UniCorn: A Unified Contrastive Learning Approach for Multi-view Molecular Representation Learning

Shikun Feng [*1] **Yuyan Ni** [*2 3] **Minghao Li** [4] **Yanwen Huang** [5] **Zhi-Ming Ma** [2] **Wei-Ying Ma** [1] **Yanyan Lan** [1 6]

Abstract

Recently, a noticeable trend has emerged in developing pre-trained foundation models in the domains of CV and NLP. However, for molecular pre-training, there lacks a universal model capable of effectively applying to various categories of molecular tasks, since existing prevalent pre-training methods exhibit effectiveness for specific types of downstream tasks. Furthermore, the lack of profound understanding of existing pre-training methods, including 2D graph masking, 2D-3D contrastive learning, and 3D denoising, hampers the advancement of molecular foundation models. In this work, we provide a unified comprehension of existing pre-training methods through the lens of contrastive learning. Thus their distinctions lie in clustering different views of molecules, which is shown beneficial to specific downstream tasks. To achieve a complete and general-purpose molecular representation, we propose a novel pre-training framework, named UniCorn, that inherits the merits of the three methods, depicting molecular views in three different levels. SOTA performance across quantum, physicochemical, and biological tasks, along with comprehensive ablation study, validate the universality and effectiveness of UniCorn.

1. Introduction

Molecular representation learning is pivotal across diverse drug discovery tasks. One important application is molecular property prediction which is promising to enable high-throughput screening of molecules with desirable properties. Following the pre-training methods in natural language processing (NLP) and computer vision (CV), a range of molecular pre-training methods has emerged to address the challenge of limited labeled molecular data. Specifically, they perform self-supervised learning (SSL) on a large number of unlabeled data, and then fine-tune on specific kinds of property prediction tasks with labeled data. Existing SSL methods can be mainly classified into three categories. The first category involves 2D graph masking (Hu et al., 2020; Hou et al., 2022; Rong et al., 2020; Xia et al., 2022), where random parts of the molecular graph are masked and the model is pre-trained to reconstruct them. The second category comprises 2D-3D contrastive learning (Stärk et al., 2022; Liu et al., 2021; 2023a; Li et al., 2022) that aligns the representations of 3D conformations and that of their corresponding 2D graph. The third category is 3D denoising (Zaidi et al., 2022; Feng et al., 2023a; Luo et al., 2023; Liu et al., 2022; Ni et al., 2024) that adds noise to the conformation and trains the model to predict the noise.

Recently, foundation models, such as ChatGPT (OpenAI, 2022), GPT4 (OpenAI, 2023), SAM (Kirillov et al., 2023), Unified-IO (Lu et al., 2022) , have caused a revolutionary shift to the field of Artificial Intelligence, which are pre-trained on large-scale data and are adopted to a broad range of downstream tasks (Li et al., 2023a). However, there still lacks a unified model in the molecular domain, that can learn a universal representation and effectively be applied to various property prediction tasks. The challenge firstly lies in that the relationship between the existing SSL methods is still under study. Moreover, existing methods are unbalanced for various downstream tasks. Generally, 3D denoising methods favor quantum chemical property prediction, while 2D graph masking and 2D-3D contrastive learning prefer biological and physicochemical property prediction. This phenomenon, also manifested in section 5.2, is hardly discussed by previous studies and the causes are still unclear. Furthermore, the relationship between the existing SSL methods is still under study, posing a challenge in leveraging their strengths to create a universal model effective for all three types of property prediction tasks.

[*]Equal contribution [1]Institute for AI Industry Research (AIR), Tsinghua University [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences [3]University of Chinese Academy of Sciences [4]Beijing Institute of Genomics, Chinese Academy of Sciences [5]Department of Pharmaceutical Science, Peking University [6]Beijing Academy of Artificial Intelligence. Correspondence to: Yanyan Lan <lanyanyan@air.tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

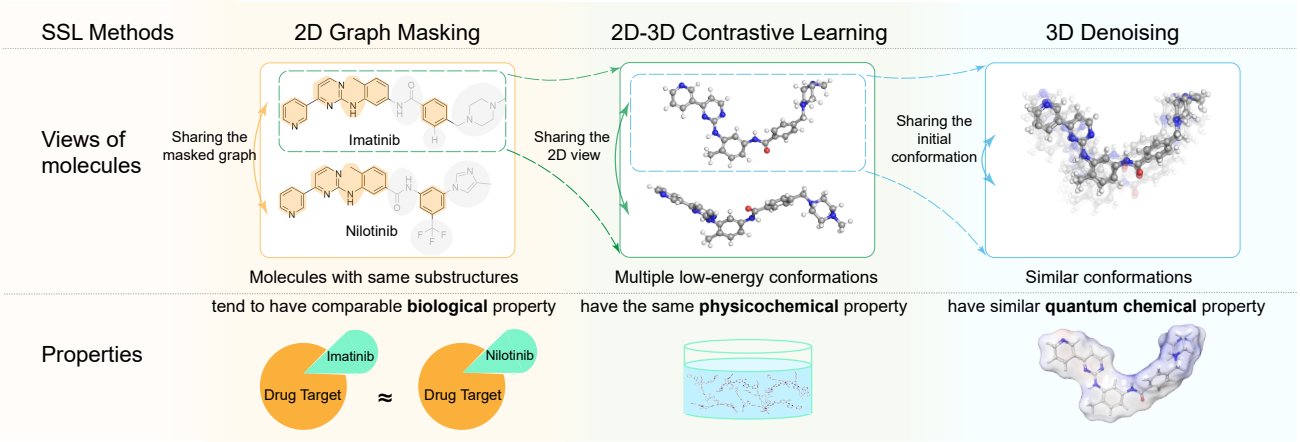

Figure 1. Correspondence between self-supervised learning (SSL) methods, views of molecules, and molecular properties. Different SSL methods cluster the molecular representations based on different levels of similarity (section 2). These clustering patterns align with the characteristics of properties at different scales (section 3).

To tackle these challenges, we provide a unified understanding of the three SSL methods through contrastive learning in section 2. In particular, 2D graph masking, 2D-3D contrastive learning, and 3D denoising can be comprehended as contrastive learning with masking, sampling multiple conformations of the same molecule, and noise addition as augmented views, respectively. In the derivation, we first of all summarize masking and denoising as reconstructive methods. Then we find the contrastive loss and the reconstructive loss are mutually upper and lower bounded by each other, indicating minimizing one loss can guarantee the other loss to be small under certain conditions and regularization. Lastly, we interpret the representation of contrastive learning as clustering different views. Therefore, the three SSL methods lead to **clustering patterns in molecular representation space at different granularity**. Importantly, clustering patterns at different granularity can exist concurrently, suggesting that the three pre-training methods are compatible, resulting in a multi-grained representation.

Moreover, we reveal that the multi-grained clustering patterns correspond to the inductive biases of different kinds of molecular properties, as shown in Figure 1. Firstly, 3D denoising aligns the representations of closely resembled conformations, in accordance with the characteristics of quantum chemical properties. Secondly, 2D-3D contrastive learning learns the invariance of multiple low-energy conformations of one molecule, corresponding to the characteristics of physicochemical properties. Thirdly, 2D graph masking clusters molecules with the same fragments, capturing characteristics of biological properties. Elaborate elucidation is provided in section 3. **This correspondence highlights the necessity to combine the three pre-training methods in order to achieve a universal representation.**

Consequently, we propose UniCorn, a **uni**fied molecular pre-training framework via **co**ntrastive lea**rn**ing, to learn multi-view molecular representations by amalgamating the strengths of existing methods into a unified pre-training framework, naturally capable of tackling quantum chemical, physicochemical, and biological properties. UniCorn takes both 2D molecular graphs and 3D molecular conformations as input, with tailored self-supervised strategies for each data type. For 2D graphs, we utilize fragments, recognized as a kind of chemical semantic component, as masking units to mask the 2D graphs and subsequently recover them. Regarding the 3D conformations, we employ torsion augmented denoising, which first augments rotatable torsions of the molecule to sample multiple chemically plausible 3D conformations, then perturbs the coordinate of the augmented molecule and predicts associated coordinate noise. Finally, cross-modal contrastive learning is introduced to align the representations of multiple 3D conformations generated by torsion augmented denoising to their shared 2D representation and further distill knowledge from 2D to 3D.

Experimental results show that UniCorn is not confined to achieving merely comparable results on tasks where existing methods excel but also consistently surpasses them across all three types of tasks. This not only demonstrates the compatibility of each pre-training module in UniCorn but also highlights their complementarity in achieving a universal molecular representation. Ablation study and visualization further validate our comprehension of SSL methods and their correlation to downstream tasks. Our contributions are summarized as follows:

- To our best knowledge, we are the first to systematically summarize existing molecular pre-training methods and conduct a thorough analysis of their associations with various downstream molecular tasks.
- Theoretically, we reveal the connection between reconstructive and contrastive methods and comprehend

them in a unified perspective through contrastive learning and representation clustering.

- Practically, we present UniCorn, a unified pre-training framework designed to learn hierarchical molecular representations applicable to a wide array of downstream tasks.
- Through exhaustive experiments on physicochemical and biological tasks in MoleculeNet and quantum tasks including QM9, MD17, and MD22, UniCorn surpasses existing molecular pre-training methods, showcasing its superior performance.

2. Unifying Reconstructive and Contrastive Methods

2.1. Abstraction and Formulation

In general, denoising and masking can be concluded as the reconstructive method that aims to reconstruct the original input from the perturbed version containing noise or masked elements. The introduction of noise or masking provides an alternative view of the original input, effectively serving as a form of data augmentation.

Contrastive learning, on the other hand, aims to align representations of different views of the input. To this end, several different families of methods have been raised, as summarized in Balestriero et al. (2023), including the deep metric learning family that usually involves negative samples and a symmetric structure (Chen et al., 2020; Oord et al., 2018), the self-distillation family that adopts a predictor to map between representations (Grill et al., 2020; Chen & He, 2021; Caron et al., 2021), and the canonical correlation analysis family that additionally regularize the covariance of the representations (Bardes et al., 2022). To show the clear correspondence between contrastive and reconstructive methods, we discuss the SimSiam like contrastive method (Chen & He, 2021) in the self-distillation family. The unification and equivalence between the three contrastive families are discussed in Tao et al. (2022); Garrido et al. (2023).

To formalize the learning objectives, we introduce $\mathbb{X} \subseteq \mathbb{R}^{d_\mathbb{X}}$, $\tilde{\mathbb{X}} \subseteq \mathbb{R}^{d_\mathbb{X}}$ to denote the sets comprising raw input data and the augmented data, respectively. We use $p(x)$ and $p(\tilde{x}|x)$ to denote the distribution of input data $x \in \mathbb{X}$ and the probability of generating an augmentation $\tilde{x} \in \tilde{\mathbb{X}}$ from x. The representations are situated within a representation space in $\mathbb{R}^{d_\mathbb{Z}}$. To facilitate a unified perspective, we define an encoder function $f_\theta : \mathbb{X} \to \mathbb{R}^{d_\mathbb{Z}}$, an aligner function $h_\psi : \mathbb{R}^{d_\mathbb{Z}} \to \mathbb{R}^{d_\mathbb{Z}}$ and a continuously differentiable decoder function $g_\phi : \mathbb{R}^{d_\mathbb{Z}} \to \mathbb{X}$, for extracting primary features from the input, aligning representations, and reconstructing the input from the representations, respectively. Here, θ, ψ and ϕ denote learnable parameters associated with these

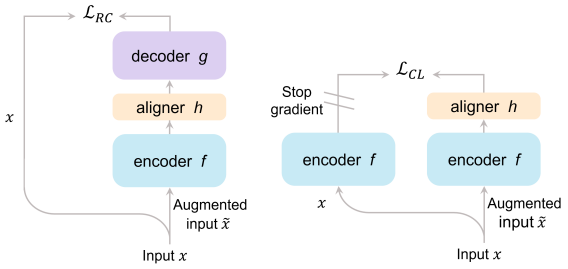

Figure 2. A unified perspective of reconstructive and contrastive methods.

functions. As a result, the reconstructive loss and contrastive loss can be expressed as:

$$\mathcal{L}_{\text{RC}} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||g_\phi(h_\psi(f_\theta(\tilde{x}))) - x||_2, \quad (1)$$

$$\mathcal{L}_{\text{CL}} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f_\theta(x))||_2, \quad (2)$$

where $||\cdot||_2$ represents the L^2 norm of vectors. While the squared L^2 norm is conventionally employed, we opt for the L^2 norm for the sake of brevity and clarity in the proof. This choice does not impact the discussion on loss functions, as the gradient vectors of these functions align in the same direction. The notation $\mathcal{SG}$ signifies stop gradient. An illustration of contrastive and reconstructive methods in a unified perspective is shown in Figure 2.

