MPerformer: An SE(3) Transformer-based Molecular Perceptron

Fanmeng Wang*
Renmin University of China
DP Technology
Beijing, China
fanmengwang@ruc.edu.cn

Hongteng Xu†
Renmin University of China
Beijing, China
hongtengxu@ruc.edu.cn

Xi Chen
DP Technology
Beijing, China
chenx@dp.tech

Shuqi Lu†
DP Technology
Beijing, China
lusq@dp.tech

Yuqing Deng
DP Technology
Beijing, China
dengyq@dp.tech

Wenbing Huang
Renmin University of China
Beijing, China
hwenbing@ruc.edu.cn

ABSTRACT

Molecular perception aims to construct 3D molecules from 3D atom clouds (i.e., atom types and corresponding 3D coordinates), determining bond connections, bond orders, and other molecular attributes within molecules. It is essential for realizing many applications in cheminformatics and bioinformatics, such as modeling quantum chemistry-derived molecular structures in protein-ligand complexes. Additionally, many molecular generation methods can only generate molecular 3D atom clouds, requiring molecular perception as a necessary post-processing. However, existing molecular perception methods mainly rely on predefined chemical rules and fail to leverage 3D geometric information, whose performance is sub-optimal fully. In this study, we propose MPerformer, an SE(3) Transformer-based molecular perceptron exhibiting SE(3)-invariance, to construct 3D molecules from 3D atom clouds efficiently. Besides, we propose a multi-task pretraining-and-finetuning paradigm to learn this model. In the pretraining phase, we jointly minimize an attribute prediction loss and an atom cloud reconstruction loss, mitigating the data imbalance issue of molecular attributes and enhancing the robustness and generalizability of the model. Experiments show that MPerformer significantly outperforms state-of-the-art molecular perception methods in precision and robustness, benefiting various molecular generation scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Multi-task learning**; • **Information systems** → **Data mining**.

KEYWORDS

Molecular Perception, SE(3) Transformer, Multi-task Pretraining, Molecular Generation

*Work done during an internship at DP Technology
†Corresponding author

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614974

ACM Reference Format:
Fanmeng Wang, Hongteng Xu, Xi Chen, Shuqi Lu, Yuqing Deng, and Wenbing Huang. 2023. MPerformer: An SE(3) Transformer-based Molecular Perceptron. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614974

1 INTRODUCTION

Molecular 3D structure modeling plays a central role in many applications of Chemistry and Biology, which is essential for exploiting the various molecular knowledge space [11, 35, 45, 62]. Typically, molecular structural data are collected from quantum chemistry calculations or crystallography measurements. Recently, with the application of artificial intelligence techniques, many learning-based methods have been proposed to generate 3D molecules [9, 32–34, 49]. However, both the real-world measurements and the generated data [24, 29, 31, 40] are often formulated as molecular 3D atom clouds (i.e., atom types and their corresponding 3D coordinates), while other molecular attributes, e.g., bond connections and bond orders, are unknown in general [8]. Such lack of molecular attributes has given rise to the concept of **molecular perception** [47, 60], which employs 3D atom clouds to predict bond connections, bond orders, and other molecular attributes within molecules, thus constructing 3D molecules with complete chemical information.

The classic molecular perception methods leverage predefined chemical rules to construct 3D molecules [1, 16, 21, 22], in which the target molecular attributes (like bond orders and formal charges) are determined by manually-designed score functions [10, 50] or heuristic searching techniques [19]. However, *these rule-based methods are sensitive to the precision of atom 3D coordinates, and some of them further rely on additional information like atom connectivity [47, 54].* As a result, these methods become unstable and even inapplicable in highly-noisy scenarios, e.g., dealing with the atom clouds generated by 3D molecular generative models. Moreover, the generalization power of these methods is limited by the predefined chemical rules, making them unsuitable for exploratory tasks like drug discovery. Recently, some learning-based methods have been proposed, e.g., the Knodle in [18] and the HIP-NN in [30]. These methods apply various machine learning models, e.g., SVM [18], decision tree [28], and neural network [17, 36], to predict bond orders. However, *existing learning-based methods ignore the multi-task nature of molecular perception, merely focusing on the prediction of*

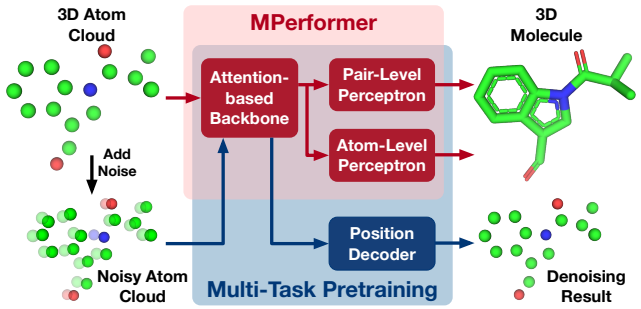


Figure 1: An illustration of MPerformer and its learning paradigm. Given an atom cloud, MPerformer derives its atom-level and pair-level representations and predicts a 3D molecule with complete chemical information. During pretraining, we connect MPerformer with a decoder, reconstructing the atom cloud from its noisy versions.

bond orders. Additionally, due to the limitations on model capacity and learning paradigm, these methods cannot fully leverage 3D geometric information contained in molecular 3D atom clouds.

In this study, we propose an effective molecular perceptron based on SE(3) Transformer, called MPerformer, providing a new learning-based molecular perception model. As illustrated in Figure 1, given a molecular 3D atom cloud, MPerformer applies an attention-based backbone model to extract SE(3)-invariant atom-level and pair-level molecular representations, respectively. Based on the learned molecular representations, we use two perception heads (a.k.a. perceptrons) to predict corresponding molecular attributes and construct the 3D molecule accordingly. Inspired by existing 3D molecular representation models like 3D-Graphormer [44] and Uni-Mol [61], we implement the attention-based backbone model in our MPerformer as the SE(3) Transformer [13], which fully leverages the 3D geometric information contained in molecular 3D atom clouds and ensures the SE(3)-invariance of the molecular representations.

To learn the proposed MPerformer effectively, we design a multi-task pretraining-and-finetuning paradigm. In the pretraining phase, we introduce a denoising task to the learning objective besides the atom-level and pair-level prediction tasks. In particular, we impose random position noise on the input atom cloud and derive the molecular representations of the noisy atom cloud through our backbone model. Accordingly, we reconstruct the original clear atom cloud by decoding the molecular representations of the noisy atom cloud and penalize the reconstruction error. Pretraining under the multi-tasks provides us with a robust and generalizable model. Finally, finetuning the model based on the atom-level and pair-level prediction tasks leads to the proposed MPerformer.

To the best of our knowledge, MPerformer is the first universal learning-based molecular perception method to construct 3D molecules with complete chemical information purely based on molecular 3D atom clouds. Experiments show that MPerformer significantly and consistently outperforms state-of-the-art molecular perception methods (including rule-based methods and learning-based ones), especially in those challenging molecular perception cases. In addition, it exhibits excellent robustness to noise in the molecular 3D coordinates. Moreover, taking MPerformer as the post-processing module of representative 3D molecular generators [24, 31, 40] can effectively improve the quality of generated molecules, further showing the great application potential of MPerformer in the practical downstream scenarios.

2 RELATED WORK

2.1 Molecular Perception Methods

In the past few decades, many rule-based methods dependent on predefined chemical rules have been proposed for molecular perception. The work in [22] first introduced a backtracking search method called COBRA to construct molecules. In [1, 16, 21], some simple geometric features (e.g., bond length and angle) were used for molecular perception. The work in [50] presented a heuristic algorithm by penalizing a scoring function. Following this strategy, some other manually-designed score functions were proposed in [10, 47, 54]. However, these rule-based methods are sensitive to the precision of atom 3D coordinates, and some even require additional information like atom connectivity. Moreover, the generalization power of these rule-based molecular perception methods is limited by the predefined chemical rules, so they cannot greatly handle various unseen cases, such as AI-generated atom clouds.

