Uni-Mol: A Universal 3D Molecular Representation Learning Framework

Gengmo Zhou[1,2]*, **Zhifeng Gao**[2*†], **Qiankun Ding**[2], **Hang Zheng**[2]
Hongteng Xu[1], **Zhewei Wei**[1], **Linfeng Zhang**[2,3], **Guolin Ke**[2†]
[1]Renmin University [2]DP Technology [3]AI for Science Institute, Beijing
{zgm2015, hongtengxu, zhewei}@ruc.edu.cn
{gaozf, dingqk, zhengh, zhanglf, kegl}@dp.tech

Abstract

Molecular representation learning (MRL) has gained tremendous attention due to its critical role in learning from limited supervised data for applications like drug design. In most MRL methods, molecules are treated as 1D sequential tokens or 2D topology graphs, limiting their ability to incorporate 3D information for downstream tasks and, in particular, making it almost impossible for 3D geometry prediction or generation. Herein, we propose Uni-Mol, a universal MRL framework that significantly enlarges the representation ability and application scope of MRL schemes. Uni-Mol is composed of two models with the same SE(3)-equivariant transformer architecture: a molecular pretraining model trained by 209M molecular conformations; a pocket pretraining model trained by 3M candidate protein pocket data. The two models are used independently for separate tasks, and are combined when used in protein-ligand binding tasks. By properly incorporating 3D information, Uni-Mol outperforms SOTA in 14/15 molecular property prediction tasks. Moreover, Uni-Mol achieves superior performance in 3D spatial tasks, including protein-ligand binding pose prediction, molecular conformation generation, etc. Finally, we show that Uni-Mol can be successfully applied to the tasks with few-shot data like pocket druggability prediction. The model and data will be made publicly available at https://github.com/dptech-corp/Uni-Mol.

1 Introduction

Recently, representation learning (or pretraining, self-supervised learning) [1, 2, 3] has been prevailing in many applications, such as BERT [4] and GPT [5, 6, 7] in Natural Language Processing (NLP), ViT [8] in Computer Vision (CV), etc. These applications have a common characteristic: unlabeled data is abundant, while labeled data is limited. As a solution, in a typical representation learning method, one first adopts a pretraining procedure to learn a good representation from large-scale unlabeled data, and then a finetuning scheme is followed to extract more information from limited supervised data.

Applications in the field of drug design share the characteristic that calls for representation learning schemes. The chemical space that a drug candidate lies in is vast, while drug-related labeled data is limited. Not surprisingly, compared with traditional molecular fingerprint based models [9, 10], recent molecular representation learning (MRL) models perform much better in most property prediction tasks [11, 12, 13]. However, to further improve the performance and extend the application scope of existing MRL models, one is faced with a critical issue. From the perspective of life science, the properties of molecules and the effects of drugs are mostly determined by their 3D structures [14,

*Equal contribution.
†Corresponding authors.

Preprint. Under review.

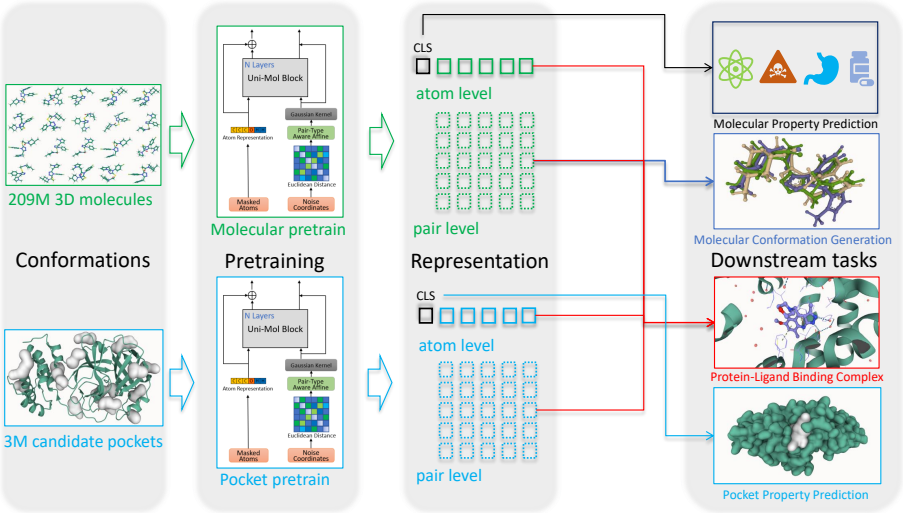

Figure 1: Schematic illustration of the Uni-Mol framework. Uni-Mol is composed of two models: a molecular pretraining model trained by 209M molecular 3D conformations; a pocket pretraining model trained by 3M candidate protein pocket data. The two models are used independently for separate tasks, and are combined when used in protein-ligand binding tasks.

15]. In most current MRL methods, one starts with representing molecules as 1D sequential strings, such as SMILES [16, 17, 18] and InChI [19, 20, 21], or 2D graphs [22, 11, 23, 12]. This may limit their ability to incorporate 3D information for downstream tasks. In particular, this makes it almost impossible for 3D geometry prediction or generation, such as, e.g., the prediction of protein-ligand binding pose [24]. Even though there have been some recent attempts trying to leverage 3D information in MRL [25, 26], the performance is less than optimal, possibly due to the small size of 3D datasets, and 3D positions can not be used as inputs/outputs during finetuning, since they only serve as auxiliary information.

In this work, we propose Uni-Mol, to our best knowledge, the first universal 3D molecular pretraining framework, which is derived from large-scale unlabeled data and is able to directly take 3D positions as both inputs and outputs. Uni-Mol consists of 3 parts. 1) *Backbone*. Based on Transformer, the invariant spatial positional encoding and pair level representation are added to better capture the 3D information. Moreover, an equivariant head is used to directly predict 3D positions. 2) *Pretraining*. We create two large-scale datasets, a 209M molecular conformation dataset and a 3M candidate protein pocket dataset, for pretraining 2 models on molecules and protein pockets, respectively. For the pretraining tasks, besides masked atom prediction, a 3D position denoising task is used for learning 3D spatial representation. 3) *Finetuning*. According to specific downstream tasks, the used pretraining models are different. For example, in molecular property prediction tasks, only the molecular pretraining model is used; in protein-ligand binding pose prediction, both two pretraining models are used. We refer to Fig. 1 for an overall schematic illustration of the Uni-Mol framework.

To demonstrate the effectiveness of Uni-Mol, we conduct experiments on a series of downstream tasks. In the molecular property prediction tasks, Uni-Mol outperforms SOTA on 14/15 datasets on the MoleculeNet benchmark. In 3D geometric tasks, Uni-Mol also achieves superior performance. For the pose prediction of protein-ligand complexes, Uni-Mol predicts 88.07% binding poses with RMSD <= 2Å, 22.81% more than popular docking methods, and ranks 1st in the docking power test on CASF-2016 [27] benchmark. Regarding molecular conformation generation, Uni-Mol achieves SOTA for both Coverage and Matching metrics on GEOM-QM9 and GEOM-Drugs [28]. Moreover, Uni-Mol can be successfully applied to tasks with very limited data like pocket druggability prediction.

2 Uni-Mol Framework

In this section, we introduce the Uni-Mol framework by showing the details of the backbone, the pretraining scheme, and the finetuning scheme. We refer to Fig. 2 for a schematic illustration of the model architecture.

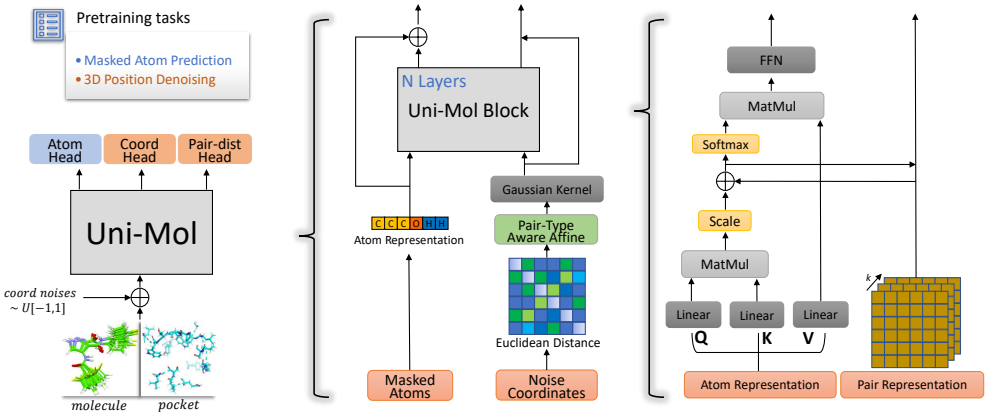

Figure 2: Left: the overall pretraining architecture. Middle: the model inputs, including atoms and spatial positional encoding created by pair Euclidean distance. Right: pair representation and its update process.

2.1 Backbone

Transformer [29] is the default backbone in representation learning. However, Transformer was originally designed for NLP tasks and cannot handle 3D spatial data directly. To tackle this, based on the standard Transformer with Pre-LayerNorm [30] backbone, we introduce several modifications.

