A Probability Contrastive Learning Framework for 3D Molecular Representation Learning

Anonymous Authors[1]

Abstract

Contrastive Learning (CL) has been applied to extract meaningful representations from molecules, facilitating various applications in molecular property prediction and drug design. However, the straightforward application of standard CL to molecular datasets may result in suboptimal performance, primarily due to the false positive/negative pairs introduced by conventional graph augmentations like node masking and subgraph removal. To address this challenge, we propose a novel probability CL framework that uses learnable weight distribution to alleviate the impact of false positive/negative pairs and allow for effective optimization through stochastic expectation maximization. Moreover, to incorporate the 3D structural information and make the CL framework 3D-aware, we adopt a transformer based encoder to integrate the 3D coordinates as input, develop a compatible molecular augmentation method called position noise injection and utilize additional 3D recovery loss. The experimental results indicate that our method outperforms existing approaches in 13 out of 15 molecular property prediction benchmarks, achieving new state-of-the-art results in average. Additionally, it excels in the protein-ligand binding task compared to standard contrastive learning and other unsupervised learning methods, underscoring its potential in practical drug design.

1. Introduction

We investigate the problem of learning representations from molecules. Molecular representation learning (MRL) has gained tremendous attention due to its critical role in learning from limited supervised data for applications such as molecular property prediction(Rong et al., 2020; Wang et al., 2022; Fang et al., 2022) and drug design.(Koukos et al., 2019; Liu et al., 2022; Méndez-Lucio et al., 2021) The model aims to learn generic representations from various augmentations of the molecules that could benefit downstream applications.

With the success of contrastive learning method in computer vision and multi-modality pretraining(He et al., 2020; Radford et al., 2021), a variety of contrastive learning methods have been proposed for molecular representation learning. MolCLR(Wang et al., 2022) introduces a contrastive learning framework for molecular representation learning, utilizing atom masking and edge removal as data augmentation and enhancing the performance of GNN models on various downstream molecular property prediction benchmarks. GraphMVP (Liu et al., 2022) considers both 2D topology and 3D geometry during pre-training, although its downstream tasks only require 2D topology.

Although existing works have proven contrastive learning to succeed in learning molecular representations. However, we propose that it still faces two drawbacks.

First, despite the proven effectiveness of contrastive loss in empirical applications for molecular representation learning, a lingering question has been largely overlooked in prior works. That is, the reliability of the "positive" and "negative" labels in augmented molecule pairs raises concerns. Most augmentations applied to molecular datasets involve removing parts of the molecular graph, such as nodes, edges, and subgraphs. Within the entire molecular dataset, there could be multiple molecules with similar structures and chemical properties labeled as negative pairs. In essence, due to their extensive volume and numerous augmentation processes, molecular datasets naturally contain numerous falsely aligned pairs.

Illustratively, Figure 1 provides an example of false positives/negatives resulting from graph augmentations in MolCLR (Wang et al., 2022). In this context, we employ two distinct graph augmentations to enhance two disparate molecules. The augmented molecule pair originating from the same molecule is categorized as positive, while other molecule pairs within the same batch are considered negative. However, as illustrated in the figure, two molecules augmented using the same augmentation method should indeed be regarded as positive, given their structural similarity. Similarly, the same molecule augmented using different augmentation methods is structurally negative, yet it is incorrectly treated as instances of positive pairs.

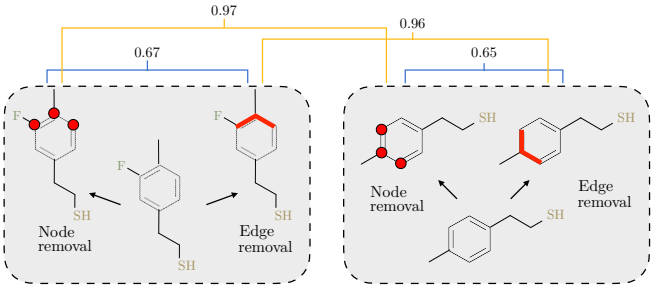

Figure 1. **Existing problem in molecular contrastive learning.** Adopt node removal and edge removal for molecular contrastive learning can lead to false positive and false negative problems. Blue lines indicate positive pairs and yellowing lines indicate negative pairs. The numbers on each line indicate the chemical similarity between the augmented pair of molecules. In this case, positive pairs indeed have lower similarity than negative pairs.

Second, existing molecular contrastive learning methods are weak in leveraging the 3D geometric information and thus are seldom used for 3D related molecular tasks. From the input side, most molecular contrastive learning methods utilize graph neural networks (GNNs) as molecular feature extractors, which have limitations in capturing overall dependencies within the input molecule and typically do not incorporate 3D positions as part of their input. From the data augmentation side, existing data augmentation methods are designed for 2D molecular graphs. From output side, the training objective of contrastive learning is not inherently 3D aware, making the knowledge it discovers less generalized in 3D space.

To overcome the aforementioned difficulties, this paper introduces modifications to existing frameworks for molecular contrastive learning. In addressing issues related to false positive and false negative pairs, we present a systematic approach to molecular contrastive learning by redefining it within a probability framework. This involves incorporating random weights for data pairs. Utilizing Bayesian methods, these random weights can be accurately determined through sampling, and the model parameters can be efficiently optimized using stochastic expectation maximization.