Besides the above rule-based methods, some learning-based approaches have been proposed for molecular perception, especially bond-order perception. The work in [18] predicted bond orders based on nonlinear Support Vector Machines (SVM). Following this strategy, other learning-based bond-order predictors were developed based on different machine learning models, e.g., decision tree [28] and neural network [30, 36]. However, due to the limitations of their model architectures and learning paradigms, these learning-based methods cannot leverage 3D geometric information fully and ignore the multi-task nature of molecular perception, thus failing to solve the molecular perception problem effectively.

Nowadays, the most widely-used molecular perception tools are the open-source tool OpenBabel [38] and the commercial software Schrödinger Maestro [4]. However, their performance is often unsatisfactory in practical application scenarios.

2.2 Molecular Representation Learning

A problem highly correlated with molecular perception is molecular representation learning. Since representation learning has already made significant progress in the field of natural language processing [2, 3, 26, 39], the early works in molecular representation learning (MRL) are mainly based on 1D sequence representation of molecules, such as SMILES-BERT [52]. Then, with the rapid development of graph neural networks [43, 56, 58] in recent years, some molecular representation methods have been developed based on molecular graphs [5, 12, 42, 53]. Recently, considering the properties of molecules are primarily determined by the 3D structure of molecules, more and more works are trying to leverage the 3D geometric information [11, 25, 44, 61]. For example, the 3D Graphormer in [44] leverages a transformer-based model to represent 3D molecules. Similarly, the work in [61] proposed a universal 3D molecular pretraining framework, called Uni-Mol. In fact, these 3D molecular representation models have provided many valuable backbones for encoding molecular 3D information.

3 PROPOSED MPERFORMER

3.1 Problem Statement

In this study, we denote a 3D molecule with N atoms as a tuple $\mathcal{M} = (\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{V}, \boldsymbol{E})$. Here, $\boldsymbol{A} = [a_i] \in \{1, ..., S\}^N$ is an integer vector representing the types of the N atoms. S is the size of the predefined atom-type dictionary based on datasets. $\boldsymbol{C} = [\boldsymbol{c}_i] \in \mathbb{R}^{N\times 3}$ is a matrix containing the 3D coordinates of the N atoms, whose row vector $\boldsymbol{c}_i$ represents the 3D coordinate of the i-th atom. $\boldsymbol{V} = \{\boldsymbol{V}^k\}_{k=1}^{K_v}$ represents a set of K_v atom-level attributes (e.g., formal charges, hydrogen numbers, and so on). These attributes are categorical data, and we represent them as binary matrices with one-hot rows, i.e., $\boldsymbol{V}^k = [\boldsymbol{v}_i^k] \in \{0,1\}^{N\times V_k}$ represents the k-th atom-level attribute, where V_k represents the number of possible values for the k-th atom-level attributes and $\boldsymbol{v}_i^k$ is the one-hot vector corresponding to the k-th atom-level attribute of the i-th atom. Similarly, $\boldsymbol{E} = \{\boldsymbol{E}^k\}_{k=1}^{K_e}$ represents a set of K_e pair-level attributes (e.g., bond connections and bond orders), where $\boldsymbol{E}^k = [\boldsymbol{e}_{ij}^k] \in \{0,1\}^{N\times N\times E_k}$ represents the k-th pair-level attribute, E_k is the number of possible values of the k-th pair-level attributes, and $\boldsymbol{e}_{ij}^k$ is the one-hot vector corresponding to the k-th pair-level attribute of the atom pair (i, j).

As aforementioned, in practice what we observed is a molecular 3D atom cloud, i.e., the atoms' types and their 3D coordinates, rather than a complete 3D molecule. Therefore, we would like to learn a molecular perceptron, denoted as ϕ, to construct the complete 3D molecule purely based on the corresponding 3D atom cloud, i.e.,

$$\widehat{\mathcal{M}} = (\boldsymbol{A}, \boldsymbol{C}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{E}}), \text{ where } \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{E}} = \phi(\boldsymbol{A}, \boldsymbol{C}). \tag{1}$$

An ideal molecular perceptron should have the following properties:

- **SE(3)-invariance.** The model output should be invariant to the rotations and translations of the input 3D atom cloud.
- **Robustness to position noise.** The atom clouds in practice are often noisy, whose 3D coordinates often have some randomness. The proposed model should be robust to the position noise imposed on the atom clouds.
- **Generalization power.** Molecular perception commonly works as the post-processing of molecular generation, which serves for exploratory scientific tasks like drug discovery. Therefore, given unseen atom clouds that correspond to new molecules, the proposed model should still be able to construct the molecules with high precision.

To achieve the above properties, we will introduce our proposed MPerformer and design an effective multi-task pretraining-and-finetuning learning paradigm for it in the following content.

3.2 Extracting Molecular Representations

As illustrated in Figure 2(a), our MPerformer leverages the SE(3) Transformer as its backbone, extracting atom-level and pair-level molecular representations from the input 3D atom clouds. In particular, given a 3D atom cloud with N atoms, we encode its atom types $\boldsymbol{A}$ by an embedding layer, i.e.,

$$\boldsymbol{X}_a^{(0)} = g_a(\boldsymbol{A}) \in \mathbb{R}^{N\times D_a}. \tag{2}$$

Here, $g_a : \{1, ..., S\} \mapsto \mathbb{R}^{D_a}$ embeds each atom type to a D_a-dimensional atom type embedding. Applying g_a to each atom leads to the initial atom-level representation, denoted as $\boldsymbol{X}_a^{(0)}$.

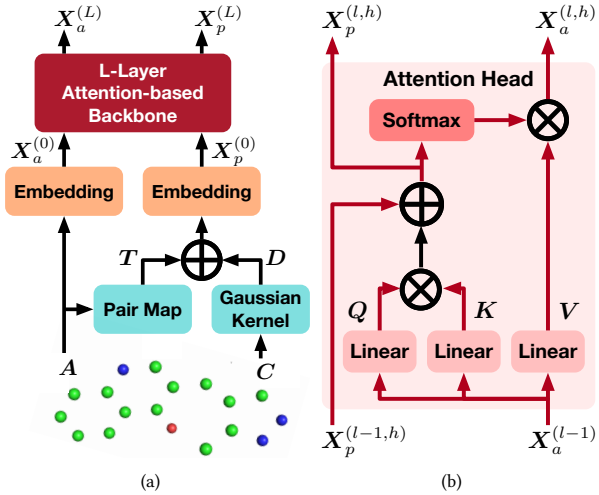


Figure 2: (a) An overview of the backbone model architecture. (b) The illustration of the h-th attention head in the l-th layer.

Additionally, we map the atom types $\boldsymbol{A}$ to a type-based relational matrix $\boldsymbol{T} = [t_{ij}] \in \mathbb{R}^{N\times N}$ and use a Gaussian kernel to encode the coordinates $\boldsymbol{C}$ as a position-based relational matrix $\boldsymbol{D} = [d_{ij}] \in \mathbb{R}^{N\times N}$. For $i, j \in \{1, .., N\}$, the elements of the two matrices are respectively represented as

$$t_{ij} = t(a_i, a_j) = S \cdot a_i + a_j, \quad d_{ij} = \exp\Big(-\frac{\|\boldsymbol{c}_i - \boldsymbol{c}_j\|^2}{2b^2}\Big), \tag{3}$$

where $t(a_i, a_j)$ is a function mapping the pair of atom types (a_i, a_j) to a scalar value, S is the size of the predefined atom-type dictionary, and b is a hyperparameter controlling the bandwidth of the Gaussian kernel. In this study, we implement $t(a_i, a_j)$ as a linear function $S \cdot a_i + a_j$. This implementation ensures that $t_{ij} \neq t_{i'j'} \Leftrightarrow a_i \neq a_{i'}$ or $a_j \neq a_{j'}$, such that the elements in the relational matrix $\boldsymbol{T}$ encodes the discriminative information for different pairs of atom types. Based on these two relational matrices, we derive the initial pair-level representation as follows:

$$\boldsymbol{X}_p^{(0)} = g_p(\boldsymbol{T} + \boldsymbol{D}) \in \mathbb{R}^{N\times N\times D_p}, \tag{4}$$

where $g_p : \mathbb{R} \mapsto \mathbb{R}^{D_p}$ is a linear map that embeds each element of $(\boldsymbol{T} + \boldsymbol{D})$ to a D_p-dimensional embedding, leading to the initial pair-level representation $\boldsymbol{X}_p^{(0)}$. Note that, we only leverage the atom type and distance information at the pair-level shown in (3), so the representations $\boldsymbol{X}_a^{(0)}$ in (2) and $\boldsymbol{X}_p^{(0)}$ in (4) are SE(3)-invariant.

