MOTIF-BASED ROTO-TRANSLATION INVARIANT TRANSFORMER FOR MOLECULAR PROPERTY PREDICTION IN 3D SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies use geometric deep learning to represent molecules and predict properties. However, they are computationally expensive in capturing long-range dependencies and ignore the non-uniformity of interatomic distances. More importantly, few of them consider injecting the biochemical structure knowledge such as functional groups into model architectures. To overcome such issues, we introduce Molformer, a variant of the Transformer for molecular representations that exploits both semantic motifs and 3D spatial information. Specifically, Molformer extracts motifs based on functional groups and learns customized embeddings to store the semantic meanings of those informative substructures. In order to fully employ 3D geometry, we adopt a convolutional position encoding to achieve roto-translation invariance, a multi-scale self-attention mechanism to capture local fine-grained patterns with increasing contextual scales, and an attentive farthest point sampling algorithm to attain the molecular representation. We validate Molformer across several domains in quantum chemistry, physiology, and biophysics. Our experiments show better or competitive performance in those datasets. Our work provides a promising way to amalgamate 3D geometric information and make better usage of informative substructures in representing molecules.

1 INTRODUCTION

Spatial structures are among the most crucial factors to decide molecular properties and understand their principles of action in the physical world. For example, 3D structures of proteins provide valuable information for inferring biological interventions, such as structure-based drug development and targeted mutagenesis (Senior et al., 2020; Jumper et al., 2021; Baek et al., 2021). In chemistry, zeolites show obvious differences in separation properties caused by subtle changes in their 3D geometric compositions (Chai et al., 2020; Pfriem et al., 2021). Apart from that, in the pharmaceutical industry, the same compounds can have different 3D structures, resulting in different solubility (Zhang et al., 2017). To sum up, capturing 3D spatial structures is essential to accurately forecast molecular properties. Based on these facts, researchers have studied molecular representation learning techniques (Rao et al., 2019) to include 3D spatial information (Zhavoronkov et al., 2019).

The dominant 3D molecular models are Graph Neural Networks (GNNs) and 3D Convolutional Neural Networks (3DCNNs) (Derevyanko et al., 2018; Pagès et al., 2019; Townshend et al., 2019). GNNs create edges by using either chemical bonds or finding the neighbors of each node within a distance cutoff (Zhang et al., 2020b). They encode pairwise connectivity of atoms and require running multiple hops for an atom to reach to another. 3DCNNs encode translational and permutational symmetries, but need to stack deep layers to build direct connections between distant regions, incurring significant computational costs. In contrast, Transformers rely on the self-attention mechanism to capture long-term dependencies in parallel (Hernández & Amigó, 2021). Meanwhile, Equivariant Neural Networks (ENNs) (Thomas et al., 2018) have emerged as a new class of methods, where geometric transformations of their inputs lead to well-defined transformations of outputs. Some ENNs adopt Transformers as the backbone but fail to surmount the intrinsic drawbacks of this architecture, including its insensibility to local patterns among non-uniformly distancing atoms and its ineffi-

ciency to aggregate atom features. Some other Transformer-based methods have been proposed to fuse distance and graph neighbourhood information (Maziarka et al., 2020; 2021). However, they take no consideration of employing motifs, which are frequently-occurring substructures in molecules and can be leveraged to uncover global graph properties.

In this work, we present the Molformer on the basis of all preceding analysis. For the sake of injecting chemical domain knowledge, we construct a motif-template vocabulary based on functional groups and adopt trainable motif embeddings to maintain the semantic meanings of those essential substructures. Then with both motifs and atoms as input, Molformer operates on a fully-connected graph with direct connections between remote regions (Veličković et al., 2017; Joshi, 2020), which reduces computational burden of multi-hop GNNs and stacked 3DCNNs. However, this characteristic limits Molformer's capacity in exploiting local structures and leads to poor generalization in unseen cases (Qi et al., 2017). Therefore, we propose a Multi-scale Self-Attention (MSA) module to recognize fine-grained patterns from neighborhoods. Moreover, we introduce a roto-translation invariant Convolutional Position Encoding (CPE) to depict position relationships among atoms and their adjacencies. After that, to retain a comprehensive representation of the entire molecule, we propose an Attentive Farthest Point Sampling (AFPS) module that selects important atoms with the assistance of the attention score map.

To summarize, our contributions are as follows:

- To the best of our knowledge, we are the foremost to incorporate motifs with knowledge of functional groups into a Transformer architecture for 3D molecular representation learning.
- We propose a novel MSA to extract local patterns, a roto-translation invariant CPE method to encode relative distance at a linear computational time cost, and a simple yet effective downsampling algorithm to gather molecular representations.
- We show significant improvements on several benchmarks in three domains. Code and all datasets are available at `https://github.com/smiles724/Molformer`.

2 PRELIMINARIES

Problem Definition. A molecule $\boldsymbol{S} = (\boldsymbol{E}, \boldsymbol{P})$ has N atoms and C atom classes, where $\boldsymbol{E} = \{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\} \in \mathbb{R}^{N \times C}$ contains the one-hot atom representations and $\boldsymbol{P} = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_N\} \in \mathbb{R}^{N \times 3}$ contains the 3D coordinates of each atom. Each one-hot $\boldsymbol{e}_i$ can be converted to a dense vector $\boldsymbol{x}_i = \boldsymbol{e}_i \boldsymbol{W}^E$, with $\boldsymbol{x}_i \in \mathbb{R}^{d_{model}}$ and $\boldsymbol{W}^E \in \mathbb{R}^{C \times d_{model}}$ being the embedding matrix. The 3D coordinates of the atom i is a three-dimensional vector $\boldsymbol{p}_i = [p_i^x, p_i^y, p_i^z]$. A representation learning model f acts on $\boldsymbol{S}$, obtaining its representation $\boldsymbol{r} = f(\boldsymbol{S})$. Then $\boldsymbol{r}$ is forwarded to a prediction model g and attain the prediction of a biochemical property $\hat{y} = g(\boldsymbol{r})$.