To extend the application of contrastive loss to 3D tasks in drug design, we introduce the following modifications: First, we employ a molecule encoder based on the Transformer architecture instead of GNN to embed 3D global context into molecular features, taking both atom type and atom 3D coordinates as input. This choice not only enhances the model's understanding of the molecular 3D geometry but also ensures scalability to larger datasets and more intricate molecular architectures. Second, we also adopt a new 3D data augmentation technique called position noise injection to generate positive and negative pairs for molecular contrastive learning. Position noise injection randomly injects noise into the 3D coordinates of one or more atoms in a molecule. This method mimics the real world chemical reactions and prompts the model to learn correlations between the involvements of one molecule in various reactions. Third, we adopt two additional 3D loss functions, masked atom prediction loss and 3D position recovery loss, as a supplementary for contrastive learning loss function.

We pretrain our model on two large scale datasets, one molecular dataset and the other protein pocket dataset, and then evaluate its performance on molecular property prediction tasks MoleculeNet (Wu et al., 2018) and protein-ligand binding tasks(Koukos et al., 2019). With molecular property prediction tasks, we aim to test our model's ability in extracting useful features from molecular and with protein-ligand binding tasks, we aim to test the model's ability in a real-world 3D drug design task. Experiments show that our method outperforms all other molecular representation learning baselines, including contrastive and non-contrastive methods.

The contribution of this paper is summarized as:

- To address issues related to false positive and false negative pairs, we present a systematic approach to molecular contrastive learning by redefining it within a probability framework. We propose a novel optimization algorithm based on Bayesian method and stochastic expectation maximization.
- We investigate 3D molecular contrastive learning by explicitly using 3D coordinate as input to Transformer encoder.
- We design a new 3D augmentation method called *position noise injection* to generate positive and negative pairs for moleculars, aiding the model in learning correlations between the involvements of one molecule in various reactions.
- We adopt two additional 3D constraint loss functions to help contrastive loss more generalizable to 3D space.

2. Methods

Following existing works, we begin by elucidating the foundational setup and notation in molecular contrastive learning. In learning, one randomly samples a batch of N molecules. Subsequently, employing stochastic augmentation strategies, each molecule sample $\mathbf{x}_i$ is transformed into two augmented molecules, denoted as $(\mathbf{x}_i, \mathbf{x'}_i)$. Among these $2N$ augmented molecules, $(\mathbf{x}_i, \mathbf{x'}_i)$ is considered as positives and the other $2(N-1)$ in the same batch as negative samples. After that, a neural network encoder $f(\mathbf{x}; \boldsymbol{\theta})$ parameterized by θ is adopted to extract representation vectors z from augmented molecular examples.

Let $s_{i^+} \triangleq \mathrm{sim}(\mathbf{z}_i, \mathbf{z'}_i)$ represent the similarity score between the positive pair $(\mathbf{x}_i, \mathbf{x'}_i)$ after the encoder, and $s_{ik^-} \triangleq \mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)$ signifies the similarity score between the negative pair $(\mathbf{x}_i, \mathbf{x}_k)$, and $\mathrm{sim}(\cdot,\cdot)$ represents any positive-valued similarity metric. In this paper, we adopt the commonly used exponential cosine similarity, defined as $\mathrm{sim}(\mathbf{z}_1, \mathbf{z}_2) \triangleq e^{\mathbf{z}_1^T \mathbf{z}_2 / \|\mathbf{z}_1\| \|\mathbf{z}_2\| \tau}$, where τ denotes a temperature parameter.

2.1. Probability Weighted Contrastive Learning

In standard contrastive learning, one tries to encode data samples to a latent space such that positive pairs stay close to each other while negative pairs are pushed away. The idea is described by the following loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{k=1}^{N}[\ell(2k-1,2k) + \ell(2k,2k-1)], \text{with} \\ \ell(i,j) = -\log \frac{s_{i^+}}{s_{i^+} + \sum_{k=1}^{2N} \mathbb{I}_{[k\neq i,j]} s_{i,k^-}} \tag{1}$$

However, one issue of directly applying the contrastive learning into molecular representation learning is the potential false positive positive and negative molecular pairs, as discussed in the introduction. This could confuse the learning, ending up with sub-optimal representations. Is there a way to automatically identify and differentiate these pair data? In the following, we propose a Bayesian approach to address this issue that allows the algorithm for automatic inference of the degree of positiveness and negativeness of data pairs, involving enhancing the standard contrastive loss by incorporating learnable stochastic weights for all data pairs. To be more specific, we introduce local learnable weights, denoted as w_i^+ for each positive pair and w_{ik}^- for each negative pair. We then define a weighted contrastive loss based on these introduced weights. This modification aims to mitigate the issues by automatically assigning relatively lower weights (or no weights) to false positive and false negative pairs;

$$\mathcal{L}_w = \frac{1}{N}\sum_{k=1}^{N}[\bar{\ell}(2k-1,2k) + \bar{\ell}(2k,2k-1)], \text{with} \\ \bar{\ell}(i,j) = -\log \frac{w_i^+ s_{i^+}}{w_i^+ s_{i^+} + \sum_{k=1}^{2N} \mathbb{I}_{[k\neq i,j]} w_{ik}^- s_{ik^-}} \tag{2}$$

One problem with the above formulation, however, is that it is not realistic to compute and store all the weights in the learning process. This precaution arises from the quadratic growth in the number of weights to be calculated as the training data size increases. Furthermore, the random nature of our augmentation method further adds complexity to the pre-calculation and storage of these weights.