Given $\boldsymbol{X}_a^{(0)}$ and $\boldsymbol{X}_p^{(0)}$ as input, we further encode them by an attention-based backbone model. The backbone model contains L SE(3) Transformer-based encoding layers, each consisting of a multi-head self-attention module and a feed-forward neural network module. Figure 2(b) illustrates the architecture of the h-th head in the l-th layer. Different from the classic self-attention mechanism [48], the interactions between the atom-level and pair-level representations in each head are considered. We take the dimension of pair-level embedding D_p as the number of attention heads. For

$l = 1, ..., L$ and $h = 1, ..., D_p$, we have

$$\boldsymbol{Att}^{(l,h)} = \frac{\boldsymbol{Q}^{(l,h)}(\boldsymbol{K}^{(l,h)})^\top}{\sqrt{D_h}}, \quad \boldsymbol{X}_p^{(l,h)} = \boldsymbol{Att}^{(l,h)} + \boldsymbol{X}_p^{(l-1,h)}, \\ \text{and} \quad \boldsymbol{X}_a^{(l,h)} = \sigma\left(\boldsymbol{Att}^{(l,h)} + \boldsymbol{X}_p^{(l-1,h)}\right)\boldsymbol{V}^{(l,h)}, \tag{5}$$

where $\boldsymbol{Q}^{(l,h)} = \boldsymbol{X}_a^{(l-1)}\boldsymbol{W}_Q^{(l,h)}$, $\boldsymbol{K}^{(l,h)} = \boldsymbol{X}_a^{(l-1)}\boldsymbol{W}_K^{(l,h)}$, and $\boldsymbol{V}^{(l,h)} = \boldsymbol{X}_a^{(l-1)}\boldsymbol{W}_V^{(l,h)}$ are query, key, and value matrices of the corresponding head. $\boldsymbol{W}_Q^{(l,h)}, \boldsymbol{W}_K^{(l,h)}, \boldsymbol{W}_V^{(l,h)} \in \mathbb{R}^{D_a \times D_h}$ are linear maps, where $D_h = D_a/D_p$ is the hidden dimension of each head. $\sigma(\cdot)$ is the row-wise softmax operator. $\boldsymbol{X}_p^{(l-1,h)} \in \mathbb{R}^{N\times N}$ is the h-th slice of $\boldsymbol{X}_p^{(l-1)}$. As shown in (5), the update of the pair-level representation is based on the query-key matrix derived from the atom-level representation, and the attention map of the atom-level representation also considers the impact of the pair-level representation.

Given the output of each head, the l-th layer outputs $\boldsymbol{X}_a^{(l)} \in \mathbb{R}^{N\times D_a}$ and $\boldsymbol{X}_p^{(l)} \in \mathbb{R}^{N\times N\times D_p}$ as follows:

$$\boldsymbol{X}_a^{(l)} = \text{Concat}(\{\boldsymbol{X}_a^{(l,h)}\}_{h=1}^{D_p})\boldsymbol{W}_O^{(l)} + \boldsymbol{X}_a^{(l-1)}, \\ \boldsymbol{X}_p^{(l)} = \text{Concat}(\{\boldsymbol{X}_p^{(l,h)}\}_{h=1}^{D_p}), \tag{6}$$

where $\boldsymbol{W}_O^{(l)} \in \mathbb{R}^{D_a\times D_a}$ is a linear map and $\text{Concat}(\cdot)$ represents the concatenation operator. Stacking the above layers (i.e., (5) and (6)) L times leads to the proposed backbone model. Passing $\boldsymbol{X}_a^{(0)}$ and $\boldsymbol{X}_p^{(0)}$ through the backbone model leads to information-rick and discriminative representations, i.e., $\boldsymbol{X}_a^{(L)}$ and $\boldsymbol{X}_p^{(L)}$.

3.3 Molecular Attribute Perception

As shown in Figure 1, after the backbone model extracts final atom-level and pair-level molecular representations $\boldsymbol{X}_a^{(L)}$ and $\boldsymbol{X}_p^{(L)}$, our proposed MPerformer infers atom-level and pair-level molecular attributes by two perception heads, respectively. Here, each perception head is a collection of multi-layer perceptrons (MLPs).

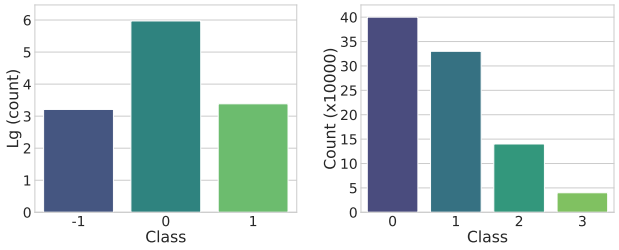
Specifically, for the K_v atom-level attributes (e.g., formal charges, hydrogen numbers, and so on), we apply K_v MLPs, denoted as $\{f_a^k\}_{k=1}^{K_v}$ to infer corresponding atom-level attributes. Passing the atom-level representation $\boldsymbol{X}_a^{(L)}$ through the MLPs can provide us with the estimated atom-level attributes. Similarly, the pair-level perception head contains K_e MLPs, i.e., $\{f_p^k\}_{k=1}^{K_e}$, each of which takes the pair-level representation $\boldsymbol{X}_p^{(L)}$ as input and predicts a corresponding pair-level attribute, e.g., bond orders, and so on. The formulations of the perception heads are shown below:

$$\boldsymbol{P}^k = [\boldsymbol{p}_i^k] = f_a^k(\boldsymbol{X}_a^{(L)}) \in \mathbb{R}^{N\times V_k}, \ \forall k = 1, ..., K_v, \\ \boldsymbol{Q}^k = [\boldsymbol{q}_{ij}^k] = f_p^k(\boldsymbol{X}_p^{(L)}) \in \mathbb{R}^{N\times N\times E_k}, \ \forall k = 1, ..., K_e, \tag{7}$$

where each MLP estimates the distributions of one attribute associated with atoms or atom pairs. $\boldsymbol{p}_i^k \in \Delta^{V_k}$ represents the distribution of the k-th atom-level attribute at the i-th atom, and $\boldsymbol{q}_{ij}^k \in \Delta^{E_k}$ represents the distribution of the k-th pair-level attribute at the atom pair (i, j), where Δ denotes the Simplex. Finally, we can predict $\widehat{\boldsymbol{V}} = \{\widehat{\boldsymbol{V}}^k\}_{k=1}^{K_v}$ (and $\widehat{\boldsymbol{E}} = \{\widehat{\boldsymbol{E}}^k\}_{k=1}^{K_e}$) by selecting the attributes with the highest probabilities for each atom (and each atom pair), i.e.,

$$\widehat{\boldsymbol{V}}^k = \arg\max(\boldsymbol{P}^k), \ \forall k = 1, ..., K_v, \\ \widehat{\boldsymbol{E}}^k = \arg\max(\boldsymbol{Q}^k), \ \forall k = 1, ..., K_e, \tag{8}$$

and accordingly, we obtain the 3D molecule with complete chemical information (attributes) as $\widehat{\mathcal{M}} = (\boldsymbol{A}, \boldsymbol{C}, \widehat{\boldsymbol{V}}, \widehat{\boldsymbol{E}})$.



(a) Formal charges (b) Hydrogen numbers

Figure 3: Histograms of formal charges and hydrogen numbers in the chemical component dictionary (CCD) [55].