Self-attention Mechanism. The Transformer (Vaswani et al., 2017) has become very successful due to its core component, self-attention. Given a set of input features $\{\boldsymbol{x}_i\}_{i=1,...,N}$, the standard dot-product attention layer is as the following:

$$\boldsymbol{q}_i = f_Q(\boldsymbol{x}_i),\ \boldsymbol{k}_i = f_K(\boldsymbol{x}_i),\ \boldsymbol{v}_i = f_V(\boldsymbol{x}_i),\ a_{ij} = \boldsymbol{q}_i \boldsymbol{k}_j^T / \sqrt{d_k},\ \boldsymbol{z}_i = \sum_{j=1}^{N} \sigma(a_{ij}) \boldsymbol{v}_j \tag{1}$$

where $\{f_Q, f_K, f_V\}$ are embedding transformations, and $\{\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i\}$ are respectively the query, key, and value vectors with the same dimension d_k. a_{ij} is the attention that the token i pays to the token j. σ denotes the *Softmax* function and $\boldsymbol{z}_i$ is the output embedding of the token i. This formula conforms to a non-local network (Wang et al., 2018), indicating its inability to capture fine-grained patterns in a local context.

Position Encoding. Self-attention is invariant to permutation of the input (Dufter et al., 2021), and position encoding ensures that the Transformer will reveal positional information. Position encoding methods can be either based on absolute positions or relative distances. The former takes the raw position information as input and is sensitive to spatial transformations. The latter manipulates the attention score by incorporating relative distances (Guo et al., 2020a; Pan et al., 2021): $a_{ij} = \boldsymbol{q}_i \boldsymbol{k}_j^T / \sqrt{d_k} + f_{\mathrm{PE}}(\boldsymbol{p}_i - \boldsymbol{p}_j)$, where $f_{\mathrm{PE}}(\cdot)$ is the position encoding function and is translation

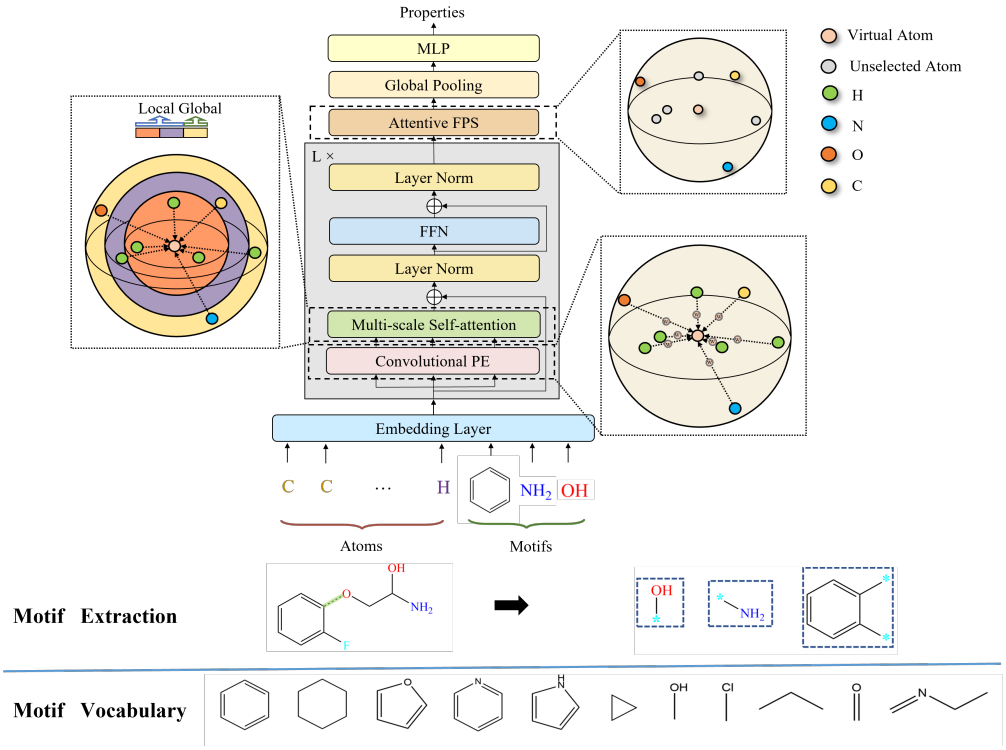

Figure 1: The overall architecture of our Molformer. FFN stands for a feed-forward network. Local features are shown in purple and orange; yellow corresponds to a global feature.

invariant. The rotation invariance can be further accomplished by taking a L2-norm $||\boldsymbol{p}_i - \boldsymbol{p}_j||_2$ (Chen et al., 2019b).

3 MOLFORMER

Molformer is based on the architecture of Transformer but adopts several significantly different and novel components (see Figure 1). First, a vocabulary of motif templates is constructed on the basis of functional groups and we extract all available motifs from each molecule. Then both atoms and motifs acquire their corresponding embeddings and are forwarded into L feature learning blocks. Each block consists of a convolutional position encoding, a multi-scale self-attention, and a feed-forward network. After that, an attentive subsampling method is utilized to adaptively aggregate the molecular presentation, which is later fed into a predictor to forecast properties in a broad range of downstream tasks.