A straightforward baseline for calculating these weights can be envisioned as follows: we can consider these weights in a binary fashion, with all weights initialized to one. In the learning process, if for some positive pairs the similarity score falls below a specified threshold, we set the corresponding weights to zero, marking these positive pairs as false positives. Conversely, if for some negative pairs the similarity score exceeds a threshold, we set the associated weights to zero, indicating false negatives. A challenge associated with this baseline method, however, lies in the establishment of a rigid similarity threshold to create a binary division of weights between zero and one. This approach proves less suitable for our molecular contrastive task as these heuristically chosen thresholds might not be optimal.

To address this challenge, we propose a principled Bayesian approach that allows adaptively inferring the optimal weights by Bayesian inference. Specifically, we treat the weights to be random variables and assign appropriate priors to them. We consider two types of priors: a Bernoulli prior to model weights as binary random variables and a Gamma prior to represent them as positive values. For simplicity, we model positive weights using the Gamma distribution and negative weights using either the Gamma distribution or the Bernoulli distribution, as expressed by the following formulas:

Option 1 - Gamma priors for continuous weighting:

$$w_i^+ \sim \mathrm{Gamma}(a_+, b_+), w_{ik}^- \sim \mathrm{Gamma}(a_-, b_-).$$

Option 2 - Bernoulli priors for selective weighting:

$$w_i^+ \sim \mathrm{Gamma}(a_+, b_+), \quad w_{ik}^- \sim \mathrm{Bernoulli}(a_-^-).$$

here, a_+, b_+, a_- and b_- are shape and rate parameters for Gamma distribution and a_-^- is the probability parameter for Bernoulli distribution.

With our reformulation, we can define a joint distribution over the global model parameter and local random weight variables w_i^+ and w_{ik}^-, as:

$$p\left(\{w_i^+\}, \{w_{ik}^-\}, \boldsymbol{\theta}; \mathcal{D}\right) \\ \propto \prod_{\mathbf{x}_i \in \mathcal{D}} \frac{w_i^+ s_{i^+}}{w_i^+ s_{ij^+} + \sum_{k=1}^{K} w_{ik}^- s_{ik^-}} p(\{w_i^+\}) p(\{w_{ik}^-\}) p(\boldsymbol{\theta}). \tag{3}$$

One problem with the above formulation, however, is that posterior inference of the weights is challenging, due to the lack of convenience posterior distributions.

Fortunately, inspired by (Chen et al., 2022), we can introduce an augmented random variable u_i that is associated to data point $\mathbf{x}_i$. Consequently, we can define an augmented joint posterior distribution of the random variables $\boldsymbol{\theta}, \mathbf{u}, \mathbf{w}$,

denoted as $p\left(\left\{w_i^+\right\}, \left\{w_{ik}^-\right\}, \boldsymbol{\theta} \mid \mathcal{D}\right)$ [1], to be

$$p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \mid \mathcal{D}) \propto \prod_{i:\mathbf{x}_i \in \mathcal{D}} w_i^+ s_i + e^{-\mathbf{u}_i w_i^+ s_i +} \prod_k e^{-u_i w_{ik}^- s_{ik}-} p\left(\left\{w_i^+\right\}\right) p\left(\left\{w_{ik}^-\right\}\right) p(\boldsymbol{\theta}), \tag{4}$$

where $\mathbf{u} \triangleq \left\{u_1, u_2, \cdots, u_{|\mathcal{D}|}\right\}$ and $\mathbf{w} \triangleq \left\{w_i^+\right\} \cup \left\{w_{ik}^-\right\}$. It is worth noting that this joint distribution is equivalent to the original distribution (3), because (3) is recovered if one marginalize out the auxiliary random variables $\mathbf{u}$ in (4). In other words, optimization thought (4) is equivalent to optimization over (3). Consequently, we can perform learning and inference based on the augmented posterior of $p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \mid \mathcal{D})$, which preserves a much convenient form for posterior inference. In the following, we propose an efficient algorithm based on stochastic expectation maximization (stochastic EM) to alternatively infer the local random variables $\mathbf{w}$ and optimize the global model parameter $\boldsymbol{\theta}$.

2.2. Efficient Inference and Learning with Stocastic Expectation Maximization

We propose a stochastic EM algorithm for efficient inference and learning of our model. Stochastic EM is a stochastic variant of the EM algorithm, which is an iterative method for finding the maximum likelihood of model parameters in statistical models when data is only partially, or when model depends on unobserved latent variables.

In our setting, the objective of stocastic EM is to maximize the posterior in equation 4. The basic idea is to alternatively 1) optimizing model parameter $\boldsymbol{\theta}$ with fixed $(\mathbf{u}, \mathbf{w})$ and 2) sampling $(\mathbf{u}, \mathbf{w})$ with fixed $\boldsymbol{\theta}$. To this end, we follow standard procedures in stochastic EM to divide the learning into three steps: Simulation, Stochastic Expectation, and Maximization. Specifically, simulation corresponds to sampling local random variables $\mathbf{u}$ and w for a batch of data; stochastic expectation then uses the sampled auxiliary random variables to update the model parameter $\boldsymbol{\theta}$ by maximizing a stochastic objective $Q(\boldsymbol{\theta})$, defined as: $Q_{t+1}(\boldsymbol{\theta}) = Q_t(\boldsymbol{\theta}) + \lambda_t \left(\log p(\boldsymbol{\theta}, \mathbf{u}, \mathbf{w} \mid \mathcal{D}) - Q_t(\boldsymbol{\theta})\right)$ at iteration $t+1$, where $\{\lambda_t\}$ is a sequence of decreasing weights. In the following, we detail the three steps.