4 MULTI-TASK PRETRAINING AND FINETUNING

As shown above, molecular perception involves predicting multiple atom-level and pair-level attributes. It is a multi-task learning problem [59] in nature, which suffers from the following challenges:

- **Data Imbalance.** Many molecular attributes exhibit significant data imbalance issues. As shown in Figure 3, for the atoms from the chemical component dictionary (CCD) [55], the formal charges of most atoms are zero, while only a small fraction have non-zero focal charges (e.g., 1 or -1), Besides, the number of the atoms disconnecting to hydrogen is significantly larger than that of the atoms connecting to three hydrogens. In addition, as shown in Table 1, there are far more null bonds (i.e., disconnected atom pairs) than any other types of bonds, especially aromatic bonds. Since the data imbalance issue can greatly harm the training of the model, we must design a reasonable loss function to suppress this data imbalance issue.
- **Sensitivity to Molecular Conformations.** Each molecule often has various conformations, and the conformations correspond to different atom clouds. Mathematically, we can treat the atom clouds of the conformations as the "noisy" versions of a standard atom cloud. Given such noisy atom clouds, we need to predict the same molecular attributes, which requires the molecular perception model robust to noise and with good generalization power.

To overcome these challenges, we design a **multi-task pretraining-and-finetuning** learning paradigm, which considers the following two learning objectives.

4.1 Attribute Prediction Loss

Because the atom-level and pair-level attributes are categorical in this study, we can consider the cross-entropy loss between the

outputs of our model and the ground truth. However, traditional cross-entropy loss treats all samples evenly, resulting in an inadequate focus on those in minority classes. To suppress the data imbalance issue, we use a multi-class focal (MCF) loss, which extends the focal loss in [23] from binary classification to multi-class cases, to replace the traditional cross-entropy loss in our work. This loss dynamically adjusts the weights of the cross-entropy losses for different samples. In particular, for k-th atom-level attribute $\boldsymbol{V}^k = [\boldsymbol{v}_i^k] \in \{0,1\}^{N\times V_k}$, our model outputs the corresponding probability matrix $\boldsymbol{P}^k = [\boldsymbol{p}_i^k] \in [0,1]^{N\times V_k}$. Our MCF loss is

$$\mathcal{L}_{MCF}(\boldsymbol{V}^k, \boldsymbol{P}^k) = -\frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{w}_i^k \odot \boldsymbol{v}_i^k, \log \boldsymbol{p}_i^k\rangle, \tag{9}$$

where $\odot$ represents the Hadamard product, $\langle\cdot,\cdot\rangle$ represents the inner product operation, and the weight vector $\boldsymbol{w}_i^k \in \mathbb{R}_+^{V_k}$ represents the nonnegative weights corresponding to the V_k classes in the k-th attribute. For the atom i, we set

$$\boldsymbol{w}_i^k = \boldsymbol{\alpha}^k \odot (\mathbf{1}_{V_k} - \boldsymbol{p}_i^k))^\gamma. \tag{10}$$

Here, $\boldsymbol{\alpha}^k = \{1 - \frac{M_j^k}{M}\}_{j=1}^{V_k} \in \mathbb{R}_+^{V_k}$, where M is the total number of atoms, and M_j^k is the number of the atoms whose k-th attribute is j, so that the minority classes have large values.

In the same way, for k-th pair-level attribute $\boldsymbol{E}^k = [\boldsymbol{e}_{ij}^k] \in \{0,1\}^{N\times N\times E_k}$, our model outputs the corresponding probability $\boldsymbol{Q}^k = [\boldsymbol{q}_i^k] \in \{0,1\}^{N\times N\times E_k}$. Our MCF loss is

$$\mathcal{L}_{MCF}(\boldsymbol{E}^k, \boldsymbol{Q}^k) = -\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}\langle \boldsymbol{w}_{ij}^k \odot \boldsymbol{e}_{ij}^k, \log \boldsymbol{q}_{ij}^k\rangle, \tag{11}$$

where $\boldsymbol{w}_{ij}^k = \boldsymbol{\alpha}^k \odot (\mathbf{1}_{E_k} - \boldsymbol{q}_{ij}^k))^\gamma$. $\boldsymbol{\alpha}^k = \{1 - \frac{M_j^k}{M}\}_{j=1}^{E_k}$, where M is the total number of atom pairs, and M_j^k is the number of the atom pairs whose k-th attribute is j.

By applying this MCF loss, our model can take the low-confident and incorrect predictions of the minority classes more seriously.

Prediction loss. Taking each attribute prediction as a learning task, we learn the prediction module (the attention-based backbone model and all MLPs) jointly. The prediction loss of our MPerformer is the weighted summation of all the MCF losses, i.e.,

$$\mathcal{L}_{pred}(\boldsymbol{P}, \boldsymbol{Q}) = \sum_{k=1}^{K_v}\beta^k\mathcal{L}_{MCF}(\boldsymbol{V}^k, \boldsymbol{P}^k) + \sum_{k=1}^{K_e}\delta^k\mathcal{L}_{MCF}(\boldsymbol{E}^k, \boldsymbol{Q}^k) \tag{12}$$

where $\boldsymbol{\beta} = [\beta^k] \in \mathbb{R}^{K_v}$ and $\boldsymbol{\delta} = [\delta^k] \in \mathbb{R}^{K_e}$ are used to control the weights of different tasks.

4.2 Atom Cloud Reconstruction Loss

Besides the prediction loss, we further introduce an atom cloud reconstruction loss to enhance the robustness of our model to the position noise in atom clouds. As illustrated in Figure 1, we first add a random noise to the atom 3D coordinates $\boldsymbol{C} \in \mathbb{R}^{N\times 3}$ to obtain the noisy atom 3D coordinates $\widetilde{\boldsymbol{C}} = [\tilde{\boldsymbol{c}}_i] \in \mathbb{R}^{N\times 3}$. Passing the noisy atom cloud $(\boldsymbol{A}, \widetilde{\boldsymbol{C}})$ through our MPerformer, we obtain i) the noisy molecular representation, including the final atom-level representation $\widetilde{\boldsymbol{X}}_a^{(L)}$, the initial pair-level representation $\widetilde{\boldsymbol{X}}_p^{(0)}$ and the final $\widetilde{\boldsymbol{X}}_p^{(L)}$, and ii) the noisy predicted molecular attributes $(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{E}})$. Here, we further consider two learning tasks.

Firstly, we hope that the molecular attributes can be predicted with high accuracy based on the noisy representations. Therefore, passing $\widetilde{\boldsymbol{X}}_a^{(L)}$ and $\widetilde{\boldsymbol{X}}_p^{(L)}$ through the perception heads ($\{f_a^k\}_{k=1}^{K_v}$ and $\{f_p^k\}_{k=1}^{K_e}$), we can get the predictions of molecular attributes and reuse the prediction loss in (12) to penalize the prediction errors.

Secondly, we design a decoder to reconstruct the original clear atom cloud from the noisy one. The decoder takes the noisy coordinates $\widetilde{\boldsymbol{C}}$ and the noisy pair-level representations as input, and outputs the reconstruction result, denoted as $\widehat{\boldsymbol{C}} = [\hat{\boldsymbol{c}}_i]$, as follows:

$$\hat{\boldsymbol{c}}_i = \tilde{\boldsymbol{c}}_i + \sum_{j=1}^{N}\frac{\psi(\tilde{\boldsymbol{x}}_{p,ij}^{(L)} - \tilde{\boldsymbol{x}}_{p,ij}^{(0)})(\tilde{\boldsymbol{c}}_i - \tilde{\boldsymbol{c}}_j)}{N}, \tag{13}$$

where $\tilde{\boldsymbol{x}}_{p,ij}^{(L)}, \tilde{\boldsymbol{x}}_{p,ij}^{(0)} \in \mathbb{R}^{D_p}$ are pair-level representations corresponding to the atom pair (i,j), and $\psi : \mathbb{R}^{D_p} \mapsto \mathbb{R}$ is an MLP mapping the residual of the pair-level representations to a scalar.

Reconstruction loss. After reconstructing the original clear atom cloud, we consider the reconstruction loss between the original clear atom cloud $\boldsymbol{C}$ and the reconstructed $\widehat{\boldsymbol{C}}$. In this study, we implement the reconstruction loss as the smooth L1 loss [51]:

$$\mathcal{L}_{rec}(\boldsymbol{C}, \widehat{\boldsymbol{C}}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3}\begin{cases}0.5|c_{ij} - \hat{c}_{ij}|^2, & \text{if } |c_{ij} - \hat{c}_{ij}| < 1,\\ |c_{ij} - \hat{c}_{ij}| - 0.5, & \text{otherwise.}\end{cases} \tag{14}$$

In the following experiments, we will show that considering the above two losses helps to enhance the robustness of our model.