3.1 TRAINABLE MOTIF-BASED EMBEDDING

Motifs are frequently-occurring substructure patterns as well as the building blocks of complex molecular structures. They usually maintain semantic meanings and have great expressiveness of the biochemical characteristics of the whole molecule (Zhang et al., 2020a). In the chemical community, researchers have developed a set of standard criterion to recognize motifs with essential functionalities in molecules (Milo et al., 2002). Despite that, few of prior studies directly incorporate those informative motifs into their model architectures. To fill this gap, we define a series of momentous substructures using external domain knowledge, and introduce a trainable motif embeddings method to fully exploit them in our Molformer.

To begin with, all motifs are first extracted according to the motif vocabulary, which is built by functional groups. Practically, we rely on RDKit (Landrum, 2013) to draw them from the SMILES (Weininger, 1988) representation of each molecule. We assume M motifs $\{m_1, ..., m_M\}$

are detected in the molecule $\boldsymbol{S}$, and each motif m_i contains a certain number of at least two atoms. Then we regard each kind of motif as a new type of token and append them to the input. Therefore, the input for our Molformer becomes $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N, \boldsymbol{x}_{m_1}, ..., \boldsymbol{x}_{m_M}\}$, where $\boldsymbol{x}_{m_i}$ is obtained through an learnable embedding matrix $\boldsymbol{W}^M \in \mathbb{R}^{C' \times d_{model}}$ and C' denotes the number of motif categories. As for the position of each motif, we adopt a weighted sum of the 3D coordinates of its component atoms as $\boldsymbol{p}_{m_i} = \sum_{x_i \in m_i} (\frac{w_i}{\sum_{x_i \in m_i} w_i}) \cdot \boldsymbol{p}_i$, where w_i are the atomic weights.

Our approach requires the model to automatically learn a customized embedding for each motif template through backpropagations, which follows a data-driven pattern. In some data-sufficient tasks, its greatest potential can be unlocked and those motif embeddings can be well trained. Nevertheless, in the case of few-shot learning or small datasets, each category of motif might only appear rare times. Those embeddings are not fully tuned and can be extremely biased and noisy, which will do little helps to the ultimate property prediction.

3.2 Convolutional Position Encoding

To enable roto-translation invariance and take fully advantage of geometric information, instead of adding a term of $f_{\text{PE}}(\boldsymbol{p}_i - \boldsymbol{p}_j)$, we propose a CPE that applies a convolutional operation to the interatomic distance $\boldsymbol{D} \in \mathbb{R}^{N \times N}$:

$$\boldsymbol{A}_{\text{cov}} = \text{Conv}_{2d}(\boldsymbol{D}) \odot \boldsymbol{A}, \tag{2}$$

where $\boldsymbol{A} = [a_{i,j}]_{i,j=1,\cdots N} \in \mathbb{R}^{N \times N}$ is the attention matrix, $\text{Conv}_{2d}(\cdot)$ denotes a 2D shallow convolutional network with a kernel size of 1×1, and $\odot$ is the element-wise product. With multi-headed self-attention, $\boldsymbol{A}_{\text{cov}}$ is expanded in the sense that $\boldsymbol{A}_{\text{cov}} \in \mathbb{R}^{H \times N \times N}$, and $\text{Conv}_{2d}(\cdot)$ has H output channels. The CPE method induces $\text{O}(N)$ convolution operations on each atom and can drastically reduce training time when the number of atoms is very large (Wu et al., 2021).

3.3 Multi-scale Self-attention

The self-attention mechanism in the Transformer is good at capturing global data patterns but ignores local context (Guo et al., 2020a). Exploiting local context has proven to be important for 3D spatial data such as 3D point clouds (Qi et al., 2017). Therefore, we impose a distance-based constraint in self-attention in order to extract multi-scaled patterns from both local and global contexts.

Guo et al. (2020b) propose to use integer-based distance to limit attention to local word neighbors, which cannot be used in molecules. This is because different types of molecules have different densities and molecules of the same type have different spatial regularity, which results in the non-uniformity of interatomic distances. Normally, small molecules have a mean interatomic distance of 1-2 Å (Angstrom, $10^{-10}m$), which is denser than large molecules like proteins with approximately 5 Å on average. To address that, we design a new multi-scale methodology to robustly capture details. Specifically, we mask atoms beyond a certain distance τ_s (a real number as opposed to an integer in Guo et al. (2020b)) at each scale s. We denote $d_{ij} = ||\boldsymbol{p}_i - \boldsymbol{p}_j||_2$ as the Euclidean distance between the i-th and j-th atom. The attention calculation is modified as:

$$a_{ij}^{\tau_s} = \frac{\boldsymbol{q}_i \boldsymbol{k}_j^T \cdot \mathbf{1}_{\{d_{ij} < \tau_s\}}}{\sqrt{d_k}}, \; \boldsymbol{z}_i^{\tau_s} = \sum_{j=1}^{N} \sigma(a_{ij}^{\tau_s}) \boldsymbol{v}_j, \tag{3}$$

where $\mathbf{1}_{\{d_{ij} < \tau_s\}}$ is the indicator function. For small molecules, Equation 3 can be complementally combined with Equation 2. Then features extracted from S different scales $\{\tau_s\}_{s=1,...,S}$ as well as the informative global feature are concatenated together to form a multi-scale representation, denoted by $\boldsymbol{z}_i' = \boldsymbol{z}_i^{\tau_1} \oplus ... \oplus \boldsymbol{z}_i^{\tau_S} \oplus \boldsymbol{z}_i^{global} \in \mathbb{R}^{(S+1)d_k}$. After that, $\boldsymbol{z}_i'$ is forwarded into a multi-layer perceptron to be compressed as $\boldsymbol{z}_i''$ with the original dimension d_k.

3.4 Attentive Farthest Point Sampling

After having the atom embeddings $\{\boldsymbol{z}_i''\}_{i=1,...,N}$, we study how to obtain the molecular representation $\boldsymbol{r}$. For GNNs, several readout functions such as set2set (Vinyals et al., 2015) and GG-NN (Gilmer et al., 2017) are invented. For Transformer architectures, one way is via a virtual atom.