Invariant spatial positional encoding Due to its permutationally invariant property, Transformer cannot distinguish the positions of inputs without positional encoding. Different with the discrete (ordinal) positions used in NLP/CV [31, 32], the positions in 3D space, i.e. coordinates, are continuous values. Besides, the positional encoding procedure needs to be invariant under global rotation and translation. To achieve that, similar to the relative positional encoding, we simply use Euclidean distances of all atom pairs, as well as pair-type aware Gaussian kernels [33]. Formally, the D-channel positional encoding of atom pair ij is denoted as

$$\boldsymbol{p}_{ij} = \{\mathcal{G}(\mathcal{A}(d_{ij}, t_{ij}; \boldsymbol{a}, \boldsymbol{b}), \mu^k, \sigma^k) | k \in [1, D]\}, \quad \mathcal{A}(d, r; \boldsymbol{a}, \boldsymbol{b}) = a_r d + b_r, \tag{1}$$

where $\mathcal{G}(d, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}}$ is a Gaussian density function with parameters μ and σ, d_{ij} is the Euclidean distance of atom pair ij, and t_{ij} is the pair-type of atom pair ij. Please note the pair-type here is not the chemical bond, and it is determined by the atom types of pair ij. $\mathcal{A}(d_{ij}, t_{ij}; \boldsymbol{a}, \boldsymbol{b})$ is the affine transformation with parameters $\boldsymbol{a}$ and $\boldsymbol{b}$, it affines d_{ij} corresponding to its pair-type t_{ij}. Except d_{ij} and t_{ij}, all remaining parameters are trainable and randomly initialized.

Pair representation By default, Transformer maintains the token(atom) level representation, which is later used in finetuning downstream tasks. Nevertheless, as the spatial positions are encoded at pair-level, we also maintain the pair-level representation, to better learn the 3D spatial representation. Specifically, the pair representation is initialized as the aforementioned spatial positional encoding. Then, to update pair representation, we use the atom-to-pair communication via the multi-head Query-Key product results in self-attention. Formally, the update of ij pair representation is denoted as

$$\boldsymbol{q}_{ij}^0 = \boldsymbol{p}_{ij}\boldsymbol{M}, \quad \boldsymbol{q}_{ij}^{l+1} = \boldsymbol{q}_{ij}^l + \{\frac{\boldsymbol{Q}_i^{l,h}(\boldsymbol{K}_j^{l,h})^T}{\sqrt{d}} | h \in [1, H]\}, \tag{2}$$

where $\boldsymbol{q}_{ij}^l$ is the pair representation of atom pair ij in l-th layer, H is the number of attention heads, d is the dimension of hidden representations, $\boldsymbol{Q}_i^{l,h}$ ($\boldsymbol{K}_j^{l,h}$) is the Query (Key) of the i-th (j-th) atom in the l-th layer h-th head, and $\boldsymbol{M} \in \mathbb{R}^{D \times H}$ is the projection matrix to make the representation the same shape as multi-head Query-Key product results.

Besides, to leverage 3D information in the atom representation, we also introduce the pair-to-atom communication, by using the pair representation as the bias term in self-attention. Formally, the

self-attention with pair-to-atom communication is denoted as

$$\text{Attention}(\boldsymbol{Q}_i^{l,h}, \boldsymbol{K}_j^{l,h}, \boldsymbol{V}_j^{l,h}) = \text{softmax}(\frac{\boldsymbol{Q}_i^{l,h}(\boldsymbol{K}_j^{l,h})^T}{\sqrt{d}} + \boldsymbol{q}_{ij}^{l-1,h})\boldsymbol{V}_j^{l,h}, \tag{3}$$

where $\boldsymbol{V}_j^{l,h}$ is the Value of the j-th atom in the l-th layer h-th head. The pair representation and atom-pair communication are firstly proposed in the Evoformer in AlphaFold [34], but the cost of Evoformer is extremely large. In Uni-Mol, as we keep them as simple as possible, the extra cost of maintaining pair representation is negligible.

SE(3)-Equivariance coordinate head With 3D spatial positional encoding and pair representation, the model can learn a good 3D representation. However, it still lacks the ability to directly output coordinates, which is essential in 3D spatial tasks. To this end, we add a simple SE(3)-equivariance head to Uni-Mol. Following the idea of EGNN [35], the design of SE(3)-equivariance head is denoted as

$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sum_{j=1}^{n} \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)c_{ij}}{n}, \quad c_{ij} = \text{ReLU}((\boldsymbol{q}_{ij}^{L} - \boldsymbol{q}_{ij}^{0})\boldsymbol{U})\boldsymbol{W}, \tag{4}$$

where n is the number of total atoms, L is the number of layers in model, $\boldsymbol{x}_i \in \mathbb{R}^3$ is the input coordinate of i-th atom, and $\hat{\boldsymbol{x}}_i \in \mathbb{R}^3$ is the output coordinate of i-th atom, $\text{ReLU}(y) = \max(0, y)$ is Rectified Linear Unit [36], $\boldsymbol{U} \in \mathbb{R}^{H \times H}$ and $\boldsymbol{W} \in \mathbb{R}^{H \times 1}$ are the projection matrices to convert pair representation to scalar.

2.2 Pretraining

For the purpose of pretraining, we generate two large-scale datasets, one composed of 3D structures of organic molecules, and another composed of 3D structures of candidate protein pockets. Then, two models are pretrained using these two datasets, respectively. As pockets are directly involved in many drug design tasks, intuitively, the pretraining on candidate protein pockets can boost the performance of tasks related to protein-ligand structures and interactions.

The molecular pretraining dataset is based on multiple public datasets (See Appendix A for more information). After normalizing and deduplicating, it contains about 19M molecules. To generate 3D conformations, we use ETKGD [37] with Merck Molecular Force Field [38] optimization in RDKit [39] to randomly generate 10 conformations for each molecule. We also generate an additional 2D conformation (based on the molecular graph), to avoid some rare cases that fail to generate 3D conformations.

The protein pocket pretraining dataset is derived from the Protein Data Bank (RCSB PDB [3]) [40], a collection of 180K 3D structures of proteins. To extract candidate pockets, we first clean the data by adding the missing side chains and hydrogen atoms; then we use Fpocket [41] to detect possible binding pockets of the proteins; and finally, we filter pockets by the number of residues in contact with and retains water molecules in the pocket. In this way, We collect a dataset composed of 3.2M candidate pockets for pretraining.

Self-supervised task is vitally important for effective learning from large-scale unlabeled data. For example, the masked token prediction task in BERT [4] encourages the model to learn the contextual information. Similar to BERT, the masked atom prediction task is used in Uni-Mol. For each molecule/pocket, we add a special atom `[CLS]`, whose coordinate is the center of all atoms, to represent the whole molecule/pocket. However, as 3D spatial positional encoding leaks chemical bonds, atom types could be inferred easily, and therefore, the masked atom prediction cannot encourage the model to learn useful information. To tackle this, as well as learning from 3D information, we design a 3D position denoising task. Particularly, uniform noises of [-1 Å, 1 Å] are added to the random 15% atom coordinates, then the spatial positional encoding is calculated based on corrupted coordinates. In this way, the masked atom prediction task becomes non-trivial. Besides, two additional heads are used to recover the correct spatial positions. 1) Pair-distance prediction. Based on pair-representation, the model needs to predict the correct Euclidean distances of the atoms pairs with corrupted coordinates. 2) Coordinate prediction. Based on SE(3)-Equivariance coordinate head, the model needs to predict the correct coordinates for the atoms with corrupted coordinates.

[3] http://www.rcsb.org/

Both 2 pretraining models use the same self-supervised tasks described above, and Figure 2 is the illustration of the overall pretraining framework. For the detailed configurations of pretraining, please refer to Appendix C.

2.3 Finetuning

To be consistent with pretraining, we use the same data prepossessing pipeline during finetuning. For molecules, as multiple random conformations can be generated in a short time, we can use them as data augmentation in finetuning to improve performance and robustness. Some molecules may fail to generate 3D conformations, and we use their molecular graph as 2D conformation. For tasks that provide atom coordinates, we use them directly and skip the 3D conformation generation process. As there are 2 pretraining models and several types of downstream tasks, we should properly use them in the finetuning stage. According to the task types, and the involvement of protein or ligand, we can categorize them as follow.

Non-3D prediction tasks These tasks do not need to output 3D conformations. Examples include molecular property prediction, molecule similarity, pocket druggability prediction, protein-ligand binding affinity prediction, etc. Similar to NLP/CV, we can simply use the representation of `[CLS]` which represents the whole molecule/pocket, or the mean representation of all atoms, with a linear head to finetune on downstream tasks. In the tasks with pocket-molecule pair, we can concatenate their `[CLS]` representations, and then finetune with linear head.

3D prediction tasks of molecules or pockets These tasks need to predict a 3D conformation of the input, such as molecular conformation generation. Different with the fast conformation generation method used in Uni-Mol, molecular conformation generation task usually requires running advanced sampling and semi-empirical density functional theory (DFT) to account for the ensemble of 3D conformers that are accessible to a molecule, and this is very time-consuming. Therefore, there are many recent works that train the model to fast generate conformations from molecular graph [42, 43, 44, 45]. While in Uni-Mol, this task straightforwardly becomes a conformation optimization task: generate a new conformation based on a different input conformation. Specifically, in finetuning, the model supervised learns the mapping from Uni-Mol generated conformations to the labeled conformations. Moreover, the optimized conformations can be generated end-to-end by SE(3)-Equivariance coordinate head.