2.2. Mutual Upper and Lower Bounds

In this section, we demonstrate that, under certain conditions and regularization techniques, reconstructive and contrastive losses can serve as mutual upper and lower bounds for each other. This implies that minimizing one loss inherently results in the optimization of the other.

Theorem 2.1 (Relations between reconstructive and contrastive loss)**.** *We introduce an additional loss aimed at regularizing the decoder to approximate the inverse of the encoder.*

$$\mathcal{L}_{reg} = \mathbb{E}_{p(x)}||g_\phi(\mathcal{SG}(f_\theta(x))) - x||_2 \quad (3)$$

When λ_{max} and λ_{min} are non-zero, we derive the following conclusions:

① $$\mathcal{L}_{CL} + \frac{1}{\lambda_{max}}\mathcal{L}_{reg} \geq \frac{1}{\lambda_{max}}\mathcal{L}_{RC} \quad (4)$$

indicating that updating the encoder and aligner via contrastive learning and updating the decoder by the regularization loss, guarantees a small reconstructive loss.

② $$\mathcal{L}_{RC} + \mathcal{L}_{reg} \geq \lambda_{min}\mathcal{L}_{CL}, \quad (5)$$

indicating that updating the entire network through reconstructive learning and selectively partitioning the decoder by regularization loss guarantees a small contrastive loss. Kindly note that the partition of the encoder and decoder from the entire network does not affect $\mathcal{L}_{RC}$, however it is crucial for contrastive learning. Therefore, the regularization term can be optimized by finding the optimal partition

of the encoder and decoder from the entire network. And the contrastive loss is defined by the partitioned encoder and aligner.

Here, λ_{max} and λ_{min} are constants defined in the proof. The proof and discussions regarding the condition are elaborated in appendix B.

2.3. Clustered Representations

Previous studies (Huang et al., 2023; Assran et al., 2023) have revealed the connection between clustering and symmetric contrastive learning methods, such as SimCLR (Chen et al., 2020) and VICReg (Bardes et al., 2022). However, the contrastive loss defined in equation 2 exhibits asymmetry: the two representations to be aligned comprise one as the direct output of the encoder while the other passes through an aligner. In essence, equation 2 resembles the SimSiam contrastive loss up to the square: $\mathcal{L}_{\text{CL}}^{(\text{SimSiam})} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f_\theta(x))||_2^2$, where the representations are normalized $||f_\theta(\cdot)||_2^2 = ||h_\psi(f_\theta(\cdot))||_2^2 = 1$. An analogy to k-means clustering of the SimSiam contrastive loss is discussed in Chen & He (2021), where they hypothesized $h_\psi(f_\theta(\tilde{x})) \approx \mathbb{E}_{p(\tilde{x}|x)}f_\theta(\tilde{x}) \approx f_\theta(x)$. We provide a more explicit relationship between clustering and the SimSiam loss without incorporating an additional hypothesis.

Theorem 2.2. *Minimizing the contrastive loss guarantees clustering of the augmentation representations corresponding to the same raw input data.*

$$\mathcal{L}_{CL}^{(SimSiam)} \geq \mathcal{L}_{cluster}. \tag{6}$$

$\mathcal{L}_{cluster} \triangleq \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathbb{E}_{p(\tilde{x}|x)}h_\psi(f_\theta(\tilde{x}))||_2^2$ *describes the mean distance between the samples in the cluster and the cluster center.*

The theorem indicates minimizing the SimSiam contrastive loss guarantees a small clustering loss. The proof is provided in section B.3. Consequently, we can comprehend the reconstructive and contrastive methods from a unified perspective, as they both cluster different views of the input.

3. Delving into the Task-specific Preferences of Existing Methods

The previously introduced theorems have provided a unified comprehension of all three molecular SSL methods via contrastive learning and clustering of the representations. In molecular pre-training, the pivotal distinction among these methods lies in data augmentation approaches, which emphasize different views of molecules. Overall, we find that the clustered representations of each SSL method tend to align with a specific category of downstream tasks, which is commonly regarded as desirable for better generalization (Tian et al., 2020; Wang et al., 2022b; HaoChen et al., 2021; Huang et al., 2023), as well as efficient few-shot transfer learning (Galanti et al., 2022a;b).

3.1. 2D Graph Masking

In our comprehension of 2D graph masking, the data augmentation is randomly masking atoms, edges, or fragments in the molecular graph (Hu et al., 2020; Hou et al., 2022; Rong et al., 2020; Feng et al., 2023c). When similar molecules share a majority of the same substructures, their augmentation can overlap by masking non-identical regions and retaining shared substructures. As a result, these similar molecules are connected in the augmentation graph and can be clustered during pre-training (Wang et al., 2022b).

This clustering pattern is consistent with a fundamental principle in medicinal chemistry: molecules with similar substructures often exhibit comparable biological activities (Johnson et al., 1990; Dean, 1995; Willett et al., 1998). An example in point is the FDA-approved breakthrough drugs, imatinib (Druker et al., 2001) and nilotinib (Breccia & Alimena, 2010), shown in Figure 1. Sharing identical fragments, these two drugs are both orally available, potent, small-molecule inhibitors targeting breakpoint cluster region-Abelson (BCR-ABL) with some comparable clinical outcomes, including the overall frequency of adverse events, ten-year progression-free survival rates, and ten-year overall survival (Kantarjian et al., 2021). Their shared substructures contribute greatly to their similar biological properties both in vitro and in vivo (Manley et al., 2005).

3.2. 2D-3D Contrastive Learning

Recent methods advocate the alignment of 2D and 3D representations of the same molecule during pre-training to effectively leverage both 2D and 3D information (Stärk et al., 2022; Liu et al., 2021; 2023a; Li et al., 2022), Among them using multiple conformations of the same molecule tends to achieve higher performance (Stärk et al., 2022; Liu et al., 2021). According to a corollary of Theorem 2.2 in appendix B.4, the cross-modal contrastive learning clusters diverse conformations of the same molecule in the 3D representation space.

This clustering pattern also aligns with the current understanding in the field of chemistry that certain physicochemical properties of the system are resilient to conformational changes of a single molecule within. There is an overall statistical distribution describing the system with molecules interconverting between different conformations (Allen et al., 1996; Brameld et al., 2008). As illustrated in Figure 1, many physicochemical properties reflect the features of such distribution, including water solubility and octanol/water distribution coefficient, i.e., ESOL and Lipophilicity tasks in MoleculeNet dataset (Wu et al., 2018) respectively, captur-

ing the averaged behavior of molecular conformations rather than individual molecular conformation (Le Questel, 2023).

3.3. 3D Denoising

In our comprehension of 3D denoising, the data augmentation is adding Gaussian coordinate noise to the molecular conformation. As a result, 3D denoising clusters the representations of the noisy conformations around their equilibrium input. As demonstrated in Figure 1, this clustering pattern benefits quantum chemical properties as the small scale of perturbation ensures that the augmented conformations share similar properties to the raw conformation. This is because the quantum chemical properties are generally influenced by the spatial distribution of charges in molecules (Kauzmann, 2013) , and thus undergo slight alterations when conformational changes are minor. For instance, the dipole moment in QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) dataset, a common quantum chemistry property indicating the separation of positive and negative charges in a molecule, is determined by the sum of products of charge magnitudes and their location vectors (LANDAU & LIFSHITZ, 1975). For similar conformations, meaning their atomic nuclei positions are largely unchanged, the spatial arrangement of the electron cloud exhibits negligible changes, resulting in similar dipole moments.

4. UniCorn

To learn a complete representation of molecules aimed at capturing hierarchical clustering patterns essential for covering molecular task preferences, we propose UniCorn which contains fragment masking, torsion augmented denoising, and cross-modal distillation modules that capture different scales of molecular information, as illustrated in Figure 3.

4.1. Preliminary Notations

A molecule can be either represented as a 2D molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ or 3D conformation $\mathcal{C} = (\mathcal{V}, \mathcal{X})$, where vertexes $\mathcal{V}$ signify the atom types, edges $\mathcal{E}$ denote the chemical bonds, and $\mathcal{X}$ signify the Cartesian coordinates.

4.2. Fragment Masking Module

To cluster molecules with the same chemically plausible substructures, we conduct Masked Fragment Modeling (MFM) on the molecular 2D graph. A fragment is typically a small segment of a large molecule capable of existing independently while preserving meaningful chemical properties or biological functionalities (Erlanson et al., 2016; Murray & Rees, 2009; Erlanson et al., 2004). Unlike previous methods that mask atoms, edges, or tokens, We treat fragments as the basic unit for masking better fitting the characteristic of biological tasks discussed in section 3.1. Specifically, we utilize the BRICS (Degen et al., 2008) algorithm to break down the molecule into a set of fragments denoted as $\mathbf{s}$. We randomly mask a certain number of fragments, represented as $\mathbf{s}_m$, based on the mask ratio m. We then predict each atom type within the masked fragments. The masked fragment loss is defined by the following equation:

$$\mathcal{L}_m = -\mathbb{E}_{\mathcal{G},\mathcal{G}_{\setminus m}} \sum_{\mathbf{s}_m} \log \left[p_{(g_m, f^{2d})} \left(\mathbf{s}_m | \mathcal{G}_{\setminus m} \right) \right] \quad (7)$$

where f^{2d} denotes the 2D graph encoder, g_m represents the MLP head for MFM prediction, $p_{(g_m, f^{2d})}$ stands for the predicted probability of masked fragments, and $\mathcal{G}_{\setminus m}$ signifies the remaining part of the graph after masking.

4.3. Torsion Augmented Denoising Module

To cluster closely resembling conformations, we utilize the denoising pre-training strategy on 3D conformations. Unlike traditional denoising (Zaidi et al., 2022) whose unperturbed conformations are equilibriums $\mathcal{C} = (\mathcal{V}, \mathcal{X})$ that come from the dataset, we innovatively introduce a torsion augmentation step before denoising to provide diverse conformations of the same molecular graph for the cross-modal distillation module which is elaborated in the next section. Specifically, we perturb the torsion angles of the rotatable single bonds in equilibrium conformation to obtain multiple conformations for the molecule, denoted as $\mathcal{C}_a = (\mathcal{V}, \mathcal{X}_a)$. Subsequently, we adopt traditional denoising (Zaidi et al., 2022) on $\mathcal{C}_a$. We apply Gaussian noise to the coordinates of each atom and yield conformation $\tilde{\mathcal{C}} = (\mathcal{V}, \tilde{\mathcal{X}})$. The denoising loss is defined as follows:

$$\mathcal{L}_{dn} = \mathbb{E}_{\tilde{\mathcal{C}},\mathcal{C}_a} ||g_d(f^{3d}(\tilde{\mathcal{C}})) - (\tilde{\mathcal{X}} - \mathcal{X}_a)||_2^2. \quad (8)$$

Here, f^{3d} is the 3D encoder, and g_d represents the MLP head utilized for predicting the coordinate noise. Please note that the target of denoising is to predict the small-scale Gaussian coordinate noise, i.e. recovering $\mathcal{C}_a$ from $\tilde{\mathcal{C}}$, rather than recovering $\mathcal{C}$. Therefore, the addition of torsion augmentations does not alter the fact that 3D denoising clusters closely resembled conformations. On the other hand, the torsion augmentation not only efficiently provides diverse low-energy conformations for the cross-modal distillation module, but also enhances 3D denoising by enlarging sampling coverage and raising force accuracy, as validated in (Feng et al., 2023a). Although the combination of the torsion augmentation and traditional denoising results in the hybrid noise strategy in fractional denoising (Feng et al., 2023a), our derivation through augmentation is distinct from the previous interpretation of force learning.