Algorithm 1 Multi-task Pretraining and Finetuning of MPerformer

Require: A 3D molecular dataset $\mathcal{D}$.
Pretraining phase:
for A batch $\mathcal{B} \subset \mathcal{D}$ **do**
  for Each $(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{V}, \boldsymbol{E}) \in \mathcal{B}$ **do**
    Add noise to $\boldsymbol{C}$ and get $\widetilde{\boldsymbol{C}}$.
    Obtain $\{\widetilde{\boldsymbol{X}}_a^{(L)}, \widetilde{\boldsymbol{X}}_p^{(L)}\}$ and predictions $\{\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{Q}}\}$ by $\phi(\boldsymbol{A}, \widetilde{\boldsymbol{C}})$.
    Obtain the prediction loss $\mathcal{L}_{pred}(\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{Q}})$ by (12).
    Reconstruct $\widehat{\boldsymbol{C}}$ by (13).
    Obtain the reconstruction loss $\mathcal{L}_{rec}(\boldsymbol{C}, \widehat{\boldsymbol{C}})$ by (14).
  end for
  $\min_{\phi,\psi}\sum_{(\boldsymbol{A},\boldsymbol{C},\boldsymbol{V},\boldsymbol{E})\in\mathcal{B}}\mathcal{L}_{pred}(\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{Q}}) + \mathcal{L}_{rec}(\boldsymbol{C}, \widehat{\boldsymbol{C}})$ via Adam and update ϕ and ψ jointly.
end for
Finetuning phase:
for A batch $\mathcal{B} \subset \mathcal{D}$ **do**
  for Each $(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{V}, \boldsymbol{E}) \in \mathcal{B}$ **do**
    Obtain predictions $\{\boldsymbol{P}, \boldsymbol{Q}\}$ by $\phi(\boldsymbol{A}, \boldsymbol{C})$.
    Obtain the prediction loss $\mathcal{L}_{pred}(\boldsymbol{P}, \boldsymbol{Q})$ by (12).
  end for
  $\min_{\phi}\sum_{(\boldsymbol{A},\boldsymbol{C},\boldsymbol{V},\boldsymbol{E})\in\mathcal{B}}\mathcal{L}_{pred}(\boldsymbol{P}, \boldsymbol{Q})$ via Adam and update ϕ.
end for
return MPerformer ϕ and decoder ψ.

Table 1: The statistics of datasets

Type	Dataset	Source	#Molecules	#Atoms	#Bonds with different types					
					Null	Single	Double	Triple	Aromatic	Total
Experimental datasets	Epdbccd	CCD	36,733	938,360	26,392,098	1,101,064	156,928	4,186	746,804	28,401,080
	Subpdb	PDBind	4,352	112,722	3,171,012	136,064	22,420	272	81,708	3,411,476
Idealized datasets	Ipdbccd	CCD	37,709	962,711	27,135,388	1,137,172	161,720	4,150	754,778	29,193,208
	Bradley1	UFF in RDKit	28,279	419,422	6,449,805	434,834	66,022	3,804	364,426	7,318,891
	Bradley2	MMFF in RDKit	27,954	416,148	6,415,902	431,658	65,878	3,794	361,312	7,278,544
	Bradley3	ETKDG in RDKit	28,306	420,590	6,511,123	436,732	66,130	3,804	364,802	7,382,591
	3dqsar	3Dqsar study	1,249	30,105	662,068	27,620	3,880	40	33,400	727,008
	Gdb1k	DFT	1,000	6,359	23,222	8,628	1,410	700	1,066	35,026

* Here, the null bond (i.e., disconnected atom pair) is also considered a type of chemical bond.

4.3 Learning Scheme

Based on the above objectives, we leverage a multi-task pretraining-and-finetuning paradigm to learn MPerformer. In the pretraining phase, given a large 3D molecular dataset (20M), e.g., that used in [61], we add position noise to each atom cloud. Based on the noisy atom clouds, we *i*) predict molecular attributes to construct 3D molecules by minimizing the prediction loss $\mathcal{L}_{pred}$ in (12) and *ii*) reconstruct the original clear atom clouds by minimizing the reconstruction loss $\mathcal{L}_{rec}$ in (14). In the finetuning phase, we only focus on molecular perception problems of the original clear atom clouds, minimizing corresponding prediction loss with a smaller learning rate. The pretraining step helps to enhance the robustness of MPerformer, and the finetuning step makes the model fit the main goal better. In summary, Algorithm 1 shows the steps of the proposed multi-task pretraining-and-finetuning paradigm clearly.

5 EXPERIMENTS

To evaluate the effectiveness of our proposed MPerformer, we apply it to construct 3D molecules from molecular 3D atom clouds and compare it with state-of-the-art molecular perception methods. Experiments on various molecular datasets consistently demonstrate that MPerformer can accurately construct 3D molecules, significantly outperforming baselines. Then we conduct experiments to evaluate the robustness of MPerformer, proving that MPerformer exhibits excellent robustness to position noise. Besides, we also carry out ablation studies to validate the rationality of our multi-task pretraining-and-finetuning paradigm. Finally, we apply MPerformer to the 3D molecular generation field. The results demonstrate that the quality of generated molecules can be effectively improved by using MPerformer, showing the great application potential of MPerformer in practical downstream scenarios.

5.1 Experimental Setup

5.1.1 Dataset. The chemical component dictionary (CCD) [55] can provide all residues and small molecular components found in Protein Data Bank (PDB) [6]. The molecular components can be categorized into two classes: the **E**xperimental molecules generated through real-world experiments and the **I**dealized molecules generated by computational software. Accordingly, we construct two datasets from the experimental and idealized molecules, called Epdbccd and Ipdbccd, respectively.

Besides, the datasets used in the previous works are also employed in our experiments, including *i*) the Bredley1, Bradley2, and Bradley3 datasets created by RDKit [20] from molecular SMILES data [7] and based on different molecular force fields (UFF, MMFF, ETKDG), *ii*) the drug molecules used in the 3Dqsar study [46], *iii*) the Gdb1k dataset optimized by density functional theory (DFT) [41], and *iv*) the Subpdb dataset sampled from the PDBbind dataset [6]. Like Epdbccd and Ipdbccd, we also categorize the above molecular datasets into Experimental and Idealized datasets. All molecular datasets used in our experiments have been processed to remove hydrogen atoms. More details of datasets are shown in Table 1.

5.1.2 Baselines. We compare MPerformer with two state-of-the-art molecular perception methods: *i*) the most commonly-used rule-based molecular perception method **OpenBabel** [38]. It has been widely integrated into much chemical software such as Avogadro [15], and *ii*) the decision tree-based molecular perception method called **Mamba** [28]. Like MPerformer, these two baselines do not require additional atom connectivity information, leading to a fair comparison. Additionally, they are open-source methods whose results can be reproduced easily.

Besides, we also compare our method with the commercial software **Schrödinger Maestro** [4]. Since this software cannot freely support large-scale numerical experiments, we only test it on some representative molecules and visualize corresponding results.

5.1.3 Metrics. Focusing on molecule-level perception accuracy, we use the following two metrics to evaluate different methods.

- **Overall Molecule-level Perception Accuracy (Overall Acc).** The most fundamental metric for molecular perception is the proportion of molecules in which all molecular attributes, including formal charges, hydrogen numbers, and bond orders, are predicted correctly. However, this metric is only available for our MPerformer and the OpenBabel.
- **Molecule-level Bond Accuracy (Bond Acc).** As aforementioned, most existing molecular perception methods only consider the prediction of the bond order (including the connectivity and chemical bond type between atom pairs).