Simulation Given the joint posterior distribution in equation 3 and the current batch of data, the posterior distributions of the local random variables $\mathbf{u}$ and $\mathbf{w}$ can be directly read out, which simply follow Gamma or Bornoulli distributions of the following forms:

$$u_i \mid \left\{w_i^+, w_{ik}^-, \boldsymbol{\theta}\right\} \sim \text{Gamma}\left(a_u, b_u + w_i^+ s_{i^+} + \sum w_{ik}^- s_{ik^-}\right), \forall i, \text{ and}$$

$$w_i^+ \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Gamma}\left(1 + a_+, u_i s_{i+} + b_+\right), \text{and}$$

$$\text{Option 1: } w_{ik}^- \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Gamma}\left(a_-, u_i s_{ik^-} + b_-\right), \forall i, k$$

$$\text{Option 2: } w_{ik}^- \mid \{\mathbf{u}, \boldsymbol{\theta}\} \sim \text{Bernoulli}\left(\frac{a_- e^{-u_i s_{ik-}}}{1 - a_- + a_- e^{-u_i s_{ik-}}}\right)$$

Stochastic Expectation We then proceed to calculate the stochastic expectation based on the simulated local random variables above. For notation simplicity, we define $Q_0(\boldsymbol{\theta}) = 0$. Then we can reformulate $Q_{t+1}(\boldsymbol{\theta})$ by decomposing the recursion, resulting in

$$Q_{t+1}(\boldsymbol{\theta}) = \sum_{\tau=0}^{t} \tilde{\lambda}_\tau \log p\left(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \mid \mathcal{D}_\tau\right), \quad \text{where } \tilde{\lambda}_\tau \triangleq \lambda_\tau \prod_{t'=\tau+1}^{t} \left(1 - \lambda_{t'}\right), \tag{5}$$

where τ indexes the minibatch and the corresponding local random variables at the current time τ.

Maximization The stochastic expectation objective provides a convenient form for stochastic optimization over time, similar to online optimization (Bent & Van Hentenryck, 2005). Specifically, at each time t, we can initialize the parameter θ from the last step, and update it by stochastic gradient ascent on the log-likelihood, $\log p\left(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \mid \mathcal{D}_\tau\right)$ calculated from the current batch of data. To reduce variance, we propose to optimize a marginal version by integrating out $\mathbf{u}_\tau$ from $p\left(\boldsymbol{\theta}, \mathbf{u}_\tau, \mathbf{w}_\tau \mid \mathcal{D}_\tau\right)$, which essentially reduces to our original weighted contrastive loss in equation 1. With the above steps, it is ready to optimize the model by stochastic EM. The detailed steps are described in Algorithm 1.

2.3. Data Augmentation

An essential aspect of contrastive learning involves employing data augmentation to generate positive and negative pairs. Given our emphasis on graph data, the data augmentation methods employed differ from those utilized in other studies focusing on image or text data.

Within our specific framework, we implement three strategies for augmenting input molecules represented as molecule graphs: atom masking, position noise injection, and sub-molecule destruction.

Atom Masking This type of augmentation has been widely used in existing molecular contrastive learning frameworks. It aims to mask atoms in a molecule with a specific ratio. When an atom is masked, its atom type is changed into a distinct mask token [MASK], and its feature is also substituted

[1] In the sense that marginalizing over the augmented random variables $\left\{w_i^+\right\}$and $\left\{w_{ik}^-\right\}$in $p\left(\theta, \mathbf{U}, \left\{w_i^+\right\}, \left\{w_{ik}^-\right\} \mid \mathcal{D}\right)$ gives back to the original $p\left(\left\{w_i^+\right\}, \left\{w_{ik}^-\right\}, \boldsymbol{\theta}; \mathcal{D}\right)$. Thus, learning and inferences on the two forms are equivalent.

Algorithm 1 Contrastive Learning with Stochastic EM

Initialize θ; set $t = 1$
for a batch of molecules in loader **do**
  Augment each molecule $\mathbf{x}_i$ into a pair $(\mathbf{x}_i, \mathbf{x'}_i)$
  Calculate positive/negative similarity scores s^+ and s^- for all the molecule pairs
  Initialize all the weights w^+ and w^- to be one
  for $k = 1$ to iter [4 in practice] **do**
    Sample u and w according to distributions
  end for
  Calculate the weighted contrastive loss in equation 2 with the sampled w on the current batch of data
  Update the model parameter by stochastic gradient descent with the calculated weighted contrastive loss
  $t = t + 1$
end for

correspondingly. This masking process compels the model to grasp the inherent atom information within molecules.

Position Noise Injection When involved in a chemical reaction, the 3D positions of a molecule are always corrupted. This occurs as the first step in a reaction, which invariably involves the breaking and forming of chemical bonds. Such processes typically lead to changes in the positions of atoms within a molecule.

To emulate real-world reactions, we propose a novel data augmentation method named Position Noise Injection. This method entails introducing random noise into the input 3D coordinates of the atoms at a specified ratio. To prevent excessively noisy positions, which would render the learning process unfeasible, we constrain the strength of the injected random noise (n) to be less than or equal to 1 $\mathring{A}$.

Sub-Molecule Destruction The process of sub-molecule destruction may be conceptualized as a fusion of atom masking and position noise injection. Sub-molecule destruction commences with the selection of a randomly chosen origin atom. The destruction procedure starts by masking the nearest neighbors of the initial atom, followed by the nearest neighbors of those neighbors, until the count of masked atoms attains a specified ratio relative to the total number of atoms. Subsequently, noise is introduced to the positions of the masked atoms. In all experimental scenarios, we adhere to the default setting, wherein 25 percent of the atoms in a molecule are masked and destroyed.