4.4. Cross-modal Distillation Module

To achieve a hierarchical molecular representation, we employ cross-modal contrastive learning, distilling knowledge

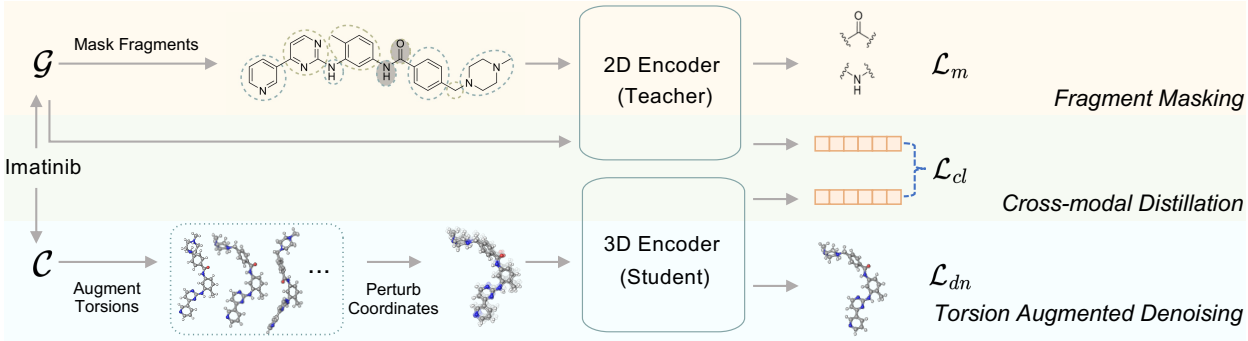

Figure 3. The UniCorn architecture we have ultimately reached, after exploring the association between each pre-training method and downstream tasks, with the goal of approaching unified molecular representations. The **Top** illustrates the Fragment Masking Module, wherein a 2D molecular graph is masked by fragments and subsequently recovered. The **Bottom** showcases the Torsion Augmented Denoising Module. This module operates in two steps: initially augmenting 3D conformers by perturbing rotatable torsions, and then introducing Gaussian coordinate noise for denoising. Finally, the **Middle** introduces the Cross-modal Distillation Module, responsible for distilling knowledge from 2D to 3D to achieve a hierarchical molecular representation.

from the 2D to 3D. Specifically, the encoded 3D conformation representation is denoted as $\boldsymbol{z}^c = h_p(f^{3d}(\tilde{\mathcal{C}}))$, where h_p refers to the alignment head used to align dimensions with 2D embeddings. $\tilde{\mathcal{C}}$ is the noisy conformation introduced in the torsion augmented denoising module. Simultaneously, the encoded 2D molecular graph representation is denoted as $\boldsymbol{z}^g = f^{2d}(\mathcal{G})$. The cross-modal contrastive loss is expressed as

$$\mathcal{L}_{cl} = -\mathbb{E}_{\mathcal{G},\tilde{\mathcal{C}}}[\log \frac{e^{\cos(\boldsymbol{z}^g,\boldsymbol{z}^c)/\tau}}{\sum_{\boldsymbol{z}_j^c} e^{\cos(\boldsymbol{z}^g,\boldsymbol{z}_j^c)/\tau}} + \log \frac{e^{\cos(\boldsymbol{z}^c,\boldsymbol{z}^g)/\tau}}{\sum_{\boldsymbol{z}_j^g} e^{\cos(\boldsymbol{z}^c,\boldsymbol{z}_j^g)/\tau}}], \quad (9)$$

where $\boldsymbol{z}_j^c \in \mathbb{N}^c \cup \boldsymbol{z}^c$ and $\boldsymbol{z}_j^g \in \mathbb{N}^g \cup \boldsymbol{z}^g$, $\mathbb{N}^c$ and $\mathbb{N}^g$ denotes in-batch negative samples of 2D graphs and 3D conformations, $cos(,)$ represents the cosine similarity function, and τ is the temperature hyperparameter. In different epochs, the same input $(\mathcal{G}, \mathcal{C})$ produces diverse torsion-augmented conformations $\mathcal{C}_a$ that are low-energy conformations largely different from each other. Therefore, the contrastive learning clusters dissimilar conformations of the same molecule.

It's important to highlight that conventional approaches primarily leverage 3D data to assist 2D networks in addressing 2D downstream tasks (Stärk et al., 2022; Liu et al., 2021). However, to achieve hierarchical clustering representation, we need to distill 2D knowledge into the 3D network, and not the other way around. This stems from the fact that the most fine-grained hierarchy, distinguishing different equilibriums, cannot be reflected in 2D inputs. This difference reflects the uniqueness of our motivation and methodology.

The ultimate loss, denoted as $\mathcal{L}$, is a sum of the three previously defined losses: $\mathcal{L} = \mathcal{L}_{dn} + \mathcal{L}_m + \mathcal{L}_{cl}$. As a result, the learned 3D representation contains knowledge of three distinct hierarchical views of molecules, promising to benefit molecular properties at different scales. Therefore after pre-training, we fine-tune the encoder f^{3d} to adapt it for diverse downstream tasks. Pseudocodes and hyperparameter settings can be found in appendix E and C.3.

5. Experiments

5.1. Settings

We use the pre-trained 3D encoder as the backbone for downstream tasks, but employing different heads for different tasks, we employ a simple 2-layer MLP head for MoleculeNet tasks, and use an equivariant head as defined in (Thölke & De Fabritiis, 2022) for quantum tasks including QM9, MD17 and MD22.

Our hyperparameter selection methods are in line with previous methods. We manually select pre-training and fine-tuning hyperparameters for QM9, MD17, and MD22 to ensure a consistent decrease in pre-training loss and optimal performance on the validation set. For tasks within MoleculeNet, we utilize grid search to identify the most suitable hyperparameters. We provide the selected results and search space in Appendix C.3.

5.1.1. Datasets

Our pre-training dataset consists of 15 million molecules sourced from (Nakata & Shimazaki, 2017; Zhou et al., 2023). We systematically assess the effectiveness of our model across two different task categories: quantum mechanical tasks, including QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012), MD17 (Chmiela et al., 2017) as well as MD22 (Chmiela et al., 2023), and physicochemical and biological property prediction tasks, focusing on the MoleculeNet dataset (Wu et al., 2018). Details about downstream datasets can be found in appendix C.1. Since we utilize

the 3D encoder to fine-tune for downstream tasks, we rely on 3D conformations generated by RDKit for MoleculeNet data, following previous methods (Yu et al., 2023; Zhou et al., 2023; Fang et al., 2022).

5.1.2. Baselines

Our baselines cover three typical categories within molecular pre-training. The first category comprises denoising methods, including Transformer-M (Luo et al., 2023), SE(3)-DDM (Liu et al., 2022), 3D-EMGP (Jiao et al., 2023), Coord (Zaidi et al., 2022), Frad (Feng et al., 2023a), and Coord(NonEq) (Wang et al., 2023). The second category consists of multimodal pre-training methods that aim to fusion molecular 2D graphs and 3D conformations, such as 3D InfoMax (Stärk et al., 2022), GraphMVP (Liu et al., 2021), MoleculeSDE (Liu et al., 2023a), MoleBLEND (Yu et al., 2023), and MoleculeJAE (Du et al., 2023). Notably, 3D InfoMax, GraphMVP, and MoleculeSDE employ contrastive learning between 2D graphs and 3D conformations as a pre-training strategy. The final category encompasses masking methods, which include AttrMask (Hu et al., 2020), GraphMAE (Hou et al., 2022), GROVER (Rong et al., 2020) and Mole-BERT (Xia et al., 2022). Masking methods typically take 2D graphs as input and may not be applicable to quantum downstream tasks that require precise conformation.

5.2. Main Experimental Results

5.2.1. Quantum Tasks

Table 1 presents the performance of 12 regression tasks in QM9. Upon comparing various pre-training methods, an evident observation is that the 3D denoising approach significantly outperforms multimodal pre-training methods on average. This observation confirms our previous conjecture that the 3D denoising task is more beneficial to quantum tasks. Further supporting this finding is that denoising holds a physical interpretation, equivalent to learning an approximate force field for molecules (Zaidi et al., 2022). Notably, our methods outperform existing denoising baselines, achieving the best performance in 10 out of 12 tasks. This achievement can be attributed to the complementary nature of chemical bond information in 2D graph to 3D conformation, enhancing the performance of quantum tasks. This insight is further validated by (Luo et al., 2023; Yu et al., 2023).

Tables 3 presents the performance results for force prediction tasks on MD17. UniCorn outperforms existing denoising and multimodal methods, establishing a new state-of-the-art performance on six out of eight molecules. Additionally, we conduct experiments with UniCorn on more challenging force prediction tasks in MD22, as detailed in sections A.1, to demonstrate its universality.

5.2.2. Biological and Physicochemical Tasks

Table 2 illustrates the performance of our method across 8 biological classification tasks in MoleculeNet. Impressively, our approach attains state-of-the-art results in 7 out of the 8 tasks. On average, our method outperforms the second-best method, MoleBLEND, by a substantial margin (78.4 vs 76.2). Furthermore, Table 4 demonstrates that our method achieves the best performance across all 3 physicochemical regression tasks in MoleculeNet.

The improvement observed in our method, as compared to general masking and multimodal methods, can be attributed to the intrinsic connection between macroscopic properties and microscopic quantum properties, which has been thoroughly elucidated by (Beaini et al., 2023; Sun et al., 2022). The quantum properties of molecules, describing internal electronic motion and atomic nucleus vibrations at the microscopic level, influence the interaction behavior of the molecules with others. These interactions, in turn, affect the physicochemical and biological properties of the molecules. Therefore, the incorporation of denoising task proves instrumental in not only enhancing quantum task performance but also contributing to the improvement in macroscopic properties.

In summary, UniCorn surpasses previous state-of-the-art methods, delivering optimal results across 33 out of 38 molecular tasks that span a wide range of quantum, physicochemical, and biological domains. This superiority not only emphasizes the compatibility of each pre-training strategy but also highlights their complementary nature, thereby contributing to the learning of a universal molecular representation.

5.3. Ablation Study

5.3.1. Loss Study

To further validate our proposition that distinct pre-training losses exhibit preferences for specific downstream tasks, we conduct two sets of experiments. Firstly, we perform pre-training with and without denoising, followed by fine-tuning on three QM9 tasks. The results are presented in Table 5, revealing a significant decline in performance without the denoising loss. This underscores the vital role of denoising for quantum chemical tasks. Secondly, we pre-train the model without the masked fragment loss and the cross-modal distillation loss, followed by fine-tuning on 4 biological and 2 physicochemical tasks in MoleculeNet. The results, shown in Table 6, demonstrate that incorporating fragment masking and cross-modal distillation can boost performance on biological and physicochemical tasks.

Table 1. Performance (MAE, ↓) on QM9 quantum tasks. The best results are in bold.