3D prediction tasks of protein-ligand pairs This is one of the most important tasks in structure-based drug design. The task is to predict the complex structure of a protein binding site and a molecular ligand. Besides the conformation changes of the pocket and the molecule themselves, we also need to consider how the molecule lays in the pocket, that is, the additional 6 degrees (3 rotations and 3 translations) of freedom of a rigid movement. In principle, with Uni-Mol, we can predict the complex conformation by the SE(3)-Equivariant coordinate head in an end-to-end fashion. However, this is unstable as it is very sensitive to the initial docking positions of molecular ligand. Herein, to get rid of the initial positions, we use a scoring function based optimization method in this paper. In particular, the molecular representation and pocket representation are firstly obtained from their own pretraining models by their own conformations; then, their representations are concatenated as the input of an additional 4-layer Uni-Mol encoder, which is finetuned to learn the pair distances of all atoms in molecule and pocket. With the predicted pair-distance matrix as the scoring function, we use a simple differential evolution algorithm [46] to sample and optimize the complex conformations. More details can be found in Appendix C.

3 Experiments

To verify the effectiveness of our proposed Uni-Mol model, we conduct extensive experiments on multiple downstream tasks, including molecular property prediction, molecular conformation generation, pocket property prediction, and protein-ligand binding pose prediction. Besides, we also conduct several ablation studies. Due to space restrictions, we leave the detailed experimental settings and ablation studies to Appendix C.

3.1 Molecular property prediction

Datasets and setup MoleculeNet [47] is a widely used benchmark for molecular property prediction, including datasets focusing on different levels of properties of molecules, from quantum

Table 1: Uni-Mol performance on molecular property prediction classification tasks

Classification (ROC-AUC %, higher is better ↑)									
Datasets	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	PCBA	MUV
# Molecules	2039	1513	1478	7831	8575	1427	41127	437929	93087
# Tasks	1	1	2	12	617	27	1	128	17
D-MPNN	71.0(0.3)	80.9(0.6)	90.6(0.6)	75.9(0.7)	65.5(0.3)	57.0(0.7)	77.1(0.5)	86.2(0.1)	78.6(1.4)
Attentive FP	64.3(1.8)	78.4(0.022)	84.7(0.3)	76.1(0.5)	63.7(0.2)	60.6(3.2)	75.7(1.4)	80.1(1.4)	76.6(1.5)
N-Gram$_{RF}$	69.7(0.6)	77.9(1.5)	77.5(4.0)	74.3(0.4)	-	66.8(0.7)	77.2(0.1)	-	76.9(0.7)
N-Gram$_{XGB}$	69.1(0.8)	79.1(1.3)	87.5(2.7)	75.8(0.9)	-	65.5(0.7)	78.7(0.4)	-	74.8(0.2)
PretrainGNN	68.7(1.3)	84.5(0.7)	72.6(1.5)	78.1(0.6)	65.7(0.6)	62.7(0.8)	79.9(0.7)	86.0(0.1)	81.3(2.1)
GROVER$_{base}$	70.0(0.1)	82.6(0.7)	81.2(3.0)	74.3(0.1)	65.4(0.4)	64.8(0.6)	62.5(0.9)	76.5(2.1)	67.3(1.8)
GROVER$_{large}$	69.5(0.1)	81.0(1.4)	76.2(3.7)	73.5(0.1)	65.3(0.5)	65.4(0.1)	68.2(1.1)	83.0(0.4)	67.3(1.8)
GraphMVP	72.4(1.6)	81.2(0.9)	79.1(2.8)	75.9(0.5)	63.1(0.4)	63.9(1.2)	77.0(1.2)	-	77.7(0.6)
MolCLR	72.2(2.1)	82.4(0.9)	91.2(3.5)	75.0(0.2)	-	58.9(1.4)	78.1(0.5)	-	79.6(1.9)
GEM	72.4(0.4)	85.6(1.1)	90.1(1.3)	78.1(0.1)	69.2(0.4)	**67.2(0.4)**	80.6(0.9)	86.6(0.1)	81.7(0.5)
Uni-Mol	**72.9(0.6)**	**85.7(0.2)**	**91.9(1.8)**	**79.6(0.5)**	**69.6(0.1)**	65.9(1.3)	**80.8(0.3)**	**88.5(0.1)**	**82.1(1.3)**

Table 2: Uni-Mol performance on molecular property prediction regression tasks

Regression (lower is better ↓)						
	RMSE			MAE		
Datasets	ESOL	FreeSolv	Lipo	QM7	QM8	QM9
# Molecules	1128	642	4200	6830	21786	133885
# Tasks	1	1	1	1	12	3
D-MPNN	1.050(0.008)	2.082(0.082)	0.683(0.016)	103.5(8.6)	0.0190(0.0001)	0.00814(0.00001)
Attentive FP	0.877(0.029)	2.073(0.183)	0.721(0.001)	72.0(2.7)	0.0179(0.001)	0.00812(0.00001)
N-Gram$_{RF}$	1.074(0.107)	2.688(0.085)	0.812(0.028)	92.8(4.0)	0.0236(0.0006)	0.01037(0.00016)
N-Gram$_{XGB}$	1.083(0.082)	5.061(0.744)	2.072(0.030)	81.9(1.9)	0.0215(0.0005)	0.00964(0.00031)
PretrainGNN	1.100(0.006)	2.764(0.002)	0.739(0.003)	113.2(0.6)	0.0200(0.0001)	0.00922(0.00004)
GROVER$_{base}$	0.983(0.090)	2.176(0.052)	0.817(0.008)	94.5(3.8)	0.0218(0.0004)	0.00984(0.00055)
GROVER$_{large}$	0.895(0.017)	2.272(0.051)	0.823(0.010)	92.0(0.9)	0.0224(0.0003)	0.00986(0.00025)
GraphMVP	1.029(0.033)	-	0.681(0.010)	-	-	-
MolCLR	1.271(0.040)	2.594(0.249)	0.691(0.004)	66.8(2.3)	0.0178(0.0003)	-
GEM	0.798(0.029)	1.877(0.094)	0.660(0.008)	58.9(0.8)	0.0171(0.0001)	0.00746(0.00001)
Uni-Mol	**0.788(0.029)**	**1.620(0.035)**	**0.603(0.010)**	**41.8(0.2)**	**0.0156(0.0001)**	**0.00467(0.00004)**

mechanics and physical chemistry to biophysics and physiology. Following previous work GEM [13], we use scaffold splitting for the dataset and report the mean and standard deviation of the results for three random seeds.

Baselines We compare Uni-Mol with multiple baselines, including supervised and pretraining baselines. D-MPNN [48] and AttentiveFP [49] are supervised GNNs methods. N-gram [50], PretrainGNN [22], GROVER [11], GraphMVP [25], MolCLR [12], and GEM [13] are pretraining methods. N-gram embeds the nodes in the graph and assembles them in short walks as the graph representation. Random Forest and XGBoost [51] are used as the predictor for downstream tasks.

Results Table 1 and Table 2 show the experiment results of Uni-Mol and competitive baselines, where the best results are marked in bold. Most baseline results are from the paper of GEM, except for the recent works GraphMVP and MolCLR. The results of GraphMVP are from its paper. As MolCLR uses a different data split setting (without considering chirality), we rerun it with the same data split setting as other baselines. From the results, we can summarize them as follows: 1) overall, Uni-Mol outperforms baselines on almost all downstream datasets. 2) In solubility (ESOL, Lipo), free energy (FreeSolv), and quantum mechanical (QM7, QM8, QM9) properties prediction tasks, Uni-Mol is significantly better than baselines. As 3D information is critical in these properties, it indicates that Uni-Mol can learn a better 3D representation than other baselines. 3) Uni-Mol fails to beat SOTA on the SIDER dataset. After investigation, we find Uni-Mol fails to generate 3D conformations (and rollbacks to 2D graphs) for many molecules (like natural products and peptides) in SIDER. Therefore, due to the missing 3D information, it is reasonable that Uni-Mol cannot outperform others.

In summary, by better utilizing 3D information in pretraining, Uni-Mol outperforms all previous MRL models in almost all property prediction tasks.