Algorithm 1 Pseudocode of Attentive Farthest Point Sampling

Input: The attention score matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, a Euclidean distance matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$.
Output: K sampled points.

$\tilde{\boldsymbol{A}} \leftarrow \sum_i \boldsymbol{A}_{ij} \in \mathbb{R}^N$ ▷ sum up the attention matrix along rows
$\tilde{D} \leftarrow D \in \mathbb{R}^{N\times N}$ ▷ normalize the distance matrix
$\mathcal{P} = \{x_\#\}$, $\mathcal{M} = \{1, 2, ..., N\}$
while $\text{length}(\mathcal{P}) < k$ **do**
    $x_{new} = \underset{i\in\mathcal{M}}{\text{argmax}} (\min_{j\in\mathcal{P}} \tilde{\boldsymbol{D}}_{ij} + \epsilon\tilde{\boldsymbol{A}}_i)$ ▷ pick up the atom that maximize the objective
    $\mathcal{P}$.append(x_{new}), $\mathcal{M}$.remove(x_{new})
return $\mathcal{P}$

Though as Ying et al. (2021) state, it significantly improves the performance of existing models in the leaderboard of Open Graph Benchmark (Hu et al., 2020), this way concentrates more on close adjacent atoms and less on distant ones, and can lead to inadvertent over-smoothing of information propagation (Ishiguro et al., 2019). Besides, it is difficult to locate a virtual node in 3D space and build connections to existing atoms. The other way selects a subset of atoms via a downsampling algorithm named Farthest Point Search (FPS), but it ignores atomic differences and has sensitivity to outlier points (Pan et al., 2021) as well as uncontrollable randomness. To address these issues, we propose a new algorithm named AFPS. It aims to sample atoms by not merely spatial distances, but also their significance in terms of attention scores.

Specifically, we choose the virtual atom $x_\#$ as the starting point and initialize two lists $\mathcal{P} = \{x_\#\}$ and $\mathcal{M} = \{1, ..., N\}$ to store remaining candidate points. Then the process begins with the attention score matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and the interatomic distance matrix $\boldsymbol{D} \in \mathbb{R}^{N\times N}$. It can be easily proved that each row of $\boldsymbol{A}$ sums up to 1 after the $Softmax$ operation along columns, i.e. $\sum_j \boldsymbol{A}_{ij} = 1$ for $\forall i \in [N]$. In order to obtain the importance of each atom in the self-attention computation, we accumulate $\boldsymbol{A}$ along rows and get $\tilde{\boldsymbol{A}} = \sum_i \boldsymbol{A}_{ij} \in \mathbb{R}^N$. Besides, we adopt the min-max normalization to rescale the distance matrix $\boldsymbol{D}$ into values between 0 and 1, and obtain $\tilde{\boldsymbol{D}} = \frac{\boldsymbol{D}-\min \boldsymbol{D}}{\max \boldsymbol{D}-\min \boldsymbol{D}}$.

After the above preprocess, we repeatedly move a point x_{new} from $\mathcal{M}$ to $\mathcal{P}$, which ensures that x_{new} is as far from $\mathcal{P}$ as possible by maximizing $\tilde{\boldsymbol{D}}_{ij}$ and also plays a crucial role in attention computation by maximizing $\tilde{\boldsymbol{A}}_i$. Mathematically, the AFPS aims to achieve the following objective:

$$\max \sum_{i\in\mathcal{M}} (\min_{j\in\mathcal{P}\setminus\{i\}} \tilde{\boldsymbol{D}}_{ij} + \epsilon\tilde{\boldsymbol{A}}_i) \tag{4}$$

where ϵ is a hyperparameter to balance those two different goals. This process is repeated until $\mathcal{P}$ has reached K points. Algorithm 1 provides a greedy approximation solution to solve this AFPS optimization objective for sake of computational efficiency.

After that, sampled features $\{\boldsymbol{z}_i''\}_{i\in P}$ are gathered by a Global Average Pooling layer (Lin et al., 2013) to attain the molecular representation $\boldsymbol{r} \in \mathbb{R}^{d_k}$.

Remarkably, our proposed AFPS has considerable difference and superiority over a body of previous hierarchical learning approaches (Eismann et al., 2020; 2021). Their subsampling operations are mainly designed for protein complexities, which have more uniform structures than small molecules. To be specific, they hierarchically use alpha carbons as the intermediate set of points and aggregate information at the level of those carbons for the entire complex. However, the structures of small molecules have no such a stable paradigm, and we provide a universal methodology to adaptively subsample atoms without any prior assumptions on the atom arrangement.

4 Experiments

4.1 Experimental Setup

We conduct extensive experiments on both small and large molecules (proteins) with various targets, including quantum chemistry, physiology, and biophysics. Table 1 summarises information of

benchmark datasets, such as the number of tasks and task types, the number of molecules and atom classes, the minimum and maximum number of atoms, and the density (mean interatomic distances) of all molecules.

Table 1: Key statistics of datasets from three different categories.