Methods	Models	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$< R^2 >$ (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{molK}$)
Multimodal	3D InfoMax	0.0280	0.057	25.9	21.6	42.1	0.141	1.67	13.30	13.81	13.62	13.73	0.030
	GraphMVP	0.0270	0.056	25.8	21.6	42.0	0.136	1.61	13.07	13.03	13.31	13.43	0.029
	MoleculeSDE	0.0260	0.054	25.7	21.4	41.8	0.151	1.59	12.04	12.54	12.05	13.07	0.028
	MoleculeJAE	0.0270	0.056	26.0	21.6	42.7	0.141	1.56	10.70	10.81	10.70	11.22	0.029
	MoleBLEND	0.0370	0.060	21.5	19.2	34.8	0.417	1.58	11.82	12.02	11.97	12.44	0.031
3D Denoising	Transformer-M	0.0370	0.041	17.5	16.2	27.4	0.075	**1.18**	9.37	9.41	9.39	9.63	0.022
	SE(3)-DDM	0.0150	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
	3D-EMGP	0.0200	0.057	21.3	18.2	37.1	**0.092**	1.38	8.60	8.60	8.70	9.30	0.026
	Frad	0.0100	0.037	15.3	13.7	27.8	0.342	1.42	5.33	5.62	5.55	6.19	0.020
	UniCorn	**0.0085**	**0.036**	**13.0**	**11.9**	**24.9**	0.326	1.40	**3.99**	**3.95**	**3.94**	**5.09**	**0.019**

Table 2. Performance (ROC-AUC %, ↑) on MoleculeNet biological classification tasks. The best results are in bold.

Methods	Models	BBBP	Tox21	MUV	BACE	ToxCast	SIDER	ClinTox	HIV	Avg.
Graph Masking	AttrMask	65.0±2.3	74.8±0.2	73.4±2.0	79.7±0.3	62.9±0.1	61.2±0.1	87.7±1.1	76.8±0.5	72.7
	GROVER	70.0±0.1	74.3±0.1	67.3±1.8	82.6±0.7	65.4±0.4	64.8±0.6	81.2±3.0	62.5±0.9	71.0
	GraphMAE	72.0±0.6	75.5±0.6	76.3±2.4	83.1±0.9	64.1±0.3	60.3±1.1	82.3±1.2	77.2±1.0	73.9
	Mole-BERT	71.9±1.6	76.8±0.5	78.6±1.8	80.8±1.4	64.3±0.2	62.8±1.1	78.9±3.0	78.2±0.8	74.0
Multimodal	3D InfoMax	69.1±1.0	74.5±0.7	74.4±2.4	79.7±1.5	64.4±0.8	60.6±0.7	79.9±3.4	76.1±1.3	72.3
	GraphMVP	68.5±0.2	74.5±0.4	75.0±1.4	76.8±1.1	62.7±0.1	62.3±1.6	79.0±2.5	74.8±1.4	71.7
	MoleculeSDE	71.8±0.7	76.8±0.3	80.9±0.3	79.5±2.1	65.0±0.2	60.8±0.3	87.0±0.5	78.8±0.9	75.1
	MoleBLEND	73.0±0.8	77.8±0.8	77.2±2.3	83.7±1.4	66.1±0.0	**64.9±0.3**	87.6±0.7	79.0±0.8	76.2
	UniCorn	**74.2±1.1**	**79.3±0.5**	**82.6±1.0**	**85.8±1.2**	**69.4±1.1**	64.0±1.8	**92.1±0.4**	**79.8±0.9**	**78.4**

Table 3. Performance (MAE, ↓) on MD17 force prediction tasks (kcal/mol/ Å). The best results are in bold. *: SE(3)-DDM employs the Benzene dataset from Chmiela et al. (2018), which differs from the version utilized in our work (Chmiela et al., 2017).

Models	Aspirin	Benzene	Ethanol	Malonal -dehyde	Naphtha -lene	Salicy -lic Acid	Toluene	Uracil
MoleculeJAE	1.289	0.345	0.365	0.613	0.498	0.712	0.480	0.463
MoleculeSDE	1.112	0.304	0.282	0.520	0.455	0.725	0.515	0.447
SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
Coord	0.211	0.169	0.096	**0.139**	0.053	0.109	0.058	**0.074**
Frad	0.209	0.199	0.091	0.142	0.053	0.108	0.054	0.076
UniCorn	**0.168**	**0.165**	**0.086**	0.152	**0.046**	**0.098**	**0.052**	0.084

Table 4. Performance (RMSE, ↓) on MoleculeNet physicochemical regression tasks. The best results are in bold.

Models	ESOL	FreeSolv	Lipo
AttrMask	1.112±0.048	-	0.730±0.004
GROVER	0.983±0.090	2.176±0.052	0.817±0.008
3D InfoMax	0.894±0.028	2.337±0.227	0.695±0.012
GraphMVP	1.029±0.033	-	0.681±0.010
MoleBLEND	0.831±0.026	1.910±0.163	0.638±0.004
UniCorn	**0.817±0.034**	**1.555±0.075**	**0.591±0.016**

5.3.2. Hierarchical Feature Visualization

We construct a hierarchical molecular dataset that includes trajectory-level labels representing adjacent similar conformations within the same molecular dynamic trajectory, as well as molecule-level labels denoting multiple trajectories share the same molecule, and scaffold-level labels indicating similar molecules have the same scaffold. A detailed description of the dataset construction process is provided in section C.2. Subsequently, we employ t-SNE (Van der Maaten & Hinton, 2008) for visualizing their UniCorn representations.

Table 5. Performance (MAE, ↓) on QM9. The top results are in bold.

QM9	ϵ_{LOMO} (meV)	ϵ_{HOMO} (meV)	$\Delta\epsilon$(meV)
Train from scratch	16.7	17.6	31.3
UniCorn w/o Denoising	14.7	16.8	31.0
UniCorn	**11.9**	**13.0**	**24.9**

Table 6. Performance (ROC-AUC %, ↑ and RMSE, ↓) on MoleculeNet. The top results are in bold. M&C denotes fragment masking and cross-modal distillation.

MoleculeNet	BBBP↑	BACE↑	Tox21↑	ToxCast↑	ESOL↓	Lipo↓
Train from scratch	68.9 (2.4)	83.6 (1.8)	76.1 (0.7)	65.1 (0.1)	1.083(0.030)	0.730(0.016)
UniCorn w/o M&C	69.3 (0.9)	81.4 (0.5)	77.8 (1.4)	65.8 (0.7)	0.825(0.036)	0.623(0.009)
UniCorn	**74.2 (1.1)**	**85.8 (1.2)**	**79.3 (0.5)**	**69.4 (1.1)**	**0.817(0.034)**	**0.591(0.016)**

The clustering results are illustrated in Figure 4. Moving from left to right, we present the trajectory-level, molecule-level, and scaffold-level clustering results. Given that these

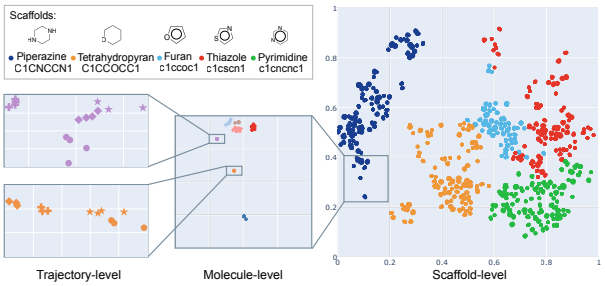

Figure 4. The visualization showcases hierarchical molecular representations learned by UniCorn, illustrating its ability to achieve effective clustering across different levels.

levels involve clustering at different resolutions, we have chosen specific areas for a closer examination in the figure. At the scaffold level, all five scaffolds are well-separated in general. Notably, heterocycles with similar properties have closer interclass distances, such as non-aromatic rings being situated on the left while aromatic rings are positioned on the right. Zooming into the molecular level, we find the conformations of different molecules are well-clustered. Moreover, when focusing on specific molecules' conformations, we find that similar conformations tend to have closely positioned representations. As a result, UniCorn's molecular representation demonstrates clustering at various levels, affirming the compatibility of the three pre-training modules and their successful realization of multi-view representations.

We also visualize the representations pre-trained by UniCorn and other SSL models across various downstream tasks in section A.2. It's important to note that these representations are not fine-tuned, directly reflecting the efficacy of pre-training methods. Once again, the results validate the task-specific preference of existing methods and highlight UniCorn's consistent competence across diverse downstream tasks.

6. Conclusion

To address the lack of effective unified models in molecular pre-training, we begin with a theoretical unification of three primary SSL methods: masking, denoising, and contrastive learning, from a contrastive perspective. This comprehension enables us to interpret the three categories of pre-training methods as clustering molecular views at different levels. Subsequently, we expound on how these different levels of views align with specific downstream tasks, thereby elucidating the task-specific preferences of existing methods. Finally, in pursuit of a universal representation beneficial for diverse downstream tasks, we introduce UniCorn, a novel approach capable of comprehending molecular views at multiple levels. UniCorn demonstrates superior performance compared to existing SSL methods across all three types of tasks. Ablation and visualization analyses further illustrate the compatibility, complementarity, and universality of the multi-view molecular representation learned by UniCorn.

Our work has opened avenues for further exploration in several directions. Firstly, beyond property prediction tasks, whether the UniCorn representation can enhance generation tasks is worth studying. Secondly, our comprehensive understanding of SSL methods offers a distinctive perspective for elucidating denoising and masking techniques, potentially inspiring advancements in noise addition and masking strategies. Thirdly, various SSL methods have emerged in other closely related domains (Yang et al., 2022; Wang et al., 2022a; Zhang et al., 2022b; Rives et al., 2021), yet their understanding and relevance to downstream tasks remain unclear. We aspire that our work can stimulate efforts to establish the connection between pre-training methods and downstream tasks, thereby fostering the development of new foundational models in these domains.

Acknowledgements

This work is supported by National Key R&D Program of China No.2021YFF1201600 and Beijing Academy of Artificial Intelligence (BAAI).

We thank anonymous reviewers for constructive and helpful discussions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen, F. H., Harris, S. E., and Taylor, R. Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. *Journal of computer-aided molecular design*, 10:247–254, 1996.

Assran, M., Balestriero, R., Duval, Q., Bordes, F., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., and Ballas, N. The hidden uniform cluster prior in self-supervised learning. 2023.

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. 2022.

Beaini, D., Huang, S., Cunha, J. A., Moisescu-Pareja, G., Dymov, O., Maddrell-Mander, S., McLean, C., Wenkel, F., Müller, L., Mohamud, J. H., et al. Towards foundational models for molecular learning on large-scale multi-task datasets. *arXiv preprint arXiv:2310.04292*, 2023.

Brameld, K. A., Kuhn, B., Reuter, D. C., and Stahl, M. Small molecule conformational preferences derived from crystal structure data. a medicinal chemistry focused analysis. *Journal of chemical information and modeling*, 48 (1):1–24, 2008.

Breccia, M. and Alimena, G. Nilotinib: a second-generation tyrosine kinase inhibitor for chronic myeloid leukemia. *Leukemia research*, 34(2):129–134, 2010.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Chmiela, S., Sauceda, H. E., Müller, K.-R., and Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9 (1):1–10, 2018.

Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., and Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023. doi: 10.1126/sciadv.adf0873. URL https://www.science.org/doi/abs/10.1126/sciadv.adf0873.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

Dean, P. M. *Defining molecular similarity and complementarity for drug design*, pp. 1–23. Springer Netherlands, Dordrecht, 1995. ISBN 978-94-011-1350-2. doi: 10.1007/978-94-011-1350-2_1. URL https://doi.org/10.1007/978-94-011-1350-2_1.

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(10):1503–1507, 2008.

Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., et al. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine*, 344(14): 1031–1037, 2001.