2.4. Backbone Transformer and Additional Loss Function

In the realm of learning molecular representations, there are two widely recognized backbone models: graph neural networks (GNN) (Hu et al., 2020; Li et al., 2021), and Transformer (Ying et al., 2021). When GNN serves as the backbone model, locally connected graphs are commonly employed to depict molecules for efficiency reasons. Nevertheless, these locally connected graphs typically do not include 3D positions of molecules, GNNs also fall short in capturing long-range interactions among atoms. Recognizing the significance of long-range interactions in molecular representation learning, also convinced by the previous success of Transformer-based encoder, we opt for Transformer as the backbone model. The Transformer fully connects the nodes, enabling it to effectively learn potential long-range interactions.

In accordance with the latest advancements as seen in Uni-Mol(Zhou et al., 2023), we employ a molecule encoder based on the Transformer architecture. This encoder takes two inputs: atom types and atom coordinates. We straightforwardly use the representation of the [CLS] token as the final encoded representation in our contrastive learning framework, signifying the entire molecule.

We have integrated two additional loss functions into our framework: the masked atom recovery loss and the position recovery loss.

Simultaneously, the position recovery loss aims to restore the accurate positions in the presence of injected noise. As atom positions are equivariant to translation and rotation, we employ an equivariant head(Satorras et al., 2021), to predict the precise position of the atom.

3. Related works

Contrastive learning (CL) As a popular self-supervised learning paradigm, CL focuses on learning semantically informative representations for downstream tasks (Li et al., 2022; Chuang et al., 2020; You et al., 2020; Hu et al., 2022). The most widely used loss function is InfoNCE (van den Oord et al., 2018) which pulls in the representations between positive sample pairs while pushing away that between negative sample pairs.

Molecular representation learning Representation learning on large-scale unlabeled molecules attracts much attention recently. SMILES-BERT (Wang et al., 2019) is pretrained on SMILES strings of molecules using BERT. Subsequent works are mostly pretraining on 2D molecular topological graphs (Li et al., 2021; Rong et al., 2020). MolCLR (Wang et al., 2022) applies data augmentation to molecular graphs at both node and graph levels, using a self-supervised contrastive learning strategy to learn molecular representations. Further, several recent works try to leverage the 3D spatial information of molecules, and focus on contrastive or transfer learning between 2D topology and 3D geometry of molecules. For example, GraphMVP (Liu et al., 2022) proposes a contrastive learning GNN-based framework between 2D topology and 3D geometry. GEM (Fang

et al., 2022) uses bond angles and bond length as additional edge attributes to enhance 3D information. Uni-Mol(Zhou et al., 2023) is a universal 3D molecular pretraining framework that significantly enlarges the representation ability and application scope in drug design.

Protein-ligand binding pose prediction Docking interaction between protein and ligand is one of the most important things to understand in structure-based drug design because it allows for leading identification and guiding molecular optimization, and has been developed for the past decades. Tools such as AutoDock4 (Morris et al., 2009), AutoDock Vina (Trott & Olson, 2010; Eberhardt et al., 2021), and Smina (Koes et al., 2013) are among the most used docking programs. Also, machine learning-based docking methods, such as $\Delta_{Vina}\mathrm{RF}_{20}$ (Wang & Zhang, 2017) and DeepDock (Méndez-Lucio et al., 2021) have also been developed to predict proteinligand binding poses and assess protein-ligand binding affinity. Equibind (Stärk et al., 2022) is a recent graph deep learning based methods.

Noisy Pairs in Contrastive Learning: Noisy data pair problem have been found and studied in visual contrastive learning community. NLIP (Huang et al., 2023) enforces the pairs with larger noise probability to have fewer similarities in embedding space to improve the model training. (Han et al., 2022) apply noise estimation component to adjust the consistency between different modalities for the action recognition task. RINCE (Hoffmann et al., 2022) uses a ranked ordering of positive samples to improve InfoNCE loss. Yet another line of research studies the false positive pair problem in multi-view contrastive learning, which aims to handle misalignment between multi-view data from multiple input (Zhang et al., 2019; Poklukar et al., 2022; Yang et al., 2022). DCP (Lin et al., 2022) leverages the maximization of mutual information to conduct consistency learning across different views. MFLVC (Xu et al., 2022) proposes to learn multi-level features for multiple views. DSIMVC (Tang & Liu, 2022) establishes a theoretical framework to reduce the risk of clustering performance degradation from semantic inconsistent views. Although satisfactory results are achieved in many cases, the noisy data pair problem in molecular contrastive learning literature has not been well studied and solved.

Stochastic Expectation Maximization Stochastic EM (Nielsen, 2000) stands as a pivotal algorithm in machine learning and probabilistic modeling for large-scale Bayesian inference. Building upon the foundations of the classical Expectation-Maximization (EM) algorithm (Lin, 2011), Stochastic EM offers an efficient solution for parameter estimation in situations involving vast datasets or latent variables, e.g., to maximize the log-likelihood of $p(\mathbf{z}, \mathcal{D} \mid \boldsymbol{\theta})$, where $\mathcal{D}$ is the dataset, $\mathbf{z}$ is the local random variable and $\boldsymbol{\theta}$ is the global model parameter. By leveraging the power of mini-batch sampling, Stochastic EM strikes a balance between computational scalability and estimation accuracy. It has found widespread utility in various domains, including clustering (Allassonnière & Chevallier, 2021), topic modeling (Zaheer et al., 2016), and latent variable modeling(Zhang & Chen, 2020), making it an indispensable tool to cope with complex probabilistic models and extensive data and a natural fit to our problem.

