for everyone. *Journal of chemical information and modeling* 55, 11 (2015), 2324–2337.
- [51] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. 2019. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000* (2019).
- [52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849* (2019).
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

- [54] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [55] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep Graph Infomax. In *ICLR (Poster)*.
- [56] Izhar Wallach, Michael Dzamba, and Abraham Heifets. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).
- [57] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. 2020. Multi-view Graph Contrastive Representation Learning for Drug-Drug Interaction Prediction. *arXiv preprint arXiv:2010.11711* (2020).
- [58] David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences* 29, 2 (1989), 97–101.
- [59] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. 2020. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149* (2020).
- [60] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* 9, 2 (2018), 513–530.
- [61] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3733–3742.
- [62] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. 2020. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659* (2020).
- [63] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry* 63, 16 (2019), 8749–8760.
- [64] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
- [65] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2017. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*. 285–294.
- [66] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* 59, 8 (2019), 3370–3388.
- [67] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6210–6219.
- [68] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894* (2019).
- [69] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* 33 (2020).
- [70] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. Xgmn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 430–438.
- [71] Kai Zhang, Yaokang Zhu, Jun Wang, and Jie Zhang. 2019. Adaptive structural fingerprints for graph attention networks. In *International Conference on Learning Representations*.
- [72] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep Graph Contrastive Representation Learning. In *ICML Workshop on Graph Representation Learning and Beyond*. <http://arxiv.org/abs/2006.04131>
- [73] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph Contrastive Learning with Adaptive Augmentation. In *Proceedings of The Web Conference 2021 (Ljubljana, Slovenia) (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3442381.3449802>