Table 3: Uni-Mol performance on molecular conformation generation

Dataset	QM9				Drugs			
Methods	COV(↑, %)		MAT(↓, Å)		COV(↑, %)		MAT(↓, Å)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
RDKit	83.26	90.78	0.3447	0.2935	60.91	65.70	1.2026	1.1252
CVGAE	0.09	0.00	1.6713	1.6088	0.00	0.00	3.0702	2.9937
GraphDG	73.33	84.21	0.4245	0.3973	8.27	0.00	1.9722	1.9845
CGCF	78.05	82.48	0.4219	0.3900	53.96	57.06	1.2487	1.2247
ConfVAE	80.42	85.31	0.4066	0.3891	53.14	53.98	1.2392	1.2447
ConfGF	88.49	94.13	0.2673	0.2685	62.15	70.93	1.1629	1.1596
GeoMol	71.26	72.00	0.3731	0.3731	67.16	71.71	1.0875	1.0586
DGSM	91.49	95.92	0.2139	0.2137	78.73	94.39	1.0154	0.9980
DMCG	96.34	99.53	0.2065	0.2003	**96.69**	100.00	0.7223	0.7236
GeoDiff	91.68	95.82	0.2099	0.2026	89.13	97.88	0.8629	0.8529
Uni-Mol	**98.68**	**100.00**	**0.1806**	**0.1510**	92.69	**100.00**	**0.6596**	**0.6215**

3.2 Molecular conformation generation

Datasets and setup Following the settings in previous works [43, 52], we use GEOM-QM9 and GEOM-Drugs [53] dataset to perform conformation generation experiments. As described in Sec. 2.3, in this task, Uni-Mol optimizes its generative conformations to the labeled ones. To construct the finetuning data, we first randomly generate 10 conformations. Then, for each of them, we calculate the RMSD between it and labeled conformations, and choose the one with minimal RMSD as its optimizing target. For the inference in the test set, we generate the same number of conformations (twice the number of labeled conformations) as previous works do. And we use the same metrics, Coverage (COV) and Matching (MAT). Higher COV means better diversity, while lower MAT means higher accuracy.

Baselines We compare Uni-Mol with 10 competitive baselines. RDKit [37] is a traditional conformation generation method based on distance geometry. The rest baseline can be categorized into two classes. GraphDG [42], CGCF[43], ConfVAE [54], ConfGF [52], and DGSM [55] combine generative models with distance geometry, which first generates interatomic distance matrices and then iteratively generates atomic coordinates. CVGAE [44], GeoMol [45], DMCG [56], and GeoDiff [57] directly generate atomic coordinates.

Results The results are shown in Table 3. We report the mean and median of COV and MAT on GEOM-QM9 and GEOM-Drugs datasets. ConfVAE [54], GeoMol[45], DGSM [55], DMCG [56], GeoDiff's [57] results are from their papers, respectively. Other baseline results are from ConfGF's paper. As shown in Table 3, Uni-Mol exceeds existing baselines in both COV and MAT metrics on both datasets. Although Uni-Mol outperforms SOTA, we suspect that the above benchmark cannot satisfy the real-world demand of conformation generation tasks in the field of drug design. Since the ensemble of molecular conformations in biological systems is different from that in a vacuum or general solution environment, the ensemble of bioactive conformation must be considered in order to apply the conformation generation model in the context of drug design, while the GEOM dataset just ignores this. Establishing a reasonable benchmark will be crucial in this research direction.

3.3 Pocket property prediction

Datasets and setup Druggability, the ability of a candidate protein pocket to produce stable binding to a specific molecular ligand, is one of the most critical properties of a candidate protein pocket. However, this task is very challenging due to the very limited supervised data. For example, NRDLD [58], a commonly used dataset, only contains 113 data samples. Therefore, besides NRDLD, we construct a regression dataset for benchmarking pocket property prediction performance. Specifically, based on Fpocket tool, we calculate Fpocket Score, Druggability Score, Total SASA, and Hydrophobicity Score for the selected 164,586 candidate pockets. Model is trained to predict these scores. To avoid leaking, the selected pockets are not overlapped with the candidate protein pocket dataset used in Uni-Mol pretraining.

Baselines On the NRDLD dataset, we compare Uni-Mol with 6 previous methods evaluated in [59]. Accuracy, recall, precision, and F1-score are used as metrics for this classification task. On our created benchmark dataset, as there are no appropriate baselines, we use an additional Uni-Mol model

Table 4: Uni-Mol performance on pocket property prediction

Dataset	Classification (higher is better ↑) NRDLD						Regression (lower is better ↓) Fpocket Scores		
Methods	Cavity-DrugScore	Volsite	DrugPred	PockDrug	TRAPP-CNN	**Uni-Mol**	Methods	Uni-Mol$_{\text{random}}$	**Uni-Mol**
Accuracy	0.82	0.89	0.89	0.865	**0.946**	**0.946**	MSE$_{\text{Fpocket}}$	0.621(0.004)	**0.551(0.008)**
Recall	-	-	-	0.957	0.913	**1.000**	MSE$_{\text{Druggability}}$	0.601(0.02)	**0.499(0.007)**
Precision	-	-	-	0.846	**1.000**	0.920	MSE$_{\text{Total SASA}}$	0.197(0.008)	**0.129(0.005)**
F1-score	-	-	-	0.898	0.955	**0.958**	MSE$_{\text{Hydrophobicity}}$	0.0357(0.017)	**0.0127(0.0005)**

without pretraining, denoted as Uni-Mol$_{\text{random}}$, to check the performance brought by pretraining on pocket property prediction. MSE (mean square error) is used as the metric.

Results As shown in Table 4, Uni-Mol shows the best accuracy, recall, and F1-score on NRDLD, the few-show dataset. In our created benchmark dataset, the pretraining Uni-Mol model largely outperforms the non-pretraining one on all four scores. This indicates that pretraining on candidate protein pockets indeed brings improvement in pocket property prediction tasks.

Unlike Molecular property prediction, due to the very limited supervised data, pocket property prediction gained much less attention. Therefore, we also plan to release our created benchmark dataset, and hopefully, it can help future research.

3.4 Protein-ligand binding pose prediction

Datasets and setup As mentioned above, protein-ligand binding pose prediction is one of the most important tasks in drug design. And Uni-Mol combines both the molecular and pocket pretraining models to learn a distance matrix based scoring function, and then sample and optimize the complex conformations. For the benchmark dataset, referring to the previous works [27, 60], we use CASF-2016 as the test set. For the training data used in finetuning, we use PDBbind General set v.2020 [61] (19,443 protein-ligand complexes), excluding complexes that already exist in the CASF-2016.

Two benchmarks are conducted: 1) Docking power, the default metric to benchmark the ability of a scoring function in CASF-2016. Specifically, it tests whether a scoring function can distinguish the ground truth binding pose from a set of decoys or not. For each ground truth, CASF-2016 provides 50 100 decoy conformations of the same ligand. Scoring functions are applied to rank them, and the ground truth binding pose is expected to be the top 1. 2) Binding pose accuracy. Specifically, we use the semi-flexible docking setting: keep the pocket conformation fixed, while the conformation of the ligand is fully flexible. We evaluate the RMSD between the predicted binding pose and the ground truth. Following previous works, we use the percentage of results that are below predefined RMSD thresholds as metrics.

Baselines For docking power benchmark, the baselines are DeepDock [60] and the top 10 scoring functions reported in [27], including both conventional scoring functions and machine learning-based ones. For the binding pose accuracy, the baselines are Autodock Vina [62, 63], Vinardo [64], Smina [65], and AutoDock4 [66].

Results From the docking power benchmark results shown in Figure 3, Uni-Mol ranks the 1st, with the top 1 success rate of 91.6%. For comparison, the previous top scoring function AutoDock Vina [62, 63] achieves 90.2% of the top 1 success rate in this benchmark. From the binding pose accuracy results shown in Table 5, Uni-Mol also surpasses all other baselines. Notably, Uni-Mol outperforms the second best method by 22.81% under the threshold of 2Å. This result indicates that Uni-Mol can effectively learn the 3D information from both molecules and pockets, as well as the interaction in 3D space of them. Even without pretraining, Uni-Mol (denoted as Uni-Mol$_{\text{random}}$) is also better than other baselines. This demonstrates the effectiveness of Uni-Mol backbone, as it effectively learns the 3D information by limited data.

In summary, by combining molecular and pocket pretraining models, Uni-Mol significantly outperforms the widely used docking tools in the protein-ligand binding tasks.

4 Related work

Molecular representation learning Representation learning on large-scale unlabeled molecules attracts much attention recently. SMILES-BERT [18] is pretrained on SMILES strings of molecules using BERT [4]. Subsequent works are mostly pretraining on 2D molecular topological graphs [23, 11]. MolCLR [12] applies data augmentation to molecular graphs at both node and graph levels, using

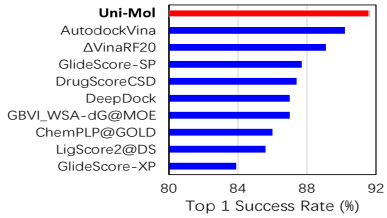

Figure 3: Docking power evaluation on CASF-2016 (Top 10 methods)

	Ligand RMSD % Below Threshold ↑					
Methods	0.5 Å	1.0 Å	1.5 Å	2.0 Å	3.0 Å	5.0 Å
Autodock Vina	23.86	44.21	57.54	64.56	73.68	84.56
Vinardo	23.51	41.75	57.54	62.81	69.82	76.84
Smina	23.51	47.37	59.65	65.26	74.39	82.11
Autodock4	7.02	21.75	31.58	35.44	47.02	64.56
Uni-Mol$_{\text{random}}$	14.04	49.47	65.26	75.44	87.02	98.60
Uni-Mol	**24.91**	**70.53**	**84.21**	**88.07**	**94.74**	**98.95**

Table 5: Uni-Mol performance on binding pose prediction

a self-supervised contrastive learning strategy to learn molecular representations. Further, several recent works try to leverage the 3D spatial information of molecules, and focus on contrastive or transfer learning between 2D topology and 3D geometry of molecules. For example, GraphMVP [25] proposes a contrastive learning GNN-based framework between 2D topology and 3D geometry. GEM [13] uses bond angles and bond length as additional edge attributes to enhance 3D information. As aforementioned, due to the inability of handling 3D information, most previous representation learning models cannot be used in the important 3D prediction tasks.