4. Experiments

In this section, we conduct our experiments on two different tasks, molecular property prediction, and protein ligand bounding pose estimation, and then compare our method with Uni-Mol, the current state of the art method in molecular representation learning, and other strong baselines. We first introduce the pretraining stage , then move on to finetuning on downstream tasks.

4.1. Large-Scale Dataset in Pretraining

For the purpose of pretraining, we follow Uni-Mol to pretrain on two large-scale datasets, one composed of organic molecules, and another composed of protein pockets. Two models are pretrained using these two datasets respectively. As pockets are directly involved in many drug design tasks, intuitively, the pretraining on candidate protein pockets can boost the performance of tasks related to protein-ligand structures and interactions.

The molecular pretraining dataset is based on multiple public datasets. After normalizing and deduplicating, it contains about 19M molecules.

The protein pocket pretraining dataset is derived from the Protein Data Bank (RCSB PDB) (Berman et al., 2000), a collection of 180K 3D structures of proteins. Fpocket is used (Guilloux et al., 2009) to detect possible binding pockets of the proteins. In this way, We have a dataset of 3.2M candidate pockets for pretraining.

4.2. Molecular Property Prediction & Baselines

MoleculeNet (Wu et al., 2018) is a popular benchmark for molecular property prediction, including datasets focusing on different molecular properties, from quantum mechanics and physical chemistry to biophysics and physiology.

We compare our method with multiple baselines, including contrastive and non-contrastive baselines. Uni-Mol (Zhou et al., 2023) is current state of the art method in molecular property prediction, MolCLR (Wang et al., 2022) is a baseline model using standard contrastive learning method, and GEM (Fang et al., 2022)is another cutting-edge work. Random Forest and XGBoost (Chen & Guestrin, 2016) are used as predictors for downstream tasks.

Table 1. ROC_AUC on molecular property prediction classification tasks (Higher is better)

Datasets	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	PCBA	MUV	MEAN
# Molecules	2039	1513	1478	7831	8575	1427	41127	437929	93078	
# Tasks	1	1	2	12	617	27	1	128	17	
D-MPNN	71.0	80.9	90.6	75.9	65.5	57.0	77.1	86.2	78.6	75.9
Attentive FP	64.3	78.4	84.7	76.1	63.7	60.6	75.7	80.1	76.6	73.8
N-GramRF	69.7	77.9	77.5	74.3	–	66.8	77.2	–	76.9	74.3
N-GramXGB	69.1	79.1	87.5	75.8	–	65.5	78.7	–	74.8	75.8
PretrainGNN	68.7	84.5	72.6	78.1	65.7	62.7	79.9	86.0	81.3	75.8
GraphMVP	72.4	81.2	79.1	75.9	63.1	63.9	77.0	–	77.7	73.3
GEM	72.4	85.6	90.1	78.1	69.2	**67.2**	80.6	86.6	81.7	79.4
MolCLR	72.2	82.4	91.2	75.0	–	58.9	78.1	–	79.6	76.7
Uni-Mol	72.9	85.7	**91.9**	79.6	69.6	65.9	80.8	88.5	82.1	79.8
Ours (Gamma)	**76.7**	**88.2**	89.4	**80.1**	**69.9**	63.6	**83.0**	**89.6**	79.0	**80.0**
Ours (Bernoulli)	73.7	84.3	85.3	79.8	68.8	64.9	80.8	89.3	**82.9**	78.9

Table 2. Performance on molecular property prediction regression tasks (Lower is better)

Datasets	ESOL	FreeSolv	Lipo	QM7	QM8	QM9	MEAN (RMSE)	MEAN (MAE)
# Molecules	1128	642	4200	6830	21786	133885		
# Metric	RMSE↓			MAE↓				
D-MPNN	1.050	2.082	0.683	103.5	0.0190	0.00814	1.272	34.509
GROVERlarge	0.895	2.272	0.823	92.0	0.0224	0.00986	1.33	30.67
MolCLR	1.271	2.594	0.691	66.8	0.0178	-	1.519	-
GraphMVP	1.029	-	0.681	-	-	-	-	-
GEM	0.798	1.877	0.660	58.9	0.0171	0.00746	1.112	19.642
Uni-Mol	0.788	1.480	0.603	41.8	0.0156	0.00467	0.957	13.940
Ours (Gamma)	0.775	1.420	**0.590**	**38.5**	**0.0142**	**0.00395**	0.928	**12.839**
Ours (Bernoulli)	**0.664**	**1.358**	0.626	55.6	0.0154	0.0056	**0.883**	18.541

As indicated in Table 1, our method outperforms the standard contrastive learning baseline model MolCLR by a significant margin, demonstrating the effectiveness of our proposed approach to deal with augmentation noise. Additionally, we outperform Uni-Mol and GEM, the current state-of-the-art methods, with an average gain of 1.3 percent in classification tasks and 7.6 percent in regression tasks. This substantiates that our approach facilitates more flexible training with a higher tolerance for false positive and false negative data pairs, thereby enhancing the model's performance in molecular representation learning. Due to resource constraints, we trained the Bernoulli version of our model for only 420,000 steps, around half of the training steps for the Gamma prior version. This is the reason that our Bernoulli prior version is slightly worse. However, we anticipate that when training long enough, our Bernoulli prior version can quickly catch up, if not better than our Gamma prior version.

In summary, by mitigating the false positive and false negative pair problem in molecular contrastive learning, our method outperforms all previous MRL models in almost all property prediction tasks.