Category	Dataset	Tasks	Task Type	Molecules	Atom Class	Min. Atoms	Max. Atoms	Density (Å)	Metric
Quantum Chemistry	QM7	1	regression	7,160	5	4	23	2.91	MAE
	QM8	12	regression	21,786	5	3	26	1.54	MAE
	QM9	12	regression	133,885	5	3	28	1.61	MAE
Physiology	BBBP	1	classification	2,039	13	2	132	2.64	ROC-AUC
	ClinTox	2	classification	1,478	27	1	136	2.83	ROC-AUC
Biophysics	PDBind[1]	1	regression	11,908	23	115	1,085	5.89	RMSE
	BACE	1	classification	1,513	8	10	73	3.24	ROC-AUC

Datasets. We test Molformer on a series of small molecule datasets, containing QM7 (Blum & Reymond, 2009), QM8 (Ramakrishnan et al., 2015), QM9 (Ramakrishnan et al., 2014), BBBP (Martins et al., 2012), ClinTox (Gayvert et al., 2016), and BACE (Subramanian et al., 2016) [2]. QM7 is a subset of GDB-13 and composed of 7K molecules with up to 5 heavy atom types. QM8 and QM9 are subsets of GDB-17 with 22k molecules and 133K molecule respectively.

Additionally, we also inspect Molformer's ability of learning mutual relations between proteins and molecules on the PDBbind dataset (Wang et al., 2005). We follow Townshend et al. (2020) and split protein-ligand complexes by protein sequence identity at 30%. As for the target, we predict $pS = -\log(S)$, where S is the binding affinity in Molar unit. In addition, we only use the pocket of each protein and put pocket-ligand pairs together as the input.

For QM9, we use the exact train/validation/test split as Townshend et al. (2020). For PDBbind, 90% of the data is used for training and the rest is divided equally between validation and test like Chen et al. (2019c). For others, we adopt the scaffold splitting method with a ratio of 8:1:1 for train/validation/test as Rong et al. (2020). More implementing details can be found in Appendix A.1

Baselines For small molecules, we compare our approach with a number of state-of-the-art baselines. TF_Robust (Ramsundar et al., 2015) takes molecular fingerprints as the input. GraphConv (Kipf & Welling, 2016), Weave (Kearnes et al., 2016), MPNN (Gilmer et al., 2017), Schnet (Schütt et al., 2018), MEGNet (Chen et al., 2019c), DMPNN (Yang et al., 2019), MGCN (Lu et al., 2019), AttentiveFP (Xiong et al., 2019), DimeNet++ (Klicpera et al., 2020), SphereNet (Liu et al., 2021), and SpinConv (Shuaibi et al., 2021) are all graph convolutional models. Graph Transformer (Chen et al., 2019a), MAT (Maziarka et al., 2020), R-MAT (Maziarka et al., 2021), SE(3)-Transformer (Fuchs et al., 2020), and LieTransformer (Hutchinson et al., 2021) are Transformer-based models.

For PDBbind, we choose six baselines. DeepDTA (Öztürk et al., 2018) and DeepAffinity (Karimi et al., 2019) take in pairs of ligand and protein SMILES as input. Cormorant (Anderson et al., 2019) is an ENN that represents each atom by its absolute 3D coordinates. Schnet, 3DCNN and 3DGCN (Townshend et al., 2020) are 3D molecular representation methods.

4.2 Results on Downstream Tasks

Molecules. Table 2 and Table 3 document the overall results of Molformer and baselines on small molecules datasets, where best performance is marked bold and the second best is underlined for clear comparison. It can be discovered that Molformer achieves the lowest MAE of 11.6 on QM7 and 0.009 on QM8, beating several strong baselines including DMPNN and Graph Transformer. While

[1]The total number of proteins in the full, unsplit PDBbind is 11K, but our experiment only uses 4K proteins at 30% sequence identity. Moreover, the number of atoms is the sum of both the pocket and molecules.

[2]For BBBP, ClinTox, and BACE, we use RDKit (Landrum, 2013) to procure 3D coordinates from SMILES.

not all state-of-the-art on QM9, Molformer offers competitive performance in 5 property regression tasks, which do not require thermochemical energy subtractions. Particularly, we outperforms all Transformer-based ENNs, including SE(3)-Transformer and LieTransformer. In classification problems, we surpass all non-pretrained methods and are only inferior to the pretrained GROVE. This accords to the fact that datasets with fewer samples can gain large improvements through the self-supervised pretraining (Rong et al., 2020).

Table 2: The performance comparison. For regression tasks including QM7 and QM8, lower is better. For classification tasks including BBBP, ClinTox, and Bace, higher is better. The methods in purple are pretrained methods.

Method	QM7	QM8	BBBP	ClinTox	BACE
TF-Robust (Ramsundar et al., 2015)	120.6	0.024	0.860	0.765	0.824
GraphConv (Kipf & Welling, 2016)	118.9	0.021	0.877	0.845	0.854
Weave (Kearnes et al., 2016)	94.7	0.022	0.837	0.823	0.791
MPNN (Gilmer et al., 2017)	113.0	0.015	0.913	0.879	0.815
Schnet (Schütt et al., 2018)	74.2	0.020	0.847	0.717	0.750
DMPNN (Yang et al., 2019)	105.8	0.014	0.919	0.897	0.852
MGCN (Lu et al., 2019)	77.6	0.022	0.850	0.634	0.734
Attentive FP (Xiong et al., 2019)	126.7	0.028	0.908	0.933	0.863
Graph Transformer (Chen et al., 2019a)	47.8	0.010	0.913	-	0.880
MAT (Maziarka et al., 2020)	102.8	-	0.728	-	0.846
R-MAT (Maziarka et al., 2021)	68.6	-	0.746	-	0.871
$\text{GROVE}_{\text{large}}$ (Rong et al., 2020)	89.4	0.017	0.911	0.884	0.858
$\text{GROVE}_{\text{large}}$ (Rong et al., 2020)	72.6	0.012	**0.940**	**0.944**	**0.894**
Molformer	**11.5**	**0.009**	0.926	0.941	0.884

Table 3: Comparison of MAE on QM9. The methods in orange are Transformer-based methods.