SE(3)-Equivariant models In many-body scenarios such as potential energy surface fitting, SE-(3) equivariance is usually required. A series of SE(3) models are proposed, such as SchNet [67], tensor field networks [68], SE(3) Transformer [69], DimmNet [70], equivariant graph neural networks (EGNN) [35], and GemNet [71]. Most of these models are used in supervised learning with energy and force. In Uni-Mol, based on the standard Transformer, we introduce several minor changes to make the model SE(3)-Equivariant.

Pocket druggability prediction Druggability prediction of protein binding pockets is crucial for drug discovery as druggable pockets need to be identified at the beginning. Since proteins undergo conformation changes that might alter the druggability of pockets, it is necessary to utilize 3D spatial data beyond sequential information. Early methods, such as Volsite [72], DrugPred [58], and PockDrug [73], predict druggability based on the predefined descriptors of pockets' static structures. Later, TRAPP-CNN [59], based on 3D-CNN, proposes the analysis of proteins' conformation changes and the use of such information for druggability prediction.

Protein-ligand binding pose prediction In structure-based drug design, it is crucial to understand the interactions between protein targets and ligands. The *in vitro* estimation of the binding pose and affinity, such as docking, allows for lead identification and guides molecular optimization. In particular, docking is one of the most important approaches in structure-based drug design and has been developed for the past decades. Tools such as AutoDock4 [66], AutoDock Vina [62, 63], and Smina [65] are among the most used docking programs. Also, machine learning-based docking methods, such as $\Delta_{Vina}\text{RF}_{20}$ [74] and DeepDock [60], have also been developed to predict protein-ligand binding poses and assess protein-ligand binding affinity.

5 Conclusion

In this paper, to enlarge the application scope and representation ability of molecular representation learning (MRL), we propose Uni-Mol, the first universal large-scale 3D MRL framework. Uni-Mol consists of 3 parts: a Transformer based backbone to handle 3D data; two large-scale pretraining models to learn molecular and pocket representations respectively; finetuning strategies for all kinds of downstream tasks. Experiments demonstrate that Uni-Mol can outperform existing SOTA in various downstream tasks, especially in 3D spatial tasks.

There are 3 potential future directions. 1) Better interaction mechanisms for finetuning two pretraining models together. As the interaction between the pretraining pocket model and the pretraining molecular model is simple in the current version of Uni-Mol, we believe there is a large room for further improvement. 2) Large Uni-Mol models. As larger pretraining models often perform better, it is worthy of training a large Uni-Mol model on a bigger dataset. 3) More high-quality benchmarks. Although there have been many applications in the field of drug design, high-quality public datasets have been lacking. Many public datasets cannot satisfy real-world demand due to the low data quality. We believe the high-quality benchmarks will be the lighthouse of the entire field, and will significantly accelerate the development of drug design.

Acknowledgments and Disclosure of Funding

We thank Shuqi Lu, Yuanqi Du, Zhen Wang, Yingze Wang, Xi Chen, Zhengdan Zhu and many colleagues in DP Technology for their great help in this project.

3D structures in Fig. 1 are drawn using the web service Hermite™(https://hermite.dp.tech).

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[2] William L. Hamilton, Rex Ying, and Jure Leskovec. "Representation Learning on Graphs: Methods and Applications". In: *IEEE Data Eng. Bull.* 40.3 (2017), pp. 52–74. URL: http://sites.computer.org/debull/A17sept/p52.pdf.

[3] Daokun Zhang et al. "Network representation learning: A survey". In: *IEEE transactions on Big Data* 6.1 (2018), pp. 3–28.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[5] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[6] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[7] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[8] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[9] Qingda Zang et al. "In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning". In: *Journal of chemical information and modeling* 57.1 (2017), pp. 36–49.

[10] Minjian Yang et al. "Machine learning models based on molecular fingerprints and an extreme gradient boosting method lead to the discovery of JAK2 inhibitors". In: *Journal of Chemical Information and Modeling* 59.12 (2019), pp. 5002–5012.

[11] Yu Rong et al. "Self-Supervised Graph Transformer on Large-Scale Molecular Data". In: *Advances in Neural Information Processing Systems* 33 (2020).

[12] Yuyang Wang et al. "Molecular contrastive learning of representations via graph neural networks". In: *Nature Machine Intelligence* (2022), pp. 1–9. DOI: 10.1038/s42256-022-00447-x.

[13] Xiaomin Fang et al. "Geometry-enhanced molecular representation learning for property prediction". In: *Nature Machine Intelligence* (2022), pp. 1–8. DOI: 10.1038/s42256-021-00438-4.

[14] A Crum-Brown and TR Fraser. "The connection of chemical constitution and physiological action". In: *Trans R Soc Edinb* 25.1968-1969 (1865), p. 257.

[15] Corwin Hansch and Toshio Fujita. "p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". In: *Journal of the American Chemical Society* 86.8 (1964), pp. 1616–1626.

[16] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[17] Zheng Xu et al. "Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery". In: *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*. 2017, pp. 285–294.

[18] Sheng Wang et al. “Smiles-bert: large scale unsupervised pre-training for molecular property prediction”. In: *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*. 2019, pp. 429–436.

[19] Stephen R Heller et al. “InChI, the IUPAC international chemical identifier”. In: *Journal of cheminformatics* 7.1 (2015), pp. 1–34.

[20] Robin Winter et al. “Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations”. In: *Chemical science* 10.6 (2019), pp. 1692–1701.

[21] Jennifer Handsel et al. “Translating the InChI: adapting neural machine translation to predict IUPAC names from a chemical identifier”. In: *Journal of cheminformatics* 13.1 (2021), pp. 1–11.

[22] Weihua Hu* et al. “Strategies for Pre-training Graph Neural Networks”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=HJlWWJSFDH.

[23] Pengyong Li et al. “An effective self-supervised framework for learning expressive molecular global representations to drug discovery”. In: *Briefings in Bioinformatics* 22.6 (2021), bbab109.

[24] Panagiotis I Koukos, Li C Xue, and Alexandre MJJ Bonvin. “Protein–ligand pose and affinity prediction: Lessons from D3R Grand Challenge 3”. In: *Journal of computer-aided molecular design* 33.1 (2019), pp. 83–91.

[25] Shengchao Liu et al. “Pre-training Molecular Graph Representation with 3D Geometry”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=xQUe1pOKPam.

[26] Hannes Stärk et al. “3D Infomax improves GNNs for Molecular Property Prediction”. In: *arXiv preprint arXiv:2110.04126* (2021).

[27] Minyi Su et al. “Comparative assessment of scoring functions: the CASF-2016 update”. In: *Journal of chemical information and modeling* 59.2 (2018), pp. 895–913.

[28] Andrew L Hopkins, Colin R Groom, and Alexander Alex. “Ligand efficiency: a useful metric for lead selection.” In: *Drug discovery today* 9.10 (2004), pp. 430–431.

[29] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[30] Ruibin Xiong et al. “On Layer Normalization in the Transformer Architecture”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 10524–10533.

[31] Guolin Ke, Di He, and Tie-Yan Liu. “Rethinking Positional Encoding in Language Pre-training”. In: *International Conference on Learning Representations*. 2020.

[32] Philipp Dufter, Martin Schmitt, and Hinrich Schütze. “Position information in transformers: An overview”. In: *arXiv preprint arXiv:2102.11090* (2021).

[33] Muhammed Shuaibi et al. “Rotation invariant graph neural networks using spin convolutions”. In: *arXiv preprint arXiv:2106.09575* (2021).

[34] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[35] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. “E (n) equivariant graph neural networks”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9323–9332.

[36] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).

[37] Sereina Riniker and Gregory A Landrum. “Better informed distance geometry: using what we know to improve conformation generation”. In: *Journal of chemical information and modeling* 55.12 (2015), pp. 2562–2574.

[38] Thomas A Halgren. “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94”. In: *Journal of computational chemistry* 17.5-6 (1996), pp. 490–519.

[39] Greg Landrum et al. *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling*. 2013.

[40] Helen M Berman et al. “The protein data bank”. In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[41] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. "Fpocket: an open source platform for ligand pocket detection". In: *BMC bioinformatics* 10.1 (2009), pp. 1–11.

[42] Gregor Simm and Jose Miguel Hernandez-Lobato. "A Generative Model for Molecular Distance Geometry". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8949–8958.

[43] Minkai Xu et al. "Learning Neural Generative Dynamics for Molecular Conformation Generation". In: *International Conference on Learning Representations*. 2020.

[44] Elman Mansimov et al. "Molecular geometry prediction using a deep generative graph neural network". In: *Scientific reports* 9.1 (2019), pp. 1–13.

[45] Octavian Ganea et al. "Geomol: Torsional geometric generation of molecular 3d conformer ensembles". In: *Advances in Neural Information Processing Systems* 34 (2021).

[46] Rainer Storn and Kenneth Price. "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces". In: *Journal of global optimization* 11.4 (1997), pp. 341–359.