4.3. Protein-Ligand Binding Task

This is one of the most important tasks in structure based drug design. The task is to predict the complex structure of a protein binding site and a molecular ligand. We need to consider how ligand lays in the pocket, that is, the 6 degrees (3 rotations and 3 translations) of freedom of a rigid movement.

Following Uni-Mol, the molecular representation and pocket representation are firstly obtained from their own pretraining models by their own conformations; then, their representations are concatenated as the input of an additional 4-layer Transformer decoder, which is finetuned to learn the pair distances of all heavy atoms in molecule and pocket. Then, with the predicted pair-distance matrix as a scoring function, we first randomly place the ligand and then optimize the coordinates of its atoms by directly back-propagation the loss between current pair-distance and predicted pair-distance.

For the training data used in finetuning, we use PDBbind General set v.2020(Liu et al., 2015) (19,443 complexes).

We evaluate our method using the metric binding pose accuracy. Specifically, we keep the pocket conformation fixed, while the ligand conformation is fully flexible. We evaluate the RMSD(root mean squared distance) between the prediction and the ground truth. Following previous works, we use the percentage of results below predefined RMSD thresholds as metrics.

We compare our method with current state-of-the-art baselines, including Autodock Vina (Trott & Olson, 2010; Eberhardt et al., 2021), Vinardo (Quiroga & Villarreal, 2016), Smina (Koes et al., 2013), Autodock4 (Morris et al., 2009) and Uni-Mol (Zhou et al., 2023). The binding pose accuracy results are shown in Table 3. Not surprisingly, our model again outperforms all the baseline methods, achieving state-of-the-art results with our Gamma-prior version model.

Table 3. Performance on binding pose prediction.

Methods	1.0 Å	1.5 Å	2.0 Å	3.0 Å	5.0 Å
Autodock Vina	44.21	57.54	64.56	73.68	84.56
Vinardo	41.75	57.54	62.81	69.82	76.84
Smina	47.37	59.65	65.26	74.39	82.11
Autodock4	21.75	31.58	35.44	47.02	64.56
Uni-Mol	43.16	68.42	80.35	87.02	94.04
Ours (Bernoulli)	**48.77**	**70.18**	78.95	85.26	94.04
Ours (Gamma)	45.61	69.47	**80.70**	**88.42**	**96.84**

4.4. Qualitative analysis

Distribution of similarity scores Our method is largely motivated by the observation that previous MCL approaches neglect potential semantic dissimilarity between positive samples and that accounting for this phenomenon can improve learned molecule representations. In Figure 2(See Appendix A), we plot the distribution of similarity scores for both positive and negative samples. Figure 2 left reveals that our method yields larger similarity scores with lower variance for positive pairs compared to MolCLR baseline which uses standard contrastive learning method. Figure 2 right reveals that our method also mitigates the false negative problem in standard CL. It also shows that our method sometimes assigns lower similarity scores to positive pairs. While it may seem counter intuitive to assign lower similarity scores to positive samples, we argue that doing so is the very reason our method captures dissimilarity between positive pairs. By allowing some degree of alignment between the right set of negative examples, our method is able to minimize the inconsistencies between shared context of related positives and negatives. This in turn allows us to learn an overall more coherent representation space, resulting in increased robustness and downstream performance.

Abalation Study We conducted an ablation study to investigate the contributions of different components to our model's performance. Specifically, we ablated two components: the probability framework used to reformulate the standard contrastive loss, and the additional 3D aware loss functions.

Table 4. Abalation Study on BBBP dataset

Methods	ROC_AUC
Ours (Gamma)	**76.7**
- Additional loss	75.5
- Probability framework	73.2

Table 4 presents the results of the ablation study. As is evident, the removal of the additional loss component led to a decrease in ROC_AUC by 1.2 points. This implies that the additional loss component plays a crucial role in enhancing the model's performance. On the other hand, eliminating the probability framework component resulted in a more substantial decrease in ROC_AUC, specifically by 3.5 points. This suggests that reformulating the standard contrastive loss into a probability framework yields greater performance improvement for the model.

5. Conclusion

In this paper, we investigate an important yet unnoticeable limitation of molecular contrastive learning, where augmented graph data come with false positive and false negative data pairs. As a remedy, we propose a principled solution to molecular contrastive learning by reformulating it into a probability framework and introducing random weights for data pairs. With a Bayesian data augmentation technique, the random weights can be efficiently inferred via sampling, and the model parameter can be effectively optimized via stochastic expectation maximization.

We also extend molecular contrastive learning framework to 3D molecular tasks by explicitly using 3D coordinate as input of Transformer encoder, designing a new 3D augmentation method called position noise injection, adopting two additional 3D constraint loss functions to help contrastive loss more generalizable to 3D space.

The effectiveness of our innovative approach has been proven through rigorous evaluations on molecular property prediction and drug design benchmarks. The results also showcase the wide-ranging applicability and improved robustness of our proposed method over both standard contrastive learning method and non-contrastive learning method for learning molecular representations.

We believe our method is a valuable addition to the literature on molecular contrastive representation learning, which can further boost the performance of state-of-the-art molecular representation learning models for drug design.

References

Allassonnière, S. and Chevallier, J. A new class of stochastic em algorithms: Escaping local maxima and handling intractable sampling. *Computational Statistics & Data Analysis*, 159:107159, 2021.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Research*, 28(1): 235–242, 2000.