Table 2: The molecular perception capacity of different molecular perception methods

Metric	Method	Dataset								Challenging set		
		Epdbccd	Subpdb	Ipdbccd	Bradley1	Bradley2	Bradley3	3dqsar	Gdb1k	Molecules	Rings	Time
Bond Acc	OpenBabel	0.553	0.623	0.551	0.819	0.808	0.799	0.761	0.901	0.580	0.748	51s
	Mamba	0.548	0.834	0.817	0.994	0.931	0.980	0.915	0.948	0.628	0.900	135s
	MPerformer	**0.670**	**0.971**	**0.872**	**0.994**	**0.932**	**0.990**	**0.973**	**0.967**	**0.799**	**0.949**	**25s**
Overall Acc	OpenBabel	0.532	0.424	0.529	0.813	0.803	0.794	0.629	0.836	—	—	—
	MPerformer	**0.616**	**0.700**	**0.821**	**0.898**	**0.899**	**0.892**	**0.772**	**0.962**	—	—	—

* Mamba can only predict bond orders.

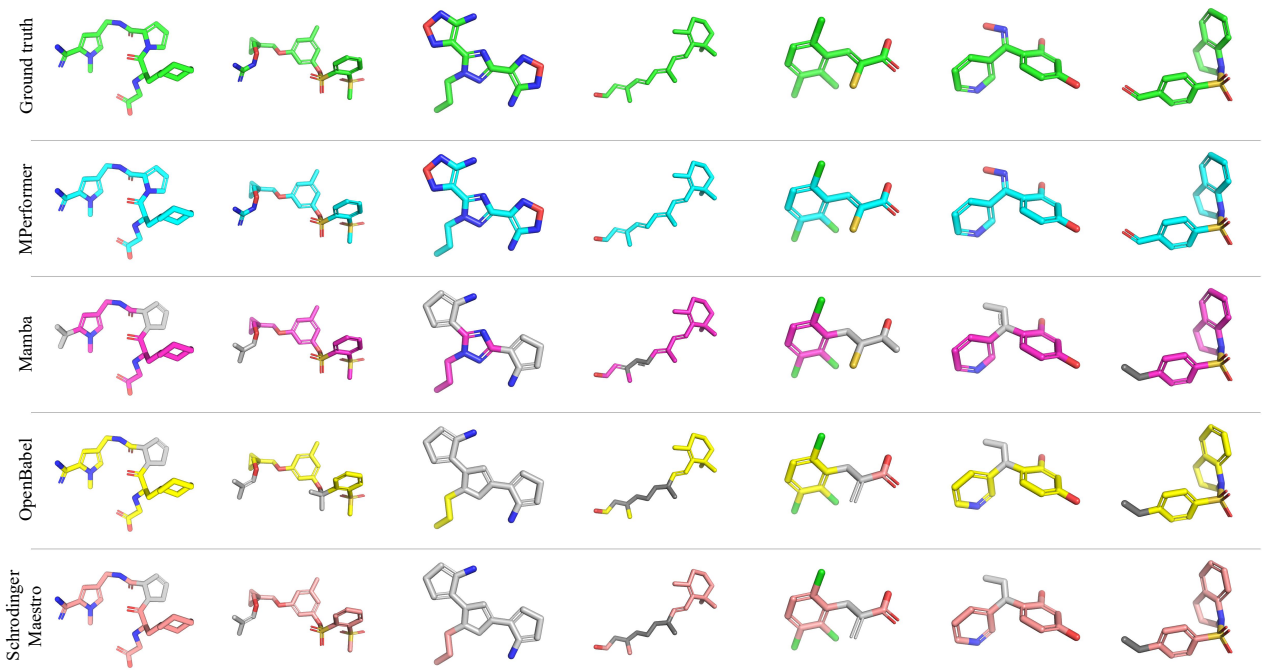


Figure 4: Several molecular perception results of MPerformer, Mamba, OpenBabel, and Schrödinger Maestro on the challenging set, where the wrongly-predicted molecular structures are colored in gray.

Therefore, besides the overall molecule-level perception accuracy, we also take the proportion of molecules in which the bond orders of all the atom pairs are predicted correctly as an evaluation metric.

5.1.4 Model architecture and hyperparameter settings. Our backbone model contains 15 encoding layers and each layer is composed of 64 attention heads. The embedding dimension of $\boldsymbol{X}_a$ and $\boldsymbol{X}_p$ are 512 and 64, respectively. We train our model on the eight Tesla V100 GPUs until convergence with a learning rate of 10^{-3} for pretraining and a learning rate of 10^{-4} for finetuning.

5.2 Comparisons

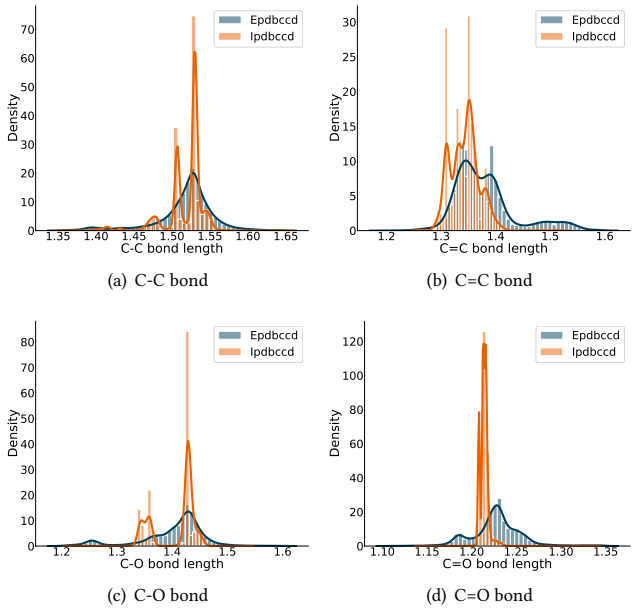
Table 2 presents the molecular perception results achieved by OpenBabel, Mamba, and our MPerformer. Note that Mamba can only predict bond orders, so we only compare it with OpenBabel and MPerformer regarding bond accuracy. In general, our MPerformer achieves the highest bond and overall accuracy consistently on all eight datasets, showing superior performance over other baselines.

To further evaluate the molecular perception capacity and generalization power of our MPerformer, we construct a challenging molecular dataset that contains 800 molecules whose bond orders are difficult to predict in practice. Besides, we train our MPerfomer on the set of all eight datasets and then evaluate it on the challenging molecular dataset.

As shown in Table 2, in addition to the molecule-level bond accuracy, we evaluate our method and the baselines on the ring-level bond accuracy, i.e., the prediction accuracy for the bonds in the rings of the molecules. Our MPerformer outperforms the baselines significantly on the challenging set, whose molecule-level bond accuracy is 0.799 and ring-level bond accuracy is 0.949, respectively.

Table 3: The robustness of different methods under different noises

Metric	Dataset	Method	No noise	Gaussian noise			Uniform noise		
				0.03	0.05	0.07	0.10	0.15	0.20
Bond Acc	Epdbccd	OpenBabel	0.553	0.528	0.466	0.363	0.532	0.492	0.427
		Mamba	0.548	0.443	0.229	0.118	0.452	0.294	0.168
		MPerformer	**0.670**	**0.654**	**0.601**	**0.525**	**0.654**	**0.624**	**0.576**
	Ipdbccd	OpenBabel	0.551	0.533	0.452	0.338	0.537	0.484	0.403
		Mamba	0.817	0.396	0.127	0.060	0.416	0.174	0.085
		MPerformer	**0.872**	**0.862**	**0.851**	**0.825**	**0.860**	**0.851**	**0.840**
Overall Acc	Epdbccd	OpenBabel	0.532	0.507	0.448	0.351	0.511	0.472	0.413
		MPerformer	**0.616**	**0.603**	**0.549**	**0.470**	**0.602**	**0.571**	**0.522**
	Ipdbccd	OpenBabel	0.529	0.510	0.434	0.324	0.514	0.463	0.386
		MPerformer	**0.821**	**0.808**	**0.798**	**0.775**	**0.807**	**0.800**	**0.785**



(a) C-C bond (b) C=C bond

(c) C-O bond (d) C=O bond

Figure 5: Visualizations of the bond length distributions of the Epdbccd and Ipdbccd datasets.

Figure 4 visualizes the results of some typical challenging molecules achieved by different methods, including Schrödinger Maestro [4], further illustrating our MPerformer's superiority.