Du, W., Chen, J., Zhang, X., Ma, Z., and Liu, S. Molecule joint auto-encoding: Trajectory pretraining with 2d and 3d diffusion. *arXiv preprint arXiv:2312.03475*, 2023.

Eastman, P., Friedrichs, M. S., Chodera, J. D., Radmer, R. J., Bruns, C. M., Ku, J. P., Beauchamp, K. A., Lane, T. J., Wang, L.-P., Shukla, D., et al. Openmm 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of chemical theory and computation*, 9(1):461–469, 2013.

Erlanson, D. A., McDowell, R. S., and O'Brien, T. Fragment-based drug discovery. *Journal of medicinal chemistry*, 47(14):3463–3482, 2004.

Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., and Jhoti, H. Twenty years on: the impact of fragments on drug discovery. *Nature reviews Drug discovery*, 15(9): 605–619, 2016.

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*, 4(2):127–134, 2022.

Feng, S., Ni, Y., Lan, Y., Ma, Z.-M., and Ma, W.-Y. Fractional denoising for 3D molecular pre-training. In *International Conference on Machine Learning*, pp. 9938–9961. PMLR, 2023a.

Feng, S., Ni, Y., Lan, Y., Ma, Z.-M., and Ma, W.-Y. Fractional denoising for 3D molecular pre-training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9938–9961. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/feng23c.html.

Feng, S., Yang, L., Ma, W., and Lan, Y. Unimap: Universal smiles-graph representation learning. *arXiv preprint arXiv:2310.14216*, 2023c.

Galanti, T., György, A., and Hutter, M. On the role of neural collapse in transfer learning. 2022a.

Galanti, T., György, A., and Hutter, M. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022b.

Gao, Z., Tan, C., Wu, L., and Li, S. Z. Cosp: Co-supervised pretraining of pocket and ligand. *arXiv preprint arXiv:2206.12241*, 2022.

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and Lecun, Y. On the duality between contrastive and non-contrastive self-supervised learning. 2023.

Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Velivckovi'c, P., Kirkpatrick, J., and Battaglia, P. W. Simple gnn regularisation for 3D molecular property prediction and beyond. In *International Conference on Learning Representations*, 2021. URL https://api.semanticscholar.org/CorpusID:247450503.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 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 (ICLR)*, 2020.

Huang, W., Yi, M., Zhao, X., and Jiang, Z. Towards the generalization of contrastive self-supervised learning. 2023.

Jiao, R., Han, J., Huang, W., Rong, Y., and Liu, Y. Energy-motivated equivariant pretraining for 3d molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8096–8104, 2023.

Johnson, M. A., Maggiora, G. M., et al. Concepts and applications of molecular similarity. 1990. URL https://cir.nii.ac.jp/crid/1130000796218783104.

Kantarjian, H. M., Hughes, T. P., Larson, R. A., Kim, D.-W., Issaragrisil, S., le Coutre, P., Etienne, G., Boquimpani, C., Pasquini, R., Clark, R. E., et al. Long-term outcomes with frontline nilotinib versus imatinib in newly diagnosed chronic myeloid leukemia in chronic phase: Enestnd 10-year analysis. *Leukemia*, 35(2):440–453, 2021.

Kauzmann, W. *Quantum chemistry: an introduction*. Elsevier, 2013.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv:2304.02643*, 2023.

LANDAU, L. and LIFSHITZ, E. Chapter 5 - constant electromagnetic fields. In LANDAU, L. and LIFSHITZ, E. (eds.), *The Classical Theory of Fields (Fourth Edition)*, volume 2 of *Course of Theoretical Physics*, pp. 89–108. Pergamon, Amsterdam, fourth edition edition, 1975. ISBN 978-0-08-025072-4. doi: https://doi.org/10.1016/B978-0-08-025072-4.50012-5. URL https://www.sciencedirect.com/science/article/pii/B9780080250724500125.

Le Questel, J.-Y. Conformations and physicochemical properties of biological ligands in various environments, 2023.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., and Gao, J. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023a.

Li, J., Wu, R., Sun, W., Chen, L., Tian, S., Zhu, L., Meng, C., Zheng, Z., and Wang, W. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1268–1279, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599546. URL https://doi.org/10.1145/3580305.3599546.

Li, S., Zhou, J., Xu, T., Dou, D., and Xiong, H. Geomgcl: Geometric graph contrastive learning for molecular property prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 4541–4549, 2022.

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*, 2021.

Liu, S., Guo, H., and Tang, J. Molecular geometry pretraining with se (3)-invariant denoising distance matching. In *International Conference on Learning Representations*, 2022.

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

Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023b.

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Luo, S., Chen, T., Xu, Y., Zheng, S., Liu, T.-Y., Wang, L., and He, D. One transformer can understand both 2D & 3D molecular data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vZTp1oPV3PC.

Manley, P. W., Cowan-Jacob, S. W., and Mestan, J. Advances in the structural biology, design and clinical development of bcr-abl kinase inhibitors for the treatment of chronic myeloid leukaemia. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1754(1-2):3–13, 2005.

Murray, C. W. and Rees, D. C. The rise of fragment-based drug discovery. *Nature chemistry*, 1(3):187–192, 2009.

Nakata, M. and Shimazaki, T. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308, 2017.

Ni, Y., Feng, S., Ma, W.-Y., Ma, Z.-M., and Lan, Y. Sliced denoising: A physics-informed molecular pre-training method. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=liKkG1zcWq.

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

OpenAI. ChatGPT. 2022. URL https://openai.com/blog/chatgpt/.

OpenAI. GPT-4 technical report. 2023. URL https://arxiv.org/abs/2303.08774.

Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., and Yi, L. Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 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. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

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

Sun, Y., Chen, Y., Ma, W., Huang, W., Liu, K., Ma, Z., Ma, W.-Y., and Lan, Y. Pemp: Leveraging physics properties to enhance molecular property prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3505–3513, 2022.

Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., and Dai, J. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14431–14440, 2022.

Thölke, P. and De Fabritiis, G. Torchmd-net: equivariant transformers for neural network based molecular potentials. *arXiv preprint arXiv:2202.02541*, 2022.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, D., Fei, Y., and Zhou, H. On pre-training language model for antibody. In *The Eleventh International Conference on Learning Representations*, 2022a.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. Smilesbert: 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., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new understanding of contrastive learning. 2022b. URL https://openreview.net/forum?id=ECvgmYVyeUz.

Wang, Y., Xu, C., Li, Z., and Barati Farimani, A. Denoise pretraining on nonequilibrium molecules for accurate and transferable neural potentials. *Journal of Chemical Theory and Computation*, 19(15):5077–5087, 2023. doi: 10.1021/acs.jctc.3c00289. URL https://doi.org/10.1021/acs.jctc.3c00289. PMID: 37390120.

Willett, P., Barnard, J. M., and Downs, G. M. Chemical similarity searching. *Journal of chemical information and computer sciences*, 38(6):983–996, 1998.

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.

Xia, J., Zhao, C., Hu, B., Gao, Z., Tan, C., Liu, Y., Li, S., and Li, S. Z. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations*, 2022.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., and Yao, J. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 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? *Advances in Neural Information Processing Systems*, 34:28877–28888, 2021.

Yu, Q., Zhang, Y., Ni, Y., Feng, S., Lan, Y., Zhou, H., and Liu, J. Unified molecular modeling via modality blending. *arXiv preprint arXiv:2307.06235*, 2023.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2022.

Zhang, Q., Wang, Y., and Wang, Y. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 27127–27139, 2022a.

Zhang, X.-C., Wu, C.-K., Yang, Z.-J., Wu, Z.-X., Yi, J.-C., Hsieh, C.-Y., Hou, T.-J., and Cao, D.-S. Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in bioinformatics*, 22(6):bbab152, 2021.

Zhang, Z., Xu, M., Jamasb, A. R., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*, 2022b.

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. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Zorich, V. A. *The Differential Calculus of Functions of Several Variables*, pp. 427–543. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-662-48792-1. doi: 10.1007/978-3-662-48792-1_8. URL https://doi.org/10.1007/978-3-662-48792-1_8.

A. More Experimental Results

A.1. MD22

In Table 7, we present the performance results for force prediction tasks on MD22. UniCorn consistently outperforms the best results across all seven tasks, which include larger molecules with more complex structures, highlighting the broad application and generalization capabilities of our approach.

Table 7. Performance (MAE, ↓) on MD22 force prediction tasks(kcal/mol/ Å). The best results are in bold.

Models	Ac-Ala3-NHMe	DHA	AT-AT	Staychose	AT-AT -CG-CG	Buckyball	Double -walled
Coord(NonEq)	0.102	0.135	0.288	0.673	0.657	1.751	2.515
Frad	0.073	0.078	0.233	0.231	0.308	0.430	0.729
UniCorn	**0.056**	**0.051**	**0.134**	**0.127**	**0.181**	**0.249**	**0.473**

A.2. Downstream Tasks Feature Visualization

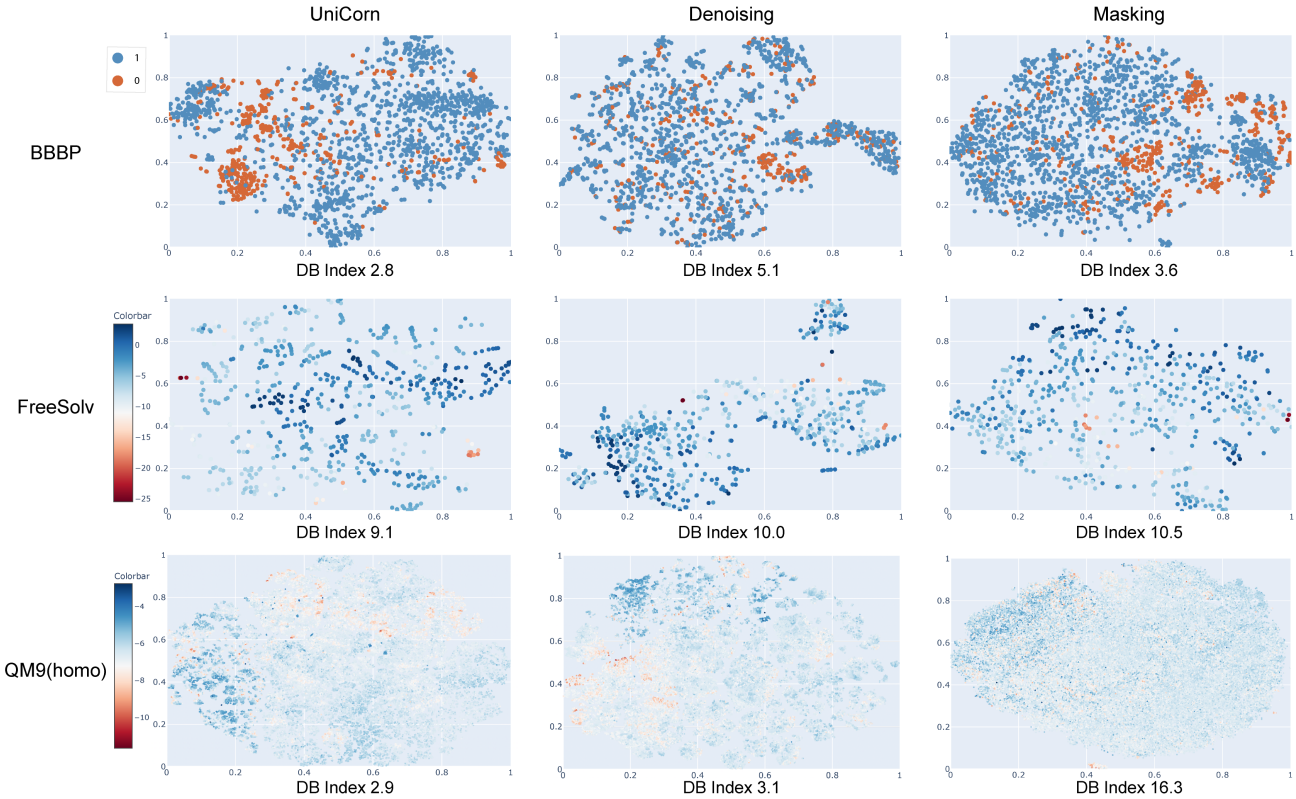

Figure 5. The clustering results of **unfine-tuned** molecular representations by three distinct methods, across three diverse tasks: BBBP (biological task), Freesolv (physicochemical task), and homo from QM9 (quantum task). The color indicates the labels of the downstream tasks—discrete binary labels for the BBBP task and continuous labels for Freesolv and QM9. Below each subfigure we present Davies–Bouldin Index to evaluate the performance of clustering results(smaller is better). While the Masking and Denoising methods exhibit a preference for biological and quantum tasks respectively, UniCorn demonstrates the capability to achieve significant clustering results across all types of tasks.

To illustrate the correlation between different pre-training methods and various types of downstream tasks analyzed in the previous section 3, we employed t-SNE to visualize the unfine-tuned features pre-trained by UniCorn, and other two representative methods: Coord (Zaidi et al., 2022) representing for denoising method, and AttrMask (Hu et al., 2020) representing for masking approach, on three typical downstream tasks: the biological task BBBP, the physical chemistry task Freesolv, and the quantum task homo from QM9. The color in the visualization denotes the ground-truth discrete or

continuous labels in downstream tasks. The clustering results are presented in Figure 5. In addition, we quantitatively calculate the Davies–Bouldin Index (Davies & Bouldin, 1979), a metric employed to assess clustering results, with lower values indicating better performance, and present these values below each corresponding subfigure.

From Figure 5, two key insights emerge. Firstly, the first row demonstrates that our methods yield meaningful clustering results across all types of tasks, where samples with the same or similar labels are clustered together. This indicates that our pre-training tasks are neither conflicting nor redundant but rather complementary. Secondly, examining the second and third rows, it is obvious that 3D denoising is more advantageous for quantum tasks compared to biological tasks, while the masking method exhibits the opposite trend: it is helpful for biological tasks but not as effective for quantum tasks. This observation serves as a solid validation of our earlier theoretical analyses concerning various pre-training tasks.

B. Missing Proofs and Analysis

B.1. Proof of Theorem 2.1

As introduced in section 2, we formalize the reconstruction loss and contrastive loss as follows:

Definition B.1 (Reconstruction loss)**.**

$$\mathcal{L}_{\text{RC}} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||g_\phi(h_\psi(f_\theta(\tilde{x}))) - x||_2 \tag{10}$$

Definition B.2 (Contrastive loss)**.**

$$\mathcal{L}_{\text{CL}} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f_\theta(x))||_2 \tag{11}$$

We also introduce two auxiliary losses to bridge the reconstructive and contrastive targets.

Definition B.3 (Regularization loss)**.**

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{p(x)}||g_\phi(\mathcal{SG}(f_\theta(x))) - x||_2 \tag{12}$$

Definition B.4 (Modified reconstruction loss)**.**

$$\mathcal{L}_{\text{RC2}} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||g_\phi(\mathcal{SG}(f_\theta(x))) - g_\phi(h_\psi(f_\theta(\tilde{x})))||_2 \tag{13}$$

The proof of Theorem 2.1 relies on the following two lemmas.

Lemma B.5. *The contrastive loss and the modified reconstruction loss are upper and lower bounded by each other when λ_{max} and λ_{min} are non-zero.*

$$\lambda_{max}\mathcal{L}_{CL} \geq \mathcal{L}_{RC2} \geq \lambda_{min}\mathcal{L}_{CL}, \tag{14}$$

where λ_{max} and λ_{min} are non-negative constants defined in the proof.

Proof. Denote vectors $a \triangleq f_\theta(x) \in \mathbb{R}^{d_\mathbb{Z}}$ and $b \triangleq h_\psi(f_\theta(\tilde{x})) \in \mathbb{R}^{d_\mathbb{Z}}$. Since g_ϕ is continuously differentiable, by the mean value formula of multivariant real-valued function (Zorich, 2015), $\forall j = 1, \cdots, d_\mathbb{X}$, $\exists \xi_j$ in the interval of a and b, s.t. $g_{\phi,j}(a) - g_{\phi,j}(b) = g'_{\phi,j}(\xi_j)^\top(a-b)$, where $g_{\phi,j}$ is the j th output dimenstion of g_ϕ and $g'_{\phi,j}(\xi_j)$ denotes the gradient with respect to ξ_j. Denote $G(x,\tilde{x}) = (g'_{\phi,1}(\xi_1), \cdots, g'_{\phi,d_\mathbb{X}}(\xi_{d_\mathbb{X}}))^\top \in \mathbb{R}^{d_\mathbb{X}\times d_\mathbb{Z}}$, $g_\phi(a) - g_\phi(b) = G(x,\tilde{x})(a-b)$. Then the modified reconstruction loss for the specific sample and its augmentation is

$$\begin{aligned} l_{\text{RC2}}(x,\tilde{x}) &\triangleq ||g_\phi(\mathcal{SG}(f_\theta(x))) - g_\phi(h_\psi(f_\theta(\tilde{x})))||_2 \\ &= \left\{[\mathcal{SG}(f_\theta(x)) - h_\psi(f_\theta(\tilde{x}))]^\top G(x,\tilde{x})^\top G(x,\tilde{x})[\mathcal{SG}(f_\theta(x)) - h_\psi(f_\theta(\tilde{x}))]\right\}^{1/2} \end{aligned} \tag{15}$$

The matrix $G^\top(x,\tilde{x})G(x,\tilde{x})$ is symmetric semi-positive definite matrix, thus its largest and smallest eigenvalue is non-negative and can be denoted as $\hat{\lambda}^2_{\text{max}}(x,\tilde{x})$ and $\hat{\lambda}^2_{\text{min}}(x,\tilde{x})$, where $\hat{\lambda}_{\text{max}}$ and $\hat{\lambda}_{\text{min}}$ are also non-negative. We use $l_{\text{CL}}(x,\tilde{x})$ to denote $||\mathcal{SG}(f_\theta(x)) - h_\psi(f_\theta(\tilde{x}))||_2$. Then

$$\lambda_{\text{min}} l_{\text{CL}}(x,\tilde{x}) \leq \hat{\lambda}_{\text{min}}(x,\tilde{x}) l_{\text{CL}}(x,\tilde{x}) \leq l_{\text{RC2}}(x,\tilde{x}) \leq \hat{\lambda}_{\text{max}}(x,\tilde{x}) l_{\text{CL}}(x,\tilde{x}) \leq \lambda_{\text{max}} l_{\text{CL}}(x,\tilde{x}), \tag{16}$$

where $\lambda_{\text{min}} = \inf_{x,\tilde{x}} \lambda_{\text{min}}(x,\tilde{x})$, $\lambda_{\text{max}} = \sup_{x,\tilde{x}} \lambda_{\text{max}}(x,\tilde{x})$ are constants independent on x and $\tilde{x}$. After taking expectation with respect to $p(x,\tilde{x})$ to the terms in the inequality, we obtain the inequality of loss functions in equation 14. □

Lemma B.6.

$$\mathcal{L}_{RC2} + \mathcal{L}_{reg} \geq \mathcal{L}_{RC} \geq \mathcal{L}_{RC2} - \mathcal{L}_{reg} \tag{17}$$

Proof. This is a straightforward application of the triangle inequality with the L^2 distance. Denote vectors $c \triangleq g_\phi(f_\theta(x)) \in \mathbb{R}^{d_\mathbb{X}}$, $d \triangleq g_\phi(h_\psi(f_\theta(\tilde{x}))) \in \mathbb{R}^{d_\mathbb{X}}$. Utilizing the Minkowski inequality, we derive $||c-d||_2+||x-c||_2 \geq ||(c-d)+(x-c)||_2 = ||x-d||_2$. This implies $\mathcal{L}_{\text{RC2}} + \mathcal{L}_{\text{reg}} \geq \mathcal{L}_{\text{RC}}$. Analogously, we establish $\mathcal{L}_{\text{RC}} + \mathcal{L}_{\text{reg}} \geq \mathcal{L}_{\text{RC2}}$. □

By employing Lemma B.5 and Lemma B.6, we establish that $\lambda_{\text{max}}\mathcal{L}_{\text{CL}}+\mathcal{L}_{\text{reg}} \geq \mathcal{L}_{\text{RC2}}+\mathcal{L}_{\text{reg}} \geq \mathcal{L}_{\text{RC}}$. When $\lambda_{\text{max}} > 0$, dividing both sides of the inequality by λ_{max} concludes the proof of equation 3. Conversely, we have $\mathcal{L}_{\text{RC}} + \mathcal{L}_{\text{reg}} \geq \mathcal{L}_{\text{RC2}} \geq \lambda_{\text{min}}\mathcal{L}_{\text{CL}}$, thus completing the proof of equation 4.

B.2. Discussing the Non-zero Conditions of Theorem 2.1

Concerning λ_{max}, it equals zero if and only if $G(x, \tilde{x}) = 0$ for all x and $\tilde{x}$. In this case, $g_{\phi,j}(f_\theta(x)) = g_{\phi,j}(h_\psi(f_\theta(\tilde{x})))$ for all j, x and $\tilde{x}$, indicating a collapsed decoder that maps all features to a constant vector. This case is avoided by optimizing $\mathcal{L}_{\text{reg}}$ or $\mathcal{L}_{\text{RC}}$.

As for λ_{min}, it is non-zero if and only if $d_\mathbb{Z} \leq d_\mathbb{X}$ and $G(x, \tilde{x})$ is full column rank for all x and $\tilde{x}$. The dimension condition is usually satisfied in practice since the input of the encoder is usually high-dimensional embeddings. We can gain more insights in the linear decoder case, where G is exactly the linear decoder matrix. Denote the feature and the reconstructed output as z and $\hat{x}$ respectively. Then we have $\hat{x} = Gz = \sum_{k=1}^{d_\mathbb{Z}} G_{:,k} z_k$, where $G_{:,k}$, $k = 1, \cdots d_\mathbb{Z}$ are the column vectors of G. Full column rank means $G_{:,k}$ is linearly independent, indicating there are no redundant dimensions in the feature for reconstruction. Otherwise, there exists an elementary matrix P s.t. GP has at least one column vector $(GP)_{:,l}$ is a zero vector. Then $(P^{-1}z)_l$ is redundant in the sense that it does not affect the reconstruction.

B.3. Proof of Theorem 2.2

Proof.

$$\begin{aligned}
&\mathcal{L}_{\text{CL}}^{(\text{SimSiam})} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f_\theta(x))||_2^2 \text{ , where the representations are normalized.} \\
&= 2 - 2\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)} f_\theta(x)^\top h_\psi(f_\theta(\tilde{x})) \\
&= 2 - \mathbb{E}_{p(x)} 2 f_\theta(x)^\top \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x})) \mathrm{d}\tilde{x} \right) \\
&\geq 2 - \mathbb{E}_{p(x)} \left[||f_\theta(x)||_2^2 + \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x})) \mathrm{d}\tilde{x} \right)^\top \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x})) \mathrm{d}\tilde{x} \right) \right] \\
&= 1 - \int\int \mathbb{E}_{p(x)} p(\tilde{x}|x) p(\tilde{x}'|x) h_\psi(f_\theta(\tilde{x}))^\top h_\psi(f_\theta(\tilde{x}')) \mathrm{d}\tilde{x}\mathrm{d}\tilde{x}'.
\end{aligned} \tag{18}$$

Denote $\omega(\tilde{x}, \tilde{x}') \triangleq \mathbb{E}_{p(x)} p(\tilde{x}|x) p(\tilde{x}'|x)$ to establish a probability of generating two augmentations $\tilde{x}$ and $\tilde{x}'$ from the same input, and it is normalized $\int\int \omega(\tilde{x}, \tilde{x}') \mathrm{d}\tilde{x}\mathrm{d}\tilde{x}' = \mathbb{E}_{p(x)} \int p(\tilde{x}|x) \mathrm{d}\tilde{x} \int p(\tilde{x}'|x) \mathrm{d}\tilde{x}' = 1$. A similar concept is initially introduced in (HaoChen et al., 2021) as the weight of the augmentation graph.

$$\begin{aligned}
\mathcal{L}_{\text{CL}}^{(\text{SimSiam})} &\geq 1 - \mathbb{E}_{\omega(\tilde{x},\tilde{x}')} h_\psi(f_\theta(\tilde{x}))^\top h_\psi(f_\theta(\tilde{x}')) \\
&= \frac{1}{2}\mathbb{E}_{\omega(\tilde{x},\tilde{x}')} ||h_\psi(f_\theta(\tilde{x})) - h_\psi(f_\theta(\tilde{x}'))||_2^2 \triangleq \frac{1}{2}\mathcal{L}_{symm}
\end{aligned} \tag{19}$$

Therefore, the SimSiam loss can serve as an upper bound for the symmetric contrastive loss that aligns the positive pairs. Then we convert the pairwise alignment into the centered clustering formulation.

$$\begin{aligned}
\mathcal{L}_{\text{cluster}} &= \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)} ||h_\psi(f_\theta(\tilde{x})) - \mathbb{E}_{p(\tilde{x}|x)} h_\psi(f_\theta(\tilde{x}))||_2^2 \\
&= \mathbb{E}_{p(x)} \left(\mathbb{E}_{p(\tilde{x}|x)} ||h_\psi(f_\theta(\tilde{x}))||_2^2 - \left(\mathbb{E}_{p(\tilde{x}|x)} h_\psi(f_\theta(\tilde{x}))\right)^\top \left(\mathbb{E}_{p(\tilde{x}|x)} h_\psi(f_\theta(\tilde{x}))\right) \right) \\
&= \frac{1}{2}\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\mathbb{E}_{p(\tilde{x}'|x)} ||h_\psi(f_\theta(\tilde{x})) - h_\psi(f_\theta(\tilde{x}'))||_2^2 \\
&= \frac{1}{2}\mathbb{E}_{\omega(\tilde{x},\tilde{x}')} ||h_\psi(f_\theta(\tilde{x})) - h_\psi(f_\theta(\tilde{x}'))||_2^2 = \frac{1}{2}\mathcal{L}_{symm} \leq \mathcal{L}_{\text{CL}}^{(\text{SimSiam})}
\end{aligned} \tag{20}$$

□

B.4. Corollary of Theorem 2.2

Corollary B.7 (Relations between cross-modal contrastive learning and clustering)**.** *Theorem 2.2 still holds when the input and its augmentations employ distinct encoders or parameters, for example in the case of cross-modal contrastive learning.*

$$\mathcal{L}_{CL,cross\text{-}modal}^{(SimSiam)} \triangleq \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f'_{\theta'}(x))||_2^2 \geq \mathcal{L}_{cluster}, \tag{21}$$

where the representations are defined to be normalized: $||f'_{\theta'}(\cdot)||_2^2 = ||h_\psi(f_\theta(\cdot))||_2^2 = 1$. $\mathcal{L}_{cluster} \triangleq \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)} ||h_\psi(f_\theta(\tilde{x})) - \mathbb{E}_{p(\tilde{x}|x)}h_\psi(f_\theta(\tilde{x}))||_2^2$ *describes the mean distance between the samples in the cluster and the cluster center in the representation space of augmentations.*

Proof. The heterogeneous encoders do not alter the result of equation 18 since the encoder of the input $f'_{\theta'}$ will diminish as long as the representation is normalized.

$$\begin{aligned}
&\mathcal{L}_{\text{CL,cross-modal}}^{(\text{SimSiam})} = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f'_{\theta'}(x))||_2^2 \\
&= 2 - 2\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)} f'_{\theta'}(x)^\top h_\psi(f_\theta(\tilde{x})) \\
&= 2 - \mathbb{E}_{p(x)} 2 f'_{\theta'}(x)^\top \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x}))\mathrm{d}\tilde{x}\right) \\
&\geq 2 - \mathbb{E}_{p(x)}\left[||f'_{\theta'}(x)||_2^2 + \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x}))\mathrm{d}\tilde{x}\right)^\top \left(\int p(\tilde{x}|x) h_\psi(f_\theta(\tilde{x}))\mathrm{d}\tilde{x}\right)\right] \\
&= 1 - \int\int \mathbb{E}_{p(x)} p(\tilde{x}|x) p(\tilde{x}'|x) h_\psi(f_\theta(\tilde{x}))^\top h_\psi(f_\theta(\tilde{x}'))\mathrm{d}\tilde{x}\mathrm{d}\tilde{x}'.
\end{aligned} \tag{22}$$

Hence, we reach an identical result as expressed in Equation equation 18. The remaining proof aligns precisely with those presented in the proof of Theorem 2.2. □

C. Experimental Details

C.1. Dataset Description

QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) is a quantum chemistry dataset that offers a single equilibrium conformation along with 12 labels covering geometric, energetic, electronic, and thermodynamic properties for 134,000 stable small organic molecules comprised of CHONF atoms. The dataset is split following typical settings, resulting in a training set of 110,000 samples, a validation set of 10,000 samples, and a test set containing the remaining 10,831 samples.

MD17 (Chmiela et al., 2017) comprises molecular dynamics trajectories for 8 small organic molecules. Each molecule in the dataset is associated with 150k to almost 1M conformations, and includes total energy and force labels. Our focus lies on the challenging task of force prediction. In accordance with a standard limited data setting, the model undergoes training on a subset of 1000 samples, with 50 allocated for validation, while the remaining data is employed for testing.

MD22 (Chmiela et al., 2023) consists of molecular dynamics trajectories covering 7 molecules from four major classes of biomolecules and supramolecules, ranging from a 42-atom peptide to a double-walled nanotube containing 370 atoms. The dataset provides labels for both total energy and force. Our focus lies on the challenging task of force prediction. The dataset is split by a ratio of 8:1:1 into the train, validation, and test sets.

MoleculeNet (Wu et al., 2018) is widely recognized as a benchmark for predicting a range of molecular properties. It covers biological tasks such as BBBP, SIDER, ClinTox, Tox21, Toxcast, BACE, HIV, and MUV, as well as physicochemical challenges, including FreeSolv, ESOL, and Lipophilicity. Each dataset is split into training, validation, and test sets following an 8:1:1 ratio, implemented through the scaffold method.

C.2. Construction of Hierarchical Data

The construction of the dataset utilized in 5.3.2 begins with the selection of five common heterocyclic scaffolds, each possessing different chemical properties but exhibiting a similar structure to make the clustering more difficult. Subsequently, we randomly sample 128 molecules for each scaffold and conduct molecular simulations on each molecule using OpenMM (Eastman et al., 2013), generating four molecular dynamics trajectories. From each trajectory, we randomly

select four adjacent conformations, resulting in approximately 16 conformations per molecule. In total, this meticulously constructed dataset consists of around 10,000 conformations.

C.3. Hyperparameter Settings

Table 8. Hyperparameters for pre-training.

Parameter	Value or description
Batch size	256
Optimizer	AdamW
Adam betas	(0.9, 0.999)
Max Learning rate	0.0004
Warm up steps	10000
Learning rate decay policy	Cosine
Learning rate factor	0.8
Training steps	1500000
3D encoder layers number	8
3D encoder attention head number	8
3D encoder embedding dimension	256
2D encoder layers number	12
2D encoder attention head number	32
2D encoder embedding dimension	512

In our pre-training stage, we employ TorchMD-NET (Thölke & De Fabritiis, 2022) as the 3D encoder backbone and Graphormer (Ying et al., 2021) as the 2D encoder. The loss weights for fragment masking loss, denoising loss, and cross-modal distillation loss are set at a ratio of 1:1:1. Within the masking task, the masking ratio is configured to be 0.2. For the denoising task, the standard deviations of torsion Gaussian noise and coordinate Gaussian noise are set to 2 and 0.04, respectively. The temperature τ is set to 0.5 for the cross-modal distillation task. Additional hyperparameters associated with the network structure and pre-training process can be found in Table 8.

In line with previous methods, we employ grid search to find the optimal hyperparameters for tasks in MoleculeNet. The specific search space for each task is detailed in Table 9.

The hyperparameters for fine-tuning on QM9, MD17, and MD22 are outlined in Tables 10, 11 and 12. When performing fine-tuning, we incorporate the Noisy Node task, following the approach in (Godwin et al., 2021; Feng et al., 2023b). The loss weight for this task is set to 0.1.

C.4. The Impact of Data Accuracy and Diversity

Both the accuracy and diversity of 3D pre-training data have an impact on downstream task results. While we do not have a definitive conclusion, we have observed some phenomena that offer valuable insights for the community.

From the perspective of accuracy and diversity of 3D conformers, our pre-training data can be categorized into two distinct types. The first type comprises data sourced from the PubChemQC Project (Nakata & Shimazaki, 2017), where conformations are calculated using the DFT method, ensuring high quality. However, due to the substantial computational costs involved, this dataset contains a limited number and variety of molecules, resulting in less diversity. The second type consists of data collected from various sources via the Uni-Mol (Zhou et al., 2023), utilizing RDKit to generate conformers more efficiently. Consequently, this dataset encompasses a broader range of molecules, enhancing diversity.

Through comprehensive experiment and analysis, we conclude that the denoising task prefers accurate conformation. Conversely, for masking and contrastive learning tasks, the diversity of the dataset plays a more critical role. As demonstrated by previous 3D denoising methods (Zaidi et al., 2022; Feng et al., 2023a), the denoising task relies on highly accurate equilibrium conformations as input. As observed in Table 13, pre-training UniCorn solely with DFT-calculated data achieves comparable performance on QM9 compared with UniCorn which uses additional RDKit data. Since quantum tasks such as QM9 and MD17 primarily involve small and simple molecules, the accuracy of the data may be more important than its diversity in the denoising task. Data diversity is more important in the context of masking and 2D-3D contrastive learning pre-training tasks, because their corresponding downstream tasks in MoleculeNet contain more complex and diverse

Table 9. Search space for MoleculeNet dataset, where [...] represents continuous interval, {...} denotes discrete candidate values.

Parameter	Classification tasks	Regression tasks
Learning rate	[1e-4,1e-3]	[1e-4,1e-3]
Batch size	{16,32,64}	{8,16,32}
Epochs	{5,10,25,50}	{25,50}
Weight decay	{0,1e-5}	{0,1e-5}

Table 10. Hyperparameters for fine-tuning on QM9.

Parameter	Value or description
Train/Val/Test Splitting	110000/10000/remaining data
Batch size	128
Optimizer	AdamW
Warm up steps	10000
Max Learning rate	0.0004
Learning rate decay policy	Cosine
Learning rate factor	0.8
Cosine cycle length	300000 (500000 for tasks α, $ZPVE$, U_0, U, H, G)

molecules covering a border range of biological properties that don't need highly accurate conformation. Moreover, more diverse 3D conformations of one molecule are also beneficial for 2D-3D contrastive learning as demonstrated by GraphMVP and 3D InfoMax. As depicted in Table 14, reveals that UniCorn pre-trained with additional RDKit data outperforms UniCorn solely pre-trained with DFT data on biological and physicochemical tasks.

D. Related Work

D.1. Unified Molecular Pre-training Methods

Recently, several unified pre-training methods have been proposed to be adopted for a diverse range of molecular tasks. Zhou et al. (2023) and Fang et al. (2022) develop pre-training tasks centered around the molecular 3D structure. involving the prediction of geometry information and 3D denoising. Other multimodal pre-training methods focus on integrating molecular 2D graphs and 3D structures. Luo et al. (2023) designs a versatile Transformer structure capable of accommodating both 2D and 3D inputs, engages in denoising and quantum property prediction as pre-training tasks. Liu et al. (2023a) and Li et al. (2023a) take a unique approach by incorporating molecular 2D and 3D information into the diffusion process for representation learning. Yu et al. (2023) merges atom relations involving different modalities, into a unified relation matrix, then predicts 2D graph and 3D structure information separately. However, previous methods have fallen short in comprehending the in-depth relationship between pre-training methods and downstream tasks, and struggle to achieve satisfactory results across both microscopic and macroscopic tasks.

D.2. Unifying Masking and Contrastive Learning in Other Fields

Recently, some theoretical works have studied masking and contrastive learning in other fields. In the field of computer vision, Zhang et al. (2022a) investigates in the mask image modeling setting and lower bound the masking loss by a type of contrastive loss that aligns the mask-induced positive pairs. Qi et al. (2023) provides a unified view of masking and contrastive learning through the student-teacher paradigm. In the field of graph representation learning, Li et al. (2023b) analyzes masking graph modeling in an information-theoretic view. They first prove that graph autoencoder and contrastive learning have asymptotic equivalent solutions and then explain the benefit of graph masking by reducing redundancy between views.

E. Pseudocode for Pre-training and Fine-tuning Algorithms

We provide Algorithm 1 and Algorithm 2 presenting the pseudocode for the pre-training and fine-tuning processes of UniCorn, respectively.

Table 11. Hyperparameters for fine-tuning on MD17.

Parameter	Value or description
Train/Val/Test Splitting	950/50/remaining data
Batch size	8
Optimizer	AdamW
Warm up steps	1000
Max Learning rate	0.0005
Learning rate decay policy	ReduceLROnPlateau (Reduce Learning Rate on Plateau) scheduler
Learning rate factor	0.8
Patience	30
Min learning rate	1.00E-07
Force weight	0.8
Energy weight	0.2

Table 12. Hyperparameters for fine-tuning on MD22.

Parameter	Value or description
Train/Val/Test Splitting	80%/10%/10% split same with (Wang et al., 2023)
Batch size	4 for double-walled-nanotube 16 for Ac-Ala3-NHMe and DHA 8 for the other 4 tasks
Optimizer	AdamW
Epochs	50
Max Learning rate	0.001
Learning rate decay policy	Cosine
Warm up steps	30% steps of the first training epoch
Patience steps	70% steps of the first training epoch
Min learning rate	1.00E-07
Force weight	0.8
Energy weight	0.2

Table 13. Performance (MAE, ↓) on QM9 quantum tasks. The best results are in bold.

QM9	ϵ_{HOMO} (meV)	ϵ_{LOMO} (meV)	$\Delta\epsilon$(meV)	AVG
UniCorn (DFT)	13.2	12	25.1	16.7
UniCorn (DFT + RDkit)	**13**	**11.9**	**24.9**	**16.6**

Table 14. Performance (ROC-AUC%, ↑; RMSE, ↓) on MoleculeNet tasks. The best results are in bold.

Models	BBBP↑	BACE↑	ClinTox↑	Tox21↑	ToxCast↑	SIDER↑	FreeSolv↓	Lipo↓
UniCorn(DFT)	73.0(0.9)	83.6(0.2)	88.5(1.4)	79.2(1.8)	68.6(0.3)	63.1(0.7)	1.831(0.189)	0.589(0.008)
UniCorn(DFT + RDKit)	**74.2(1.1)**	**85.8(1.2)**	**92.1(0.4)**	**79.3(0.5)**	**69.4(1.1)**	**64.0(1.8)**	**1.555(0.075)**	**0.591(0.016)**

F. The Relationship between Pre-training Methods and Data Modalities

In this section, we discuss the relationship between pre-training methods and data modalities, although we did not explicitly refer to this in the main text. We select data modalities that are best suited for these utilized pre-training methods and focus primarily on the relationship between pre-training methods and downstream tasks. However, understanding the connection between pre-training methods and data modalities is crucial for comprehensively understanding the proposed methods.

Different pre-training strategies have certain preferences for specific modalities. For instance, previous 3D denoising

Algorithm 1 Applying UniCorn to pre-training

Require:
f^{2d}: 2D encoder
f^{3d}: 3D encoder
g_m: MLP head for 2D masking
g_d: MLP head for 3D denoising
h_p: MLP head for feature alignment between 3D and 2D representation
$\mathcal{G}$: 2D molecular graph
$\mathcal{C} = (\mathcal{V}, \mathcal{X})$: 3D molecular conformation
m: Mask ratio of 2D fragment masking
X: Unlabeled pre-training dataset
x_i: Input sample
T: Training steps

while $T \neq 0$ **do**
  x_i = dataloader(X), x_i=$(\mathcal{G}, \mathcal{C})$ ▷ random sample x_i from X, x_i contains 2D molecular graph and 3D conformation
  According to the mask ratio m, randomly mask a specific number of fragments denoted as $\mathbf{s}_m$ in $\mathcal{G}$, resulting in the masked 2d graph $\mathcal{G}_{\setminus m}$
  $\mathcal{L}_m = -\mathbb{E}_{\mathcal{G},\mathcal{G}_{\setminus m}} \sum_{\mathbf{s}_m} \log \left[p_{(g_m, f^{2d})}\left(\mathbf{s}_m | \mathcal{G}_{\setminus m}\right)\right]$ ▷ Calculate the masked fragment loss
  Introduce torsion Gaussian noise to $\mathcal{C}$ to obtain $\mathcal{C}_a = (\mathcal{V}, \mathcal{X}_a)$, then add the coordinate Gaussian noise to $\mathcal{C}_a$ to derive $\tilde{\mathcal{C}} = (\mathcal{V}, \tilde{\mathcal{X}})$
  $\mathcal{L}_{dn} = \mathbb{E}_{\tilde{\mathcal{C}},\mathcal{C}_a} ||g_d(f^{3d}(\tilde{\mathcal{C}})) - (\tilde{\mathcal{X}} - \mathcal{X}_a)||_2^2$ ▷ Calculate the 3D denoising loss
  $\mathcal{L}_{cl} = -\mathbb{E}_{\mathcal{G},\tilde{\mathcal{C}}}[\log \frac{e^{\cos(\boldsymbol{z}^g, \boldsymbol{z}^c)/\tau}}{\sum_{\boldsymbol{z}_j^c} e^{\cos(\boldsymbol{z}^g, \boldsymbol{z}_j^c)/\tau}} + \log \frac{e^{\cos(\boldsymbol{z}^c, \boldsymbol{z}^g)/\tau}}{\sum_{\boldsymbol{z}_j^g} e^{\cos(\boldsymbol{z}^c, \boldsymbol{z}_j^g)/\tau}}]$, where $\boldsymbol{z}^c = h_p(f^{3d}(\tilde{\mathcal{C}}))$, $\boldsymbol{z}^g = f^{2d}(\mathcal{G})$, $\boldsymbol{z}_j^c \in \mathbb{N}^c \cup \boldsymbol{z}^c$ and $\boldsymbol{z}_j^g \in \mathbb{N}^g \cup \boldsymbol{z}^g$, $\mathbb{N}^c$ and $\mathbb{N}^g$ denotes in-batch negative samples of 2D graphs and 3D conformations ▷ Calculate the Cross-modal contrastive loss
  Loss = $\mathcal{L}_{dn} + \mathcal{L}_m + \mathcal{L}_{cl}$
  Optimise(Loss)
  $T = T - 1$
end while

methods (Zaidi et al., 2022; Feng et al., 2023a; Ni et al., 2024) are tailored for 3D conformations, as they introduce different types of noise to the 3D structure and predict the noise aiming to approximately learn the molecular force field. In contrast, masking and contrastive learning show a slight preference for certain modalities but are generally more versatile. While masking is typically conducted on the 2D graph by masking the atoms, edges, and fragments (Hu et al., 2020; Xia et al., 2022; Feng et al., 2023c), it can also applied to masking atoms within 3D structures (Zhou et al., 2023; Zaidi et al., 2022) and masking tokens for 1D SMILES sequence (Chithrananda et al., 2020; Wang et al., 2019; Zhang et al., 2021). Contrastive learning is particularly flexible, as it can generally align different modalities. In addition to aligning 2D graphs with 3D conformations, it can also be used to align molecules with other relevant modalities such as textual descriptions (Liu et al., 2023b) or binding proteins (Gao et al., 2022). In our analysis, we implicitly take into account the preference for modality, focusing on one of the most common scenarios: denoising for 3D modality, masking for 2D modality, and contrastive learning for cross-modality in the main text.

G. Novelty and Limitations

It's essential to note that the primary novelty of this paper lies in our combination of three prevalent existing methods under a unified contrastive learning framework for pursuing a universal molecular model. Through rigorous theoretical proof and domain-expertise-driven analysis, we innovatively establish the necessity of this combination to achieve a universal representation applicable to all types of downstream tasks—an objective that previous works have failed to attain. In particular, we demonstrate for the first time that existing methods are naturally suited to different downstream tasks, shedding light on why current literature is limited to their specific preferred task. Finally, through exhaustive experiments, our approach marks the first work in the field of molecular representation learning that achieves SOTA results on tasks that span a wide range of quantum, physicochemical, and biological domains.

Algorithm 2 Applying UniCorn to fine-tuning

Require:

f^{3d}:Pre-trained 3D encoder
h_p: MLP head for property prediction
h_d: MLP head for Noisy Nodes
$\mathcal{C} = (\mathcal{V}, \mathcal{X})$: 3D molecular conformation
X: Training dataset
x_i: Input sample
y_i: Label of x_i
$\Delta x_i \sim \mathcal{N}(0, \tau^2 I_{3N})$, N is atom number of x_i
T: Training steps
$task$: the current fine-tuning downstream task, which may belong to MoleculeNet, QM9, MD17, or MD22
λ_n: Loss weight of Noisy Nodes

while $T \neq 0$ **do**
  x_i, y_i = dataloader(X), x_i=($\mathcal{C}$) ▷ random sample x_i and corresponding label y_i from X
  $y_i^{pred} = h_p(f^{3d}(\tilde{x}_i))^*$
  Loss = PropertyPredictionLoss(y_i^{pred}, y_i)
  if $task$ in QM9, MD17, MD22 **then**
    $L_d = \lambda_n ||h_d(f^{3d}(\tilde{x}_i)) - \Delta x_i||_2^{2*}$ ▷ Calculate loss of Noisy Nodes
    Loss += $\mathcal{L}_d$
  end if
  Optimise(Loss)
  $T = T - 1$
end while

*: For MoleculeNet, $\tilde{x}_i = x_i$. For QM9, MD17 and MD22, we apply Noisy Nodes regularization and the definition of $\tilde{x}_i$ follows (Feng et al., 2023a).

We anticipate potential limitations with UniCorn concerning its data requirements. Firstly, as UniCorn aims to capture multi-grained molecular representation and involves denoising, it relies on 3D equilibrium structures as pre-training inputs, which are relatively scarcer than 2D molecular data in existing datasets. Secondly, since the 3D encoder is utilized for various downstream tasks, the necessity of 3D conformation as input may require additional 3D generation during processing. Thirdly, the accuracy of 3D inputs or failure of 3D generation could impact the outcomes of downstream tasks. However, as 3D datasets expand and 3D generation capabilities improve, we expect the limitation of data requirements for our approach as well as similar previous works like Uni-Mol (Zhou et al., 2023), GEM (Fang et al., 2022), and MoleBLEND (Yu et al., 2023), to be mitigated in the future.