[47] Zhenqin Wu et al. "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical science* 9.2 (2018), pp. 513–530.

[48] Kevin Yang et al. "Analyzing learned molecular representations for property prediction". In: *Journal of chemical information and modeling* 59.8 (2019), pp. 3370–3388.

[49] Zhaoping Xiong et al. "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism". In: *Journal of medicinal chemistry* 63.16 (2019), pp. 8749–8760.

[50] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. "N-gram graph: Simple unsupervised representation for graphs, with applications to molecules". In: *Advances in neural information processing systems* 32 (2019).

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

[52] Chence Shi et al. "Learning gradient fields for molecular conformation generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9558–9568.

[53] Simon Axelrod and Rafael Gomez-Bombarelli. "GEOM, energy-annotated molecular conformations for property prediction and molecular generation". In: *Scientific Data* 9.1 (2022), pp. 1–14.

[54] Minkai Xu et al. "An end-to-end framework for molecular conformation generation via bilevel programming". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 11537–11547.

[55] Shitong Luo et al. "Predicting Molecular Conformation via Dynamic Graph Score Matching". In: *Advances in Neural Information Processing Systems* 34 (2021).

[56] Jinhua Zhu et al. "Direct molecular conformation generation". In: *arXiv preprint arXiv:2202.01356* (2022).

[57] Minkai Xu et al. "GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation". In: *International Conference on Learning Representations*. 2022.

[58] Agata Krasowski et al. "DrugPred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set". In: *Journal of chemical information and modeling* 51.11 (2011), pp. 2829–2842.

[59] Jui-Hung Yuan et al. "Druggability assessment in TRAPP using machine learning approaches". In: *Journal of Chemical Information and Modeling* 60.3 (2020), pp. 1685–1699.

[60] Oscar Méndez-Lucio et al. "A geometric deep learning approach to predict binding conformations of bioactive molecules". In: *Nature Machine Intelligence* 3.12 (2021), pp. 1033–1039.

[61] Zhihai Liu et al. "PDB-wide collection of binding data: current status of the PDBbind database". In: *Bioinformatics* 31.3 (2015), pp. 405–412.

[62] Oleg Trott and Arthur J Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of computational chemistry* 31.2 (2010), pp. 455–461.

[63] Jerome Eberhardt et al. "AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings". In: *Journal of Chemical Information and Modeling* 61.8 (2021), pp. 3891–3898.

[64] Rodrigo Quiroga and Marcos A Villarreal. "Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening". In: *PloS one* 11.5 (2016), e0155183.

[65] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. "Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise". In: *Journal of chemical information and modeling* 53.8 (2013), pp. 1893–1904.

[66] Garrett M Morris et al. "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility". In: *Journal of computational chemistry* 30.16 (2009), pp. 2785–2791.

[67] Kristof Schütt et al. "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions". In: *Advances in neural information processing systems* 30 (2017).

[68] Nathaniel Thomas et al. "Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds". In: *arXiv preprint arXiv:1802.08219* (2018).

[69] Fabian Fuchs et al. "Se (3)-transformers: 3d roto-translation equivariant attention networks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1970–1981.

[70] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. "Directional Message Passing for Molecular Graphs". In: *International Conference on Learning Representations (ICLR)*. 2020.

[71] Johannes Klicpera, Florian Becker, and Stephan Günnemann. "GemNet: Universal Directional Graph Neural Networks for Molecules". In: *Advances in Neural Information Processing Systems*. 2021.

[72] Jérémy Desaphy et al. *Comparison and druggability prediction of protein–ligand binding sites from pharmacophore-annotated cavity shapes*. 2012.

[73] Alexandre Borrel et al. "PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties". In: *Journal of chemical information and modeling* 55.4 (2015), pp. 882–895.

[74] Cheng Wang and Yingkai Zhang. "Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest". In: *Journal of computational chemistry* 38.3 (2017), pp. 169–177.

[75] Teague Sterling and John J Irwin. "ZINC 15–ligand discovery for everyone". In: *Journal of chemical information and modeling* 55.11 (2015), pp. 2324–2337.

[76] Anna Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery". In: *Nucleic acids research* 40.D1 (2012), pp. D1100–D1107.

[77] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. "A data-driven approach to predicting successes and failures of clinical trials". In: *Cell chemical biology* 23.10 (2016), pp. 1294–1301.

[78] Matthias Rupp et al. "Fast and accurate modeling of molecular atomization energies with machine learning". In: *Physical review letters* 108.5 (2012), p. 058301.

[79] Bharath Ramsundar et al. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media, 2019.

[80] Paul CD Hawkins. "Conformation generation: the state of the art". In: *Journal of Chemical Information and Modeling* 57.8 (2017), pp. 1747–1756.

[81] Ziyao Li et al. "Conformation-guided molecular representation with hamiltonian neural networks". In: *International Conference on Learning Representations*. 2020.

A Pretraining data

Molecular dataset The pretraining datasets we use consist of two parts: one part is a database collection of 12 million molecules that can be synthesized and purchased (See Table 6), and the other part is taken from a previous work [23], whose molecules are collected from the ZINC [75] and ChemBL [76] databases. After normalizing and duplicating, we obtain 19 million molecules as our pretraining dataset. For each molecule, we add random conformer augmentations with ten 3D conformers generated by RDKit and one 2D graph to avoid ETKDG patterns missing match.

Candidate protein pocket dataset The pretraining dataset for candidate protein pockets is derived from the Protein Data Bank (RCSB PDB [4]) [40], a collection of 180K structural data of proteins. We first pre-process the raw data by adding missing side chains and hydrogen atoms, and then we use Fpocket [41] to detect candidate binding pockets of the proteins. After filtering the raw pockets by the number of residues they have contact with (10~25) and including water molecules inside the pockets, we collect a pretraining dataset of 3,291,739 candidate pockets.

B Downstream data supplements

Molecular property prediction We conduct experiments on the MoleculeNet[47] benchmark in the molecular property prediction task. MoleculeNet is a widely used benchmark for molecular property prediction. The details of the 15 datasets we used are described below.

- **BBBP** Blood-brain barrier penetration (BBBP) contains the ability of small molecules to penetrate the blood-brain barrier.
- **BACE** This dataset contains the results of small molecules as inhibitors of binding to human β-secretase 1 (BACE-1).
- **ClinTox** This dataset contains the toxicity of the drug in clinical trials and the status of the drug for FDA approval[77].
- **Tox21** The dataset contains toxicity measurements of 8k molecules for 12 targets.
- **ToxCast** This dataset is derived from toxicology data from in vitro high-throughput screening and contains toxicity measurements for 8k molecules against 617 targets.
- **SIDER** The Side Effect Resource (SIDER) contains side effects of drugs on 27 system organs. These drugs are not only small molecules but also some peptides with molecular weights over 1000.
- **HIV** This dataset contains 40k compounds with the ability to inhibit HIV replication.
- **PCBA** PubChem BioAssay (PCBA) is a database of small molecule bioactivities generated by high-throughput screening. This is a subset containing over 400k molecules on 128 bioassays.
- **MUV** Maximum Unbiased Validation (MUV) is another subset of PubChem BioAssay, containing 90k molecules and 17 bioassays.
- **ESOL** This dataset contains the water solubility of the compound and is a small dataset with 1128 molecules.
- **FreeSolv** The dataset contains hydration free energy data for small molecules, of which we use the experimental values as labels.
- **Lipo** Lipophilicity contains the solubility of small molecules in lipids, of which we use the octanol/water distribution coefficient as the label.
- **QM7, QM8, QM9** The molecule in QM7 contains up to 7 heavy atoms, QM8 is 8 and QM9 is 9. These datasets provide the geometric, energetic, electronic and thermodynamic properties of the molecule, which are calculated by density functional theory (DFT)[78]. QM9 contains several quantum mechanical properties of different quantitative ranges, and we select *homo*, *lumo* and *gap* of similar quantitative range, following the setup of the previous work[13].

[4] http://www.rcsb.org/

Table 6: Database collection of 12M purchasable molecules

Database	Molecules	Link
Targetmol	10,000	https://www.targetmol.com/
Chemdiv	1,613,931	https://www.chemdiv.com/
Enamine	2,734,581	https://enamine.net/
Chembridge	1,557,942	https://www.chembridge.com/
Life Chemical	509,975	https://lifechemicals.com/
Specs	208,670	https://www.specs.net/
Vitas-M	1,409,339	https://vitasmlab.biz/
InterBioScreen	48,627	https://www.ibscreen.com/
Maybridge	53,352	https://www.thermofisher.in/
Bionet-Key Organics	259,244	https://www.keyorganics.net/
Asinex	530,881	https://www.asinex.com/
UkrOrgSynthesis	688,952	https://uorsy.com/
Eximed	61,009	https://eximedlab.com/
HTS Biochemie Innovationen	58,437	https://www.hts-biochemie.de/
Princeton BioMolecular	1,532,542	https://princetonbio.com/
Otava	270,835	https://otavachemicals.com/
Alinda Chemical	202,332	https://www.alinda.ru/
Analyticon	42,664	https://www.analyticon-diagnostics.com/

Molecular corformation generation Following the settings in previous works [43, 52], we use GEOM-QM9 and GEOM-Drugs [53] dataset in this task.