Note that the experimental results on the Epdbccd dataset are often worse than those on the Ipdbccd dataset because the perception of the molecules achieved by real-world experiments is more challenging in general. This is mainly due to two reasons. Firstly, when determining the chemical information of 3D molecules through various experimental techniques, it is inevitable to be limited by the resolution of testing equipment, which may make the determined chemical information inaccurate. Secondly, the structure of a molecule is inevitably affected by various ions in its surroundings, which often leads to differences between its experimental structure and the corresponding computational ideal structure. Figure 5 shows the distributions of typical chemical bonds' length in the two datasets, visualizing the structural differences.

5.3 Robustness Analysis

Table 3 shows the robustness of different methods to the position noise in atom clouds. It lists the bond accuracy and overall accuracy of various methods on the two representative molecular datasets (the Epdbccd and Ipdbccd datasets) with and without Gaussian and Uniform noise. For each atom, the Gaussian noise imposed on its 3D coordinate is $\epsilon \sim \mathcal{N}(\mathbf{0}_3, \sigma^2 \boldsymbol{I}_3)$, where $\sigma = \{0.03, 0.05, 0.07\}$, respectively. Similarly, when imposing the Uniform noise, $\epsilon \sim \text{Unif}([0, \tau]^3)$, where $\tau = \{0.10, 0.15, 0.20\}$, respectively. Compared to the baselines, our MPerformer has stronger robustness to noise, whose performance is better on both evaluation metrics. Moreover, as the noise size increases, the baselines (especially Mamba) suffer from severe performance degradation, while the performance of MPerformer does not exhibit a significant decline, consistently maintaining high accuracy. Overall, the robustness of MPerformer makes it a promising solution to various application scenarios.

5.4 Ablation Studies

Furthermore, we carry out ablation studies to evaluate the effectiveness and rationality of our multi-task pretraining-and-finetuning learning paradigm. As illustrated in Figure 1 and Algorithm 1, in the multi-task pretraining phase, we not only predict molecular attributes by minimizing the prediction loss $\mathcal{L}_{pred}$ in (12) but also reconstruct the original atom clouds from their noisy versions by minimizing the reconstruction loss $\mathcal{L}_{rec}$ in (14). Besides applying this learning paradigm, we further consider three simplified learning paradigms for our MPerformer: *i*) learning MPerformer without pretraining, *ii*) learning MPerformer with a pretraining phase only minimizing $\mathcal{L}_{rec}$, and *iii*) learning MPerformer with a pretraining phase only minimizing $\mathcal{L}_{pred}$. The comparisons for the four learning paradigms on the noisy Epdbccd and Ipdbccd datasets are shown in Table 4. The results indicate that without the pretraining

Table 4: The robustness capability of MPerformer under different pretraining settings

Dataset	$\mathcal{L}_{pred}$	$\mathcal{L}_{rec}$	Bond Acc						Overall Acc					
			Gaussian noise			Uniform noise			Gaussian noise			Uniform noise		
			0.03	0.05	0.07	0.10	0.15	0.20	0.03	0.05	0.07	0.10	0.15	0.20
Epdbccd	X	X	0.558	0.457	0.338	0.570	0.501	0.410	0.511	0.408	0.291	0.524	0.451	0.355
	X	✓	0.653	0.585	0.474	**0.662**	0.623	0.550	0.596	0.529	0.413	0.604	0.561	0.493
	✓	X	0.651	0.594	0.507	0.655	0.616	0.572	0.600	0.538	0.447	**0.605**	0.561	0.511
	✓	✓	**0.654**	**0.601**	**0.525**	0.654	**0.624**	**0.576**	**0.603**	**0.549**	**0.470**	0.602	**0.571**	**0.522**
Ipdbccd	X	X	0.302	0.089	0.037	0.319	0.127	0.054	0.266	0.061	0.021	0.281	0.098	0.036
	X	✓	0.813	0.401	0.151	0.836	0.539	0.272	0.725	0.326	0.111	0.753	0.447	0.215
	✓	X	0.842	0.531	0.256	0.850	0.652	0.411	0.763	0.447	0.195	0.775	0.564	0.340
	✓	✓	**0.862**	**0.851**	**0.825**	**0.860**	**0.851**	**0.840**	**0.808**	**0.798**	**0.775**	**0.807**	**0.800**	**0.785**

Table 5: The performance of MPerformer in improving the quality of generated molecules

Model	Method	QED (↑)	SA (↑)	LogP
LiGAN [40]	OpenBabel (default)	0.524	0.704	1.490
	Mamba	0.522	0.704	1.495
	MPerformer	**0.527**	**0.705**	1.481
3DSBDD [31]	OpenBabel (default)	0.481	0.670	-0.010
	Mamba	0.492	0.645	0.124
	MPerformer	**0.497**	**0.695**	0.107
GraphBP [24]	OpenBabel (default)	0.461	0.504	3.490
	Mamba	0.461	0.506	3.399
	MPerformer	**0.466**	**0.506**	3.429

step, the performance of our MPerformer degrades seriously, especially in those highly-noisy scenarios. Merely considering one pretraining task leads to sub-optimal perception results in most situations. On the contrary, the proposed multi-task pretraining step can effectively enhance the robustness of MPerformer, maintaining its high accuracy under different noise levels.

5.5 Benefiting 3D Molecular Generation

As we mentioned, many existing 3D molecular generative models [14] generate 3D atom clouds rather than complete 3D molecules. As a result, they need to apply molecular perception methods as their post-processing modules, and currently, OpenBabel [38] is their default choice. Because of its superior performance and strong robustness to position noise, our MPerformer has excellent potential to replace OpenBabel and benefit 3D molecular generation tasks. In this experiment, we first generate atom clouds by applying various 3D molecular generative models [24, 31, 40] under their default settings. Then, we respectively utilize OpenBabel, Mamba, and MPerformer to reconstruct 3D molecules from the atom clouds and evaluate the quality of the generated molecules by their desired properties. In particular, the 3D molecular generative models we considered include LiGAN [40], 3DSBDD [31], and GraphBP [24]. All the methods take OpenBabel as their default post-processing modules. According to the work in [27, 31], we use the following three metrics to measure the quality of molecules: *i*) the **QED** quantitatively assessing the drug-likeness of molecules, *ii*) the **SA** quantitatively assessing the easiness of the synthesis of molecules, and *iii*) the **LogP** quantitatively assessing the octanol-water partition coefficient of molecules. A good-quality molecule should have high QED and SA, and its LogP should be between -0.4 and 5.6.

Table 5 presents the performance of MPerformer in improving the quality of generated molecules. The results indicate that for various 3D molecular generative models, replacing OpenBabel with MPerformer helps to enhance the drug-likeness and synthetic accessibility of generated molecules while keeping the reasonable octanol-water partition coefficient, which benefits many downstream applications like drug discovery.

6 CONCLUSIONS AND FUTURE WORK

We have presented MPerformer, a novel SE(3) Transformer-based molecular perceptron that can construct high-quality 3D molecules with complete chemical information from 3D atom clouds and do not dependent on additional prior knowledge. Experiments demonstrate that MPerformer significantly outperforms all other rules-based and learning-based molecular perception methods and exhibits strong robustness to position noise. Moreover, applying MPerformer helps to improve the quality of the molecules generated by various 3D generative models. In the future, we would like to apply this molecular perception method [1] to practical problems in the field of drug discovery, e.g., structure-based drug design [37] and molecular conformation perception [57].

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (62106271, 92270110), the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. Dr. Hongteng Xu thanks the support from the Beijing Key Laboratory of Big Data Management and Analysis Methods, the Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative.

[1] Code is available at https://github.com/FanmengWang/MPerformer

REFERENCES

[1] Jon C Baber and Edward E Hodgkin. 1992. Automatic assignment of chemical connectivity to organic molecules in the Cambridge Structural Database. *Journal of chemical information and computer sciences* 32, 5 (1992), 401–406.

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[3] Guy Barshatski, Galia Nordon, and Kira Radinsky. 2021. Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3727–3736.

[4] JA Bell, Y Cao, JR Gunn, T Day, E Gallicchio, Z Zhou, R Levy, and R Farid. 2012. PrimeX and the Schrödinger computational chemistry suite of programs. (2012).