Target (Unit)	ϵ_{HOMO} (eV)	ϵ_{LUMO} (eV)	$\Delta\epsilon$ (eV)	μ (D)	α (bohr^3)
MPNN (Gilmer et al., 2017)	.043	.037	.069	.030	.092
Schnet (Schütt et al., 2018)	.041	.034	.063	.033	.235
$\text{MEGNet}_{\text{full}}$ (Chen et al., 2019c)	.038	.031	.061	.040	.083
DimeNet++ (Klicpera et al., 2020)	.024	.019	.032	.029	**.043**
SphereNet (Liu et al., 2021)	.024	**.019**	**.032**	**.026**	.047
SpinConv (Shuaibi et al., 2021)	.026	.022	.047	.027	.058
SE(3)-Transformer (Fuchs et al., 2020)	.035	.033	.053	.051	.142
LieTransformer-SE(3) (Hutchinson et al., 2021)	.033	.029	.052	.061	.104
Molformer	**.021**	.026	.039	.045	.086

Protein. Table 4 reports the Root-Mean-Squared Deviation (RMSD), the Pearson correlation (R_p), and the Spearman correlation (R_s) on PDBbind. Molformer achieves the lowest RMSD among all baselines and the best Pearson and Spearman correlations. As Wu et al. (2018) claim, appropriate featurizations which contains pertinent information is significant for PDBbind. However, an important observation in our work is that deep learning approaches with the full exploitation of 3D geometric information can perform better than conventional methods like DeepDTA and DeepAffinity, which use a set of physicochemical descriptors but ignore 3D structures.

5 Ablation Study and Discussion

5.1 What are the Effects of Each Component

We investigate the effectiveness of different modules of our Molformer in Table 5. It can be observed that CPE substantially boosts model's performance compared with the naive method that immediately adds 3D coordinates as the atom input feature. In addition, AFPS is found to produce better

Table 4: Comparison of RMSD, R_p, and R_s on PDBbind.

Method	Geometry	RMSD	R_p	R_s
DeepDTA (Öztürk et al., 2018)	Non-3D	1.565	0.573	0.574
DeepAffinity (Karimi et al., 2019)	Non-3D	1.893	0.415	0.426
Schnet (Schütt et al., 2018)	3D	1.892	0.601	-
Cormorant (Anderson et al., 2019)	3D	1.429	0.541	0.532
3DCNN (Townshend et al., 2020)	3D	1.520	0.558	0.556
3DGCN (Townshend et al., 2020)	3D	1.963	0.581	0.647
Molformer	3D	**1.417**	**0.623**	**0.651**

predictions than the control group, which utilizes the virtual node as the molecular representation. Moreover, MSA significantly reduces RMSD from 17.6 to 11.6 on QM7, but its improvements in QM8 are much smaller. This phenomenon indicates that MSA is an appropriate way to alleviate the problem of inadequate training in small datasets. It endows Molformer with capability to extract local features by regulating the scope of self-attention. However, as the data size gets larger and larger, Molformer does not require the assistance of MSA to abstract local patterns, since the parameters of CPE is properly trained. What's more, the trainable motif-level embedding leads to a MAE decrease of 2.1 in QM7 and a RMSD drop of 0.011 in PDBbind, indicating its effectiveness in both small molecules and proteins.

Table 5: Effects of each module on QM7, QM8 and PDBbind (RMSD). ME stands for the trainable motif embedding method.

	CPE	AFPS	MSA	ME	QM7	QM8	PDBbind
1	-	-	-	-	63.2	0.0205	1.925
2	✓	-	-	-	17.6	0.0104	1.489
3	✓	✓	-	-	17.0	0.0103	1.455
4	✓	-	✓	-	11.6	**0.0098**	1.423
5	✓	-	-	✓	15.2	-	1.443
7	✓	✓	✓	-	13.7	0.0099	1.428
6	✓	✓	✓	✓	**11.5**	-	**1.417**

5.2 How Useful is the Trainable Motif-based Embeddings?

How to determine motifs are critical and crucial to our proposed trainable motif-based embeddings. In organic chemistry, a functional group is a substituent or moiety in a molecule that causes the molecule's characteristic chemical reactions. The same functional group will undergo the same or similar chemical reactions regardless of the rest of the molecule's composition (Smith, 2020). Therefore, we define motifs on the basis of functional groups and explore the contribution of four different categories. Specifically, we consider four common functional groups, including groups that contain only carbon and hydrogen (Hydrocarbons), groups that contain halogen (Haloalkanes), groups that contain oxygen, and groups that contain nitrogen (see the left part in Figure 2). The ablations (see the right part in Figure 2) demonstrate that Molformer can gain improvements from all sorts of motifs, where Hydrocarbons and Haloalkanes are the most and the least effective kinds, respectively. This is in line with the fact that Hydrocarbons occur most frequently in organic molecules. Moreover, our model achieves the best performance when all categories of the motifs are integrated, implying a promising direction to discover more effective motifs.

6 Related Works

6.1 3D Molecular Representation

Deep learning has been widely applied to predict molecular properties during past decades. Small molecules are usually represented as lower-dimensional representations such as 1D linear sequence,