Chen, C., Zhang, J., Xu, Y., Chen, L., Duan, J., Chen, Y., Tran, S., Zeng, B., and Chilimbi, T. Why do we need large batch sizes in contrastive learning? a gradient-bias perspective. In *NeurIPS 2022*, 2022.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

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. *Nature Machine Intelligence*, pp. 1–8, 2022.

Guilloux, V. L., Schmidtke, P., and Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):1–11, 2009.

Han, H., Zheng, Q., Luo, M., Miao, K., Tian, F., and Chen, Y. Noise-tolerant learning for audio-visual action recognition. *arXiv preprint arXiv:2205.07611*, 2022.

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

Hoffmann, D. T., Behrmann, N., Gall, J., Brox, T., and Noroozi, M. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 897–905, 2022.

Hu, P., Zhu, H., Lin, J., Peng, D., Zhao, Y.-P., and Peng, X. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889, 2022.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Huang, R., Long, Y., Han, J., Xu, H., Liang, X., Xu, C., and Liang, X. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 926–934, 2023.

Koes, D. R., Baumgartner, M. P., and Camacho, C. J. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

Koukos, P. I., Xue, L. C., and Bonvin, A. M. Protein–ligand pose and affinity prediction: Lessons from d3r grand challenge 3. *Journal of Computer-Aided Molecular Design*, 33(1):83–91, 2019.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., Gao, P., Xie, G., and Song, S. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in Bioinformatics*, 22 (6):bbab109, 2021.

Li, Y., Yang, M., Peng, D., Li, T., Huang, J., and Peng, X. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022.

Lin, D. An introduction to expectation-maximization. 2011.

Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., and Peng, X. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

Liu, H., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. Pdb-wide collection of binding data: Current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

Méndez-Lucio, O., Ahmad, M., del Rio-Chanona, E. A., and Wegner, J. K. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.

Nielsen, S. The stochastic em algorithm: Estimation and asymptotic results. *Bernoulli*, 6:457–489, 2000.

Poklukar, P., Vasco, M., Yin, H., Melo, F. S., Paiva, A., and Kragic, D. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pp. 17782–17800, 2022.

Quiroga, R. and Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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 *Advances in Neural Information Processing Systems*, volume 33, 2020.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) equivariant graph neural networks. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R., and Jaakkola, T. Equibind: Geometric deep learning for drug binding structure prediction. 2022.

Tang, H. and Liu, Y. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, pp. 21090–21110, 2022.

Trott, O. and Olson, A. J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Wang, C. and Zhang, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3):169–177, 2017.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 429–436, 2019.

Wang, Y., Wang, J., Cao, Z., and Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, pp. 1–9, 2022.

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. *Chemical Science*, 9(2):513–530, 2018.

Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., and He, L. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022.

Yang, Y., Zhang, J., Gao, F., Gao, X., and Zhu, H. Domfn: A divergence-oriented multi-modal fusion network for resume assessment. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1612–1620, 2022.

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 *Advances in Neural Information Processing Systems*, volume 34, 2021.

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

Zaheer, M., Wick, M., Tristan, J.-B., Smola, A., and Steele, G. Exponential stochastic cellular automata for massively parallel inference. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 966–975, Cadiz, Spain, May 2016. PMLR.

Zhang, C., Han, Z., Fu, H., Zhou, J. T., Hu, Q., et al. Cpm-nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 559–569, 2019.

Zhang, S. and Chen, Y. Computation for latent variable model estimation: A unified stochastic proximal framework. *Psychometrika*, 87:1473–1502, 2020.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: a universal 3d molecular representation learning framework. 2023.

A. Similarity Score Distribution

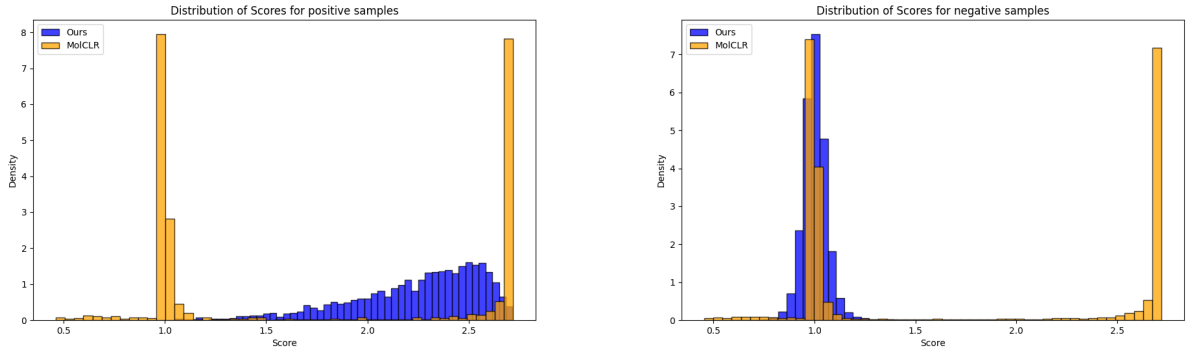

Figure 2. **Similarity Scores** – Similarity scores distribution for negative pairs in joint space after pre-training with original MolCLR loss and our proposed loss is provided. Compared to Using pretrained MolCLR model, our method yields similarity scores with lower mean and lower variance for negative pairs. While MolCLR have two peaks of negatives similarity scores around 1 and 2.7, our method concentrates them at only one peak of 1.Our method yields similarity scores with higher mean and lower variance for positive pairs. Our method concentrates at higher levels as it allows for some degree of semantic dissimilar between positives. The similarity scores are dot similarity, they are not normalized to enhance the difference for visual purposes.