[5] Roy Benjamin, Uriel Singer, and Kira Radinsky. 2022. Graph Neural Networks Pretraining Through Inherent Supervision for Molecular Property Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2903–2912.

[6] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. 2000. The protein data bank. *Nucleic acids research* 28, 1 (2000), 235–242.

[7] Jean-Claude Bradley, Andrew Lang, Antony Williams, and Evan Curtin. 2011. ONS Open Melting Point Collection. *Nature Precedings* (2011), 1–1.

[8] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. 2021. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research* 49, D1 (2021), D437–D451.

[9] Ziqi Chen, Martin Renqiang Min, Srinivasan Parthasarathy, and Xia Ning. 2021. A deep generative model for molecule optimization via one fragment modification. *Nature machine intelligence* 3, 12 (2021), 1040–1049.

[10] Anna Katharina Dehof, Alexander Rurainski, Quang Bao Anh Bui, Sebastian Böcker, Hans-Peter Lenhof, and Andreas Hildebrandt. 2011. Automated bond order assignment as an optimization problem. *Bioinformatics* 27, 5 (2011), 619–625.

[11] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* 4, 2 (2022), 127–134.

[12] Jinjia Feng, Zhen Wang, Yaliang Li, Bolin Ding, Zhewei Wei, and Hongteng Xu. 2022. MGMAE: Molecular Representation Learning by Reconstructing Heterogeneous Graphs with A High Mask Ratio. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 509–519.

[13] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. 2020. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems* 33 (2020), 1970–1981.

[14] Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. 2022. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications* 13, 1 (2022), 973.

[15] Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R Hutchison. 2012. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics* 4, 1 (2012), 1–17.

[16] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. 1997. BALI: automatic assignment of bond and atom types for protein ligands in the brookhaven protein databank. *Journal of chemical information and computer sciences* 37, 4 (1997), 774–778.

[17] Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. 2022. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*. PMLR, 8867–8887.

[18] Maria Kadukova and Sergei Grudinin. 2016. Knodle: a support vector machines-based automatic perception of organic molecules from 3D coordinates. *Journal of Chemical Information and Modeling* 56, 8 (2016), 1410–1419.

[19] Paul Labute. 2005. On the perception of molecules from 3D atomic coordinates. *Journal of chemical information and modeling* 45, 2 (2005), 215–221.

[20] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8 (2013).

[21] Elke Lang, Claus-Wilhelm von der Lieth, and Thomas Förster. 1992. Automatic assignment of bond orders based on the analysis of the internal coordinates of molecular structures. *Analytica chimica acta* 265, 2 (1992), 283–289.

[22] Andrew R Leach, Daniel P Dolata, and Keith Prout. 1990. Automated conformational analysis and structure generation: algorithms for molecular perception. *Journal of chemical information and computer sciences* 30, 3 (1990), 316–324.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[24] Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. 2022. Generating 3D Molecules for Target Protein Binding. In *International Conference on Machine Learning*.

[25] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training Molecular Graph Representation with 3D Geometry. In *International Conference on Learning Representations*.

[26] Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2020. *Representation learning for natural language processing*. Springer Nature.

[27] Siyu Long, Yi Zhou, Xinyu Dai, and Hao Zhou. 2022. Zero-Shot 3D Drug Design by Sketching and Generating. In *NeurIPS*.

[28] Christoph Loschen. 2018. Perception of Chemical Bonds via Machine Learning. (2018).

[29] Shuqi Lu, Lin Yao, Xi Chen, Hang Zheng, Di He, and Guolin Ke. 2023. 3D Molecular Generation via Virtual Dynamics. *arXiv preprint arXiv:2302.05847* (2023).

[30] Nicholas Lubbers, Justin S Smith, and Kipton Barros. 2018. Hierarchical modeling of molecular energies using a deep neural network. *The Journal of chemical physics* 148, 24 (2018), 241715.

[31] Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. 2021. A 3D generative model for structure-based drug design. *Advances in Neural Information Processing Systems* 34 (2021), 6229–6239.

[32] Youzhi Luo and Shuiwang Ji. 2022. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations (ICLR)*.

[33] Changsheng Ma, Qiang Yang, Xin Gao, and Xiangliang Zhang. 2022. Disentangled Molecular Graph Generation via an Invertible Flow Model. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1420–1429.

[34] Changsheng Ma and Xiangliang Zhang. 2021. GF-VAE: a flow-based variational autoencoder for molecule generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1181–1190.

[35] Runze Ma, Yidan Zhang, Xinye Wang, Zhenyang Yu, and Lei Duan. 2022. MORN: Molecular Property Prediction Based on Textual-Topological-Spatial Multi-View Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1461–1470.

[36] Sergey Magedov, Christopher Koh, Walter Malone, Nicholas Lubbers, and Benjamin Nebgen. 2021. Bond order predictions using deep neural networks. *Journal of Applied Physics* 129, 6 (2021), 064701.

[37] Rocco Meli. 2022. *Deep learning applications in structure-based drug discovery*. Ph. D. Dissertation. University of Oxford.

[38] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics* 3, 1 (2011), 1–14.

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[40] Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. 2022. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem Sci* 13 (7 Feb 2022), 2701–2713. https://doi.org/10.1039/D1SC05976A

[41] Bharath Ramsundar, V Pande, P Eastman, E Feinberg, J Gomes, K Leswing, A Pappu, and M Wu. 2016. Democratizing deep-learning for drug discovery, quantum chemistry, materials science and biology. *GitHub repository* (2016).

[42] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems* 33 (2020), 12559–12571.

[43] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*. PMLR, 9323–9332.

[44] Yu Shi, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. 2022. Benchmarking graphormer on large-scale molecular modeling datasets. *arXiv preprint arXiv:2203.04810* (2022).

[45] Yuancheng Sun, Yimeng Chen, Weizhi Ma, Wenhao Huang, Kang Liu, Zhiming Ma, Wei-Ying Ma, and Yanyan Lan. 2022. PEMP: Leveraging Physics Properties to Enhance Molecular Property Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3505–3513.

[46] Jeffrey J Sutherland, Lee A O'Brien, and Donald F Weaver. 2004. A comparison of methods for modeling quantitative structure- activity relationships. *Journal of Medicinal Chemistry* 47, 22 (2004), 5541–5554.

[47] Sascha Urbaczek, Adrian Kolodzik, Inken Groth, Stefan Heuser, and Matthias Rarey. 2013. Reading pdb: perception of molecules from 3d atomic coordinates. *Journal of chemical information and modeling* 53, 1 (2013), 76–87.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[49] Jike Wang, Chang-Yu Hsieh, Mingyang Wang, Xiaorui Wang, Zhenxing Wu, Dejun Jiang, Benben Liao, Xujun Zhang, Bo Yang, Qiaojun He, et al. 2021. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence* 3, 10 (2021), 914–922.

[50] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* 25, 2 (2006), 247–260.

[51] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. 2020. A comprehensive survey of loss functions in machine learning. *Annals of Data Science* (2020), 1–26.

[52] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. 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*. 429–436.

[53] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* 4, 3 (2022), 279–287.

[54] Ivan D Welsh and Jane R Allison. 2019. Automated simultaneous assignment of bond orders and formal charges. *Journal of Cheminformatics* 11, 1 (2019), 1–12.

[55] John D Westbrook, Chenghua Shao, Zukang Feng, Marina Zhuravleva, Sameer Velankar, and Jasmine Young. 2015. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* 31, 8 (2015), 1274–1278.

[56] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.

[57] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=PzcvxEMzvQC

[58] Jinliang Yuan, Hualei Yu, Meng Cao, Ming Xu, Junyuan Xie, and Chongjun Wang. 2021. Semi-supervised and self-supervised classification with multi-view graph neural networks. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 2466–2476.

[59] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (2021), 5586–5609.

[60] Yuan Zhao, Tiejun Cheng, and Renxiao Wang. 2007. Automatic perception of organic molecules based on essential structural information. *Journal of chemical information and modeling* 47, 4 (2007), 1379–1385.

[61] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*.

[62] Xinyu Zhu, Yongliang Shen, and Weiming Lu. 2022. Molecular substructure-aware network for drug-drug interaction prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4757–4761.