- **GEOM** This dataset contains 37 million accurate conformations generated for 450,000 molecules by advanced sampling and semi-empirical density flooding theory (DFT). Of these, 133,000 molecules are from QM9, and the remaining 317,000 molecules have biophysical, physiological, or physical chemistry experimental data, i.e., Drugs.

Pocket property prediction NRDLD [58] is a benchmark dataset for pocket druggability prediction. As NRDLD and other existing benchmark datasets are too small, we construct a regression dataset to benchmark pocket property prediction performance.

- **NRDLD** NRDLD contains 113 proteins, and a predefined split is provided: 76 proteins constitute the training set and 37 proteins constitute the test set. It labels 71 proteins as druggable in that they noncovalently bind small drug-like ligands [59]. The rest 42 proteins are labeled as less-druggable because none of the ligands they cocrystallized satisfy the following requirements simultaneously: the rule of five, clogP $\geq$ -2, and ligand efficiency, as defined in [28], $\geq$ 0.3 kcal mol^{-1} / heavy atom.
- **Our created benchmark dataset** The dataset contains 164,586 candidate pockets, and Fpocket scores each one of them on Fpocket Score, Druggability Score, Total SASA, and Hydrophobicity Score. These four scores are indicators of the druggability of candidate pockets. To avoid leaking, the selected pockets are not overlapped with the candidate protein pocket dataset used in Uni-Mol pretraining.

Protein-ligand binding pose prediction We use PDBbind General set v.2020 [61], excluding the complexes in CASF-2016 [27], as the training set. And CASF-2016 is used as the test set. In particular, we define the pocket for each protein-ligand pair as residues of the protein which have at least one atom within the range of 6Å from a heavy atom in the ligand. All atoms of the selected residues are included. In addition, we draw the smallest bounding box covering all of the atoms in the pocket and regard the water molecules in the bounding box as a part of the pockets, too.

- **PDBbind General set v.2020** This dataset contains 19,443 protein-ligand complexes with binding data and processed structural files originally from the Protein Data Bank (PDB). Only complexes with experimentally determined binding affinity data are included in the general set.
- **CASF-2016** CASF-2016 is the widely used benchmark for docking and scoring. This dataset, whose primary test set is known as the PDBbind Core set, contains 285 protein-ligand complexes with high quality crystal structures and reliable binding constants from PDBbind General set. For

Table 7: Uni-Mol hyperparameters setup during pre-training

Hyperparameter	Molecular pretraining	Pocket pretraining
Layers	15	15
Peak learning rate	1e-4	1e-4
Batch size	128	128
Max training steps	1M	1M
Warmup steps	10K	10k
Attention heads	64	64
FFN dropout	0.1	0.1
Attention dropout	0.1	0.1
Embedding dropout	0.1	0.1
Weight decay	1e-4	1e-4
Embedding dim	512	512
FFN hidden dim	2048	2048
Gaussian kernel channels	128	128
Mask ratio	0.15	0.15
Coordinate noise	Uniform [-1 Å, 1 Å]	Uniform [-1 Å, 1 Å]
Activation function	GELU	GELU
Learning rate decay	Linear	Linear
Adams ϵ	1e-6	1e-6
Adams (β_1, β_2)	(0.9, 0.99)	(0.9, 0.99)
Gradient clip norm	1.0	1.0
Atom loss function and its weight	Cross entropy, 1.0	Cross entropy, 1.0
Coordinate loss function and its weight	Smooth L1, 5.0	Smooth L1, 1.0
Distance loss function and its weight	Smooth L1, 10.0	Smooth L1, 1.0
Max number of atoms	256	256
Vocabulary size (atom types)	30	9

each protein-ligand complex, CASF-2016 provides 50~100 decoy molecular conformations of the same ligand for evaluation.

C Experiments details & reproduce

Molecular Pretraining setup We report the detailed hyperparameters setup of Uni-mol during pretraining in Table 7. Uni-Mol training loss is summed up by three components, atom(token) loss, coordinate loss, and pair-distance loss. Atoms are masked, and noise is added to coordinate as described in sections 2.1 and 2.2. Since the values of the above three components differ significantly, to make them have a similar influence, we enlarge the coordinate loss and distance loss.

Pocket Pretraining setup The pocket Uni-Mol model is slightly different from molecule ones during pretraining: 1) We use a residue-level masking strategy instead of the original atom-level, as residue granularity is non-redundancy and integrity in protein. 2) Only polar hydrogen is remained in pocket Uni-Mol pretraining, to reduce the number of used atoms and thus improve efficiency. 3) All weights of loss functions are set 1, as the residue-level masking strategy makes the 3D denoising task much harder. Other settings are listed in Table 7.

Molecular property prediction

- **Data split** In our experiments, referring to previous work GEM[13], we use scaffold splitting[79] to divide the dataset into training, validation, and test sets in the ratio of 8:1:1. Scaffold splitting is more challenging than random splitting as the scaffold sets of molecules in different subsets do not intersect. This splitting tests the model's generalization ability and reflects the realistic cases[47]. Since this splitting is according to the scaffold of the molecule, we find that whether or not chirality is considered when generating the scaffold using RDKit has a significant impact on the division results. From the results, the splitting considering chirality makes the task harder. The original implementation of MolCLR does not consider chirality, and we reproduce the experiment by considering it. In all experiments, we choose the checkpoint with the best validation loss, and report the results on the test-set run by that checkpoint.

Table 8: Search space for small datasets: BBBP, BACE, ClinTox, Tox21, Toxcast, SIDER, ESOL, FreeSolv, Lipo, QM7, QM8, for large datasets: PCBA, MUV, QM9, and for HIV

Hyperparameter	Small	Large	HIV
Learning rate	[5e-5, 1e-4, 4e-4, 5e-4]	[2e-5, 1e-4]	[2e-5, 5e-5]
Batch size	[32, 64, 128, 256]	[128, 256]	[128, 256]
Epochs	[40 ,60, 80, 100]	[20, 40]	[2, 5, 10]
Pooler dropout	[0.0, 0.1, 0.2, 0.5]	[0.0, 0.1]	[0.0, 0.2]
Warmup ratio	[0.0, 0.06, 0.1]	[0.0, 0.06]	[0.0, 0.1]

Table 9: Hyperparameters setup for molecular conformation generation

Learning rate	1e-4
Batch size	8
Epochs	5
Warmup ratio	0.06
Coordinate loss function and weight	MSE, 1.0
Distance loss function and weight	MSE, 1.0

- **Hyperparameter search space** Referring to previous works, we use a grid search to find the best combination of hyperparameters for the molecular property prediction task. To reduce the time cost, we set a smaller search space for the large datasets. The specific search space is shown in Table 8.

Molecular conformation generation We report the detailed hyperparameters setup for molecular conformation generation in Table 9. Since this is a 3D-related task, we only use coordinate loss and distance loss.

Pocket property prediction The hyperparameters we search are listed in Table 10.

Protein-ligand binding pose prediction

- **Data split** The training set is PDBbind General set v.2020 excluding the complexes covered CASF-2016. We perform data preprocessing, such as adding missing atoms to both proteins and ligands and manually fixing file-loading errors, before constructing the training set. And we additionally filter the complexes based on the number of residues contained in the pockets (>= 5), resulting in a training set of 18k protein-ligand complexes. The test set is CASF-2016, which contains 285 protein-ligand complexes.
- **Binding pose model architecture** As shown in Figure 4, the binding pose model is an encoder-decoder architecture consisting of two 15 layers Uni-Mol as encoder and a 4 layers Uni-Mol as decoder. The decoder Uni-Mol block follows the same setting as the pretraining ones.
- **Scoring function** To evaluate the docking power of our proposed Uni-Mol model, we construct a scoring function, composed of cross distance loss and self-distance loss, out of Uni-Mol. Cross distance loss evaluates the atom-wise distance between atoms on the pocket and ligand, and self-distance evaluates the atom-wise distance between atoms on the same ligand. The ultimate scoring function is a weighted sum of the cross distance loss and the self-distance loss, and the weights are 1.0 and 5.75 respectively.

Table 10: Search space for pocket property prediction

Hyperparameter	NRDLD	Fpocket Scores
Learning rate	[5e-5, 1e-4, 3e-4]	3e-4
Batch size	[1, 2, 4, 8, 16]	32
Epochs	40	20
Pooler dropout	[0, 0.1, 0.2, 0.3]	0
Warmup ratio	[0.0, 0.1]	0.1

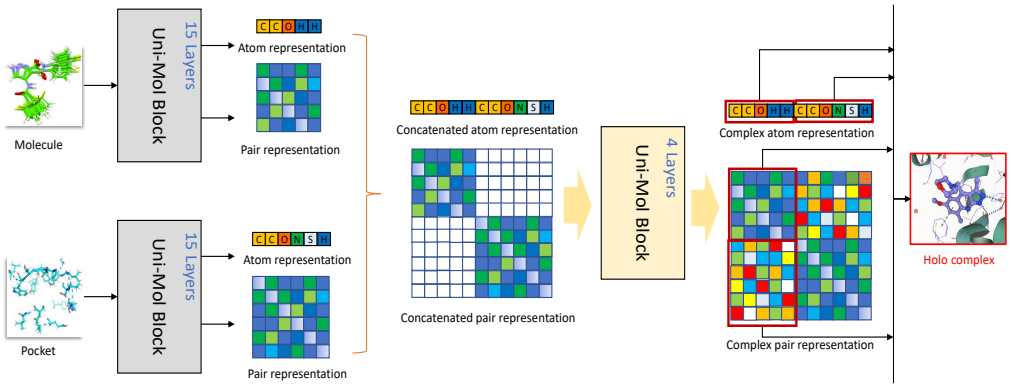

Figure 4: protein-ligand binding pose model: 1) Encoder: molecular representation and pocket representation are obtained from their own pretraining Uni-Mol models; 2) Decoder: representation is concatenated with atom and pair-level, as inputs of a 4 layers Uni-Mol block learning from scratch. 3) Output: The complex representation is used as a project layer to learn the pair distances of molecule and pocket.