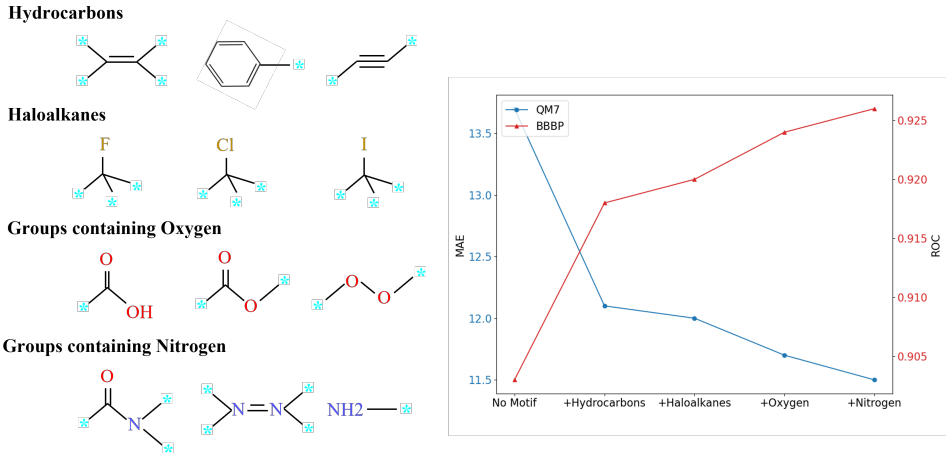

Figure 2: The left is the four different categories of motifs that we apply in Molformer based on functional groups. The right is the ablation study of those groups in QM7 and BBBP.

including amino acid sequences and SMILES (Weininger, 1988), or 2D chemical bond graphs. In spite of that, more evidence indicates that 3D space structures lead to better modelling and superior performance. 3D models becomes a popular way to capture these complex geometries in a variety of bio-molecular applications using CNNs (Anand-Achim et al., 2021; Jiménez et al., 2018) and GNNs (Cho & Choi, 2018). Nonetheless, aforementioned methods have hardly been extended to the self-attention mechanism that is proven to be good at grabbing contextual feature (Tang et al., 2018) and long-range dependencies (Vaswani et al., 2017).

Attempts have been undertaken to address that issue throughout Transformers. Initially, molecules are in the form of SMILES to obtain corresponding representations (Honda et al., 2019; Pesciullesi et al., 2020; Morris et al., 2020; Rao et al., 2021) and conduct pretraining (Chithrananda et al., 2020). Some researchers combine the characteristics of GNN and Transformer to solve generative tasks (Ingraham et al., 2019) or fulfill equivariance (Fuchs et al., 2020).

6.2 Motif-based Method

Motifs have been proven to benefit many tasks from exploratory analysis to transfer learning (Henderson et al., 2012). Various algorithms have been proposed to exploit motifs for contrastive learning (Zhang et al., 2020a), self-supervised pretraining (Rong et al., 2020; Zhang et al., 2021), and generation (Jin et al., 2020). However, none of previous work tries to embody those informative motifs in their model architectures.

7 Conclusion

In this study, we present a universal neural architecture, Molformer, for 3D molecular representations. Our model extracts motifs with semantic meanings from each molecule based on functional groups and learn customized embeddings to facilitate property predictions. Moreover, it adopts a convolutional position encoding method to make a full use of spatial information and augments the self-attention mechanism with multiplicate scales to catch local features. Furthermore, a simple but efficient downsampling algorithm is introduced to better accumulate representations of an entire molecule. Our experiments show the superiority of our model on various scientific domains.

References

Namrata Anand-Achim, Raphael Ryuichi Eguchi, Irimpan I Mathews, Carla P Perez, Alexander Derry, Russ B Altman, and Possu Huang. Protein sequence design with a learned potential. *bioRxiv*, pp. 2020–01, 2021.

Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *arXiv preprint arXiv:1906.04015*, 2019.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom: Energy-annotated molecular conformations for property prediction and molecular generation. *arXiv preprint arXiv:2006.05531*, 2020.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021. ISSN 0036-8075. doi: 10.1126/science.abj8754.

Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25): 8732–8733, 2009.

Yuchao Chai, Xue Han, Weiyao Li, Shanshan Liu, Sikai Yao, Chong Wang, Wei Shi, Ivan da Silva, Pascal Manuel, Yongqiang Cheng, et al. Control of zeolite pore interior for chemoselective alkyne/olefin separations. *Science*, 368(6494):1002–1006, 2020.

Benson Chen, Regina Barzilay, and Tommi Jaakkola. Path-augmented graph transformer network. *arXiv preprint arXiv:1905.12712*, 2019a.

Chao Chen, Guanbin Li, Ruijia Xu, Tianshui Chen, Meng Wang, and Liang Lin. Clusternet: Deep hierarchical cluster network with rigorously rotation-invariant representation for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4994–5002, 2019b.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019c.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Hyeoncheol Cho and Insung S Choi. Three-dimensionally embedded graph convolutional network (3dgcn) for molecule interpretation. *arXiv preprint arXiv:1811.09794*, 2018.

Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, 2018.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview. *arXiv preprint arXiv:2102.11090*, 2021.

Stephan Eismann, Patricia Suriana, Bowen Jing, Raphael JL Townshend, and Ron O Dror. Protein model quality assessment using rotation-equivariant, hierarchical neural networks. *arXiv preprint arXiv:2011.13557*, 2020.

Stephan Eismann, Raphael JL Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, and Ron O Dror. Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 89(5):493–501, 2021.

Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020.

Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301, 2016.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *arXiv preprint arXiv:2012.09688*, 2020a.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. Multi-scale self-attention for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7847–7854, 2020b.

Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. Rolx: structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1231–1239, 2012.

Adrián Hernández and José M Amigó. Attention mechanisms and their applications to complex systems. *Entropy*, 23(3):283, 2021.

Shion Honda, Shoi Shi, and Hiroki R Ueda. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *arXiv preprint arXiv:2005.00687*, 2020.

Michael J Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. Lietransformer: Equivariant self-attention for lie groups. In *International Conference on Machine Learning*, pp. 4533–4543. PMLR, 2021.

John Ingraham, Vikas K Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. 2019.

Katsuhiko Ishiguro, Shin-ichi Maeda, and Masanori Koyama. Graph warp module: an auxiliary module for boosting the power of graph neural networks in molecular graph analysis. *arXiv preprint arXiv:1902.01020*, 2019.

José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, pp. 4839–4848. PMLR, 2020.

Chaitanya Joshi. Transformers are graph neural networks. *The Gradient*, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021. doi: 10.1038/s41586-021-03819-2. (Accelerated article preview).

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Johannes Klicpera, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

Greg Landrum. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling, 2013.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021.

Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1052–1060, 2019.

Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

Łukasz Maziarka, Tomasz Danel, Sławomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanisław Jastrzębski. Molecule attention transformer. *arXiv preprint arXiv:2002.08264*, 2020.

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. Relative molecule self-attention transformer. *arXiv preprint arXiv:2110.05841*, 2021.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

Paul Morris, Rachel St. Clair, William Edward Hahn, and Elan Barenholtz. Predicting binding from screening assays with transformer network embeddings. *Journal of Chemical Information and Modeling*, 60(9):4191–4199, 2020.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3d oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, 2019.

Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7463–7472, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Giorgio Pesciullesi, Philippe Schwaller, Teodoro Laino, and Jean-Louis Reymond. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature communications*, 11(1):1–8, 2020.

Niklas Pfriem, Peter H Hintermeier, Sebastian Eckstein, Sungmin Kim, Qiang Liu, Hui Shi, Lara Milakovic, Yuanshuai Liu, Gary L Haller, Eszter Baráth, et al. Role of the ionic environment in enhancing the activity of reacting molecules in zeolite pores. *Science*, 372(6545):952–957, 2021.

Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

Raghunathan Ramakrishnan, Mia Hartmann, Enrico Tapavicza, and O Anatole Von Lilienfeld. Electronic spectra from tddft and machine learning in chemical space. *The Journal of chemical physics*, 143(8):084111, 2015.

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32:9689, 2019.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. *bioRxiv*, 2021.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *arXiv preprint arXiv:2007.02835*, 2020.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021.

Michael B Smith. *March's advanced organic chemistry: reactions, mechanisms, and structure*. John Wiley & Sons, 2020.

Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949, 2016.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*, 2018.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32:15642–15651, 2019.

Raphael JL Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Brandon Anderson, Stephan Eismann, Risi Kondor, et al. Atom3d: Tasks on molecules in three dimensions. *arXiv preprint arXiv:2012.04035*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Yulun Wu, Nicholas Choma, Andrew Chen, Mikaela Cashman, Érica T Prates, Manesh Shah, Verónica G Melesse Vergara, Austin Clyde, Thomas S Brettin, Wibe A de Jong, et al. Spatial graph attention and curiosity-driven policy for antiviral drug discovery. *arXiv preprint arXiv:2106.02190*, 2021.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? *arXiv preprint arXiv:2106.05234*, 2021.

Haitao Zhang, Gye Won Han, Alexander Batyuk, Andrii Ishchenko, Kate L White, Nilkanth Patel, Anastasiia Sadybekov, Beata Zamlynny, Michael T Rudd, Kaspar Hollenstein, et al. Structural basis for selectivity and diversity in angiotensin ii receptors. *Nature*, 544(7650):327–332, 2017.

Shichang Zhang, Ziniu Hu, Arjun Subramonian, and Yizhou Sun. Motif-driven contrastive learning of graph representations. *arXiv preprint arXiv:2012.12533*, 2020a.

Shuo Zhang, Yang Liu, and Lei Xie. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. *arXiv preprint arXiv:2011.07457*, 2020b.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. Motif-based graph self-supervised learning for molecular property prediction. *arXiv preprint arXiv:2110.00987*, 2021.

Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

A EXPERIMENTAL SETUP

A.1 EXPERIMENTAL DETAILS

Molformer Architecture. A standard Molformer has 6 multi-scale self-attention layers, and each layer has 3 scales and 8 heads. Normally, scales are set by $\tau = [\frac{\rho}{2}, \rho, 2\rho]$, where ρ is the density of each corresponding dataset. The number of selected atoms K and the weight ratio ϵ in AFPS is set as 4 and 0.1, respectively. We use ReLU as the activation function and a dropout rate of 0.1 for all layers. The input embedding size is 512 and the hidden size for FFN is 2048.

For BBBP and ClinTox, we use Molformer with 2 multi-scale self-attention layers with 4 heads. The scales are 0.8, 1.6, and 3.0 Å. The dropout rate is 0.2 and 0.6 for BBBP and ClinTox, respectively. For BACE, we use a standard Molformer but with a dropout rate of 0.2.

Training Details. We use Pytorch (Paszke et al., 2019) to implement Molformer and data parallelism in two GeForce RTX 3090. An Adam (Kingma & Ba, 2014) optimizer is used and a lambda scheduler is enforced to adjust it. We apply no weight decay there. Each model is trained with 300 epochs, except for PDBbind where we solely train the model for 30 epochs. For QM7 and QM8, we use a batch size of 64 and a learning rate of 10^{-4}. For QM9, we use a batch size of 256 and a learning rate of 10^{-3}. All hyper-parameters are tuned based on validation sets. For all molecular datasets, we impose no limitation on the input length and normalise the values of each regression task by mean and the standard deviation of the training set. We used grid search to tune the hyper-parameters of our model and baselines based on the validation dataset.

Motif Generation. We adopt RDKit (Landrum, 2013) to search motifs. However, QM8 and QM9 do not provide SMILES representations but only 3D coordinates, thus we cannot pull out motifs from these datasets. As for PDBbind, we only extract motifs of small molecules and leave out motifs in proteins.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 CONFORMATION CLASSIFICATION

Task and Data. In order to explore the influence of multiple conformations, we introduce a new task, conformation classification, to evaluate model's capacity to differentiate molecules with various low-energy conformations. We use the recent GEOM-QM9 (Axelrod & Gomez-