Table 11: Hyperparameters setup for binding pose prediction

Hyperparameters for finetuning	Value
Learning rate	3e-4
Batch size	32
Epochs	50
Warmup ratio	0.06
Dropout	0.2
Dist_threshold	8.0
Cross distance loss function and weight	MSE, 1.0
Holo distance loss function and weight	MSE, 1.0
Hyperparameters for sampling	**Value**
Population size	150
Max iterations	500
Dist_threshold	5.0
Mutation	(0.5, 1.0)
Recombination	0.9
Conformation size	10
Cross distance weight	1.0
Holo distance weight	5.75

- **Hyperparameter settings**

  As shown in Figure 4, Uni-Mol directly predicts protein-ligand cross distance and self-distance with MSE loss during finetuning. Dist_threshold is used to mask distances, since atoms that are more than a certain distance apart do not have interactions that would affect the binding pose. We use 10 randomly generated molecular conformations as data augmentation when sampling. Also, a lower dist_threshold is used to reduce variance in sampling with consideration of error in prediction. The details of hyperparameters are shown in Table 11.

- **Exhaustiveness search** To ensure that the comparison between Uni-Mol and popular molecular docking software is unbiased, we increase the exhaustiveness of the global search (roughly proportional to time) of the molecular docking software to observe the effect of computational complexity to docking power on CASF-2016 benchmark. And we find that when exhaustiveness is above 16, the popular molecular docking software can no longer improve the performance by increasing the computational complexity.

Table 12: Exhaustiveness study of popular docking tools on CASF-2016

		Ligand RMSD % Below Threshold ↑			
Methods	Exhaustiveness	0.5 Å	1.0 Å	1.5 Å	2.0 Å
Autodock Vina	1	21.40	35.79	47.02	52.28
Autodock Vina	8	23.86	44.21	57.54	64.56
Autodock Vina	16	25.61	45.96	60.70	66.67
Autodock Vina	32	25.96	45.96	60.00	66.32
Vinardo	1	16.84	33.33	43.16	49.82
Vinardo	8	23.51	41.75	57.54	62.81
Vinardo	16	23.51	45.26	60.70	66.67
Vinardo	32	23.86	44.56	59.30	65.61
Smina	1	23.51	39.65	50.53	56.14
Smina	8	23.51	47.37	59.65	65.26
Smina	16	**28.77**	49.47	61.40	67.72
Smina	32	28.07	51.23	61.75	67.37
Autodock4	1	4.91	18.95	26.67	28.87
Autodock4	8	7.02	21.75	31.58	35.44
Autodock4	16	6.32	24.56	34.04	38.95
Autodock4	32	6.32	23.16	34.04	38.25
Uni-Mol$_{\text{random}}$	-	14.04	49.47	65.26	75.44
Uni-Mol	-	24.91	**70.53**	**84.21**	**88.07**

D Metrics

In the conformation generation task, following previous work [80, 81], we use the Root of Mean Squared Deviations (RMSD) of heavy atoms to evaluate the difference between the generated conformation and the reference one. Before computing RMSD, the generated conformation is first aligned with the reference one, and the function Φ aligns conformations by applying rotations and translations to them:

$$\text{RMSD}(\boldsymbol{R}, \hat{\boldsymbol{R}}) = \min_{\Phi}\left(\frac{1}{n}\sum_{i=1}^{n}||\Phi(\boldsymbol{R}_i) - \hat{\boldsymbol{R}}_i||^2\right)^{\frac{1}{2}} \tag{5}$$

where $\boldsymbol{R}$ and $\hat{\boldsymbol{R}}$ are the generated and reference conformation, i is the i-th heavy atom, and n is the number of heavy atoms.

We use Coverage (COV) and Matching (MAT) to evaluate the performance of the conformation generation model. Higher COV means better diversity, while lower MAT means higher accuracy. Formally, COV and MAT are denoted as:

$$\text{COV}(S_g, S_r) = \frac{\left|\left\{\boldsymbol{R} \in S_r | \text{RMSD}(\boldsymbol{R}, \hat{\boldsymbol{R}}) < \delta, \hat{\boldsymbol{R}} \in S_g\right\}\right|}{|S_r|} \tag{6}$$

$$\text{MAT}(S_g, S_r) = \frac{1}{|S_r|}\sum_{\boldsymbol{R} \in S_r}\min_{\hat{\boldsymbol{R}} \in S_g}\text{RMSD}(\boldsymbol{R}, \hat{\boldsymbol{R}}) \tag{7}$$

where S_g and S_r are the set of generated and reference conformations, respectively, and δ is a given RMSD threshold. Following previous work [43, 52], for GEOM-QM9, the threshold is 0.5Å, and for GEOM-Drugs, the threshold value is 1.25Å.

E Ablation studies

We investigate the impact of the pair-type aware affine (PTAA) module on the molecular property prediction tasks. As described in Sec 2.1, in invariant spatial positional encoding, the PTAA is combined with the pair Euclidean distance matrix. Tables 13 and 14 show the results of the ablation

Table 13: Ablation study on pair-type with molecular property prediction classification tasks

Classification (ROC-AUC %, higher is better ↑)									
Datasets	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	PCBA	MUV
Uni-Mol w/o pair-type	66.3(1.7)	76.2(0.2)	87.1(2.3)	72.4(0.1)	62.3(0.4)	61.2(1.1)	75.8(0.5)	85.1(0.1)	80.9(0.6)
Uni-Mol	**72.9(0.6)**	**85.7(0.2)**	**91.9(1.8)**	**79.6(0.5)**	**69.6(0.1)**	**65.9(1.3)**	**80.8(0.3)**	**88.5(0.1)**	**82.1(1.3)**

Table 14: Ablation study on pair-type with molecular property prediction regression tasks

	Regression (lower is better)					
	RMSE			MAE		
Datasets	ESOL	FreeSolv	Lipo	QM7	QM8	QM9
Uni-Mol w/o pair-type	0.977(0.007)	2.053(0.053)	0.951(0.056)	45.9(1.7)	0.0156(0.0001)	0.00473(0.00004)
Uni-Mol	**0.788(0.029)**	**1.620(0.035)**	**0.603(0.010)**	**41.8(0.2)**	**0.0156(0.0001)**	**0.00467(0.00004)**

studies, and we can find that PTAA largely improves the performance of molecular property prediction. There are several possible reasons: 1) in chemicals (and physics), the interactions between two atoms are determined by their distances and types together. Given pair distance and their types, the model can distinguish different interactions, such as Van der Waals forces, covalent interactions, etc., and thus perform better. 2) PTAA enlarges the capacity of pair representation by introducing more trainable parameters, and therefore, the model learns better pair interactions in 3D space and thus performs better.

F Training Stability

With Pre-LayerNorm [30] backbone and mixed-precision training, the pretraining sometimes diverges. After investigation, we found there are large numerical values in the intermediate states when divergence happens. We hypothesize that the Final-LayerNorm layer in the Pre-LayerNorm backbone results in the problem. Specifically, Final-LayerNorm is applied to the sum of all encoder layers, denoted as

$$\boldsymbol{o}_i = \text{LayerNorm}(\boldsymbol{s}_i), \quad \boldsymbol{s}_i = \sum_{l=1}^{L} \boldsymbol{o}_i^l \tag{8}$$

where L is the number of layers, $\boldsymbol{o}_i^l$ is the output of the i-th position in the l-th layer, and $\boldsymbol{o}_i$ is the final output of the i-th position, after Final-LayerNorm. Therefore, due to normalization, $\boldsymbol{s}_i$ can be arbitrarily large (or arbitrarily small), without affecting model results. However, a too large or too small numerical value will cause the numerical unstable, especially in the mixed-precision training. To tackle this, we introduce a simple loss, to restrict the value range of $\boldsymbol{s}_i$. Formally, the loss is denoted as

$$\mathcal{L}_{norm} = \text{mean}_i \left(\max \left(\left| \|\boldsymbol{s}_i\| - \sqrt{d} \right| - \tau, 0 \right) \right), \tag{9}$$

where d is the dimension size of $\boldsymbol{s}_i$, τ is the tolerance factor. In Uni-Mol, we set $\tau = 1$, and both atom-level and pair-level representations are constrained by this loss. Besides, to avoid affecting other loss functions, we set a very small loss weight (0.01) to $\mathcal{L}_{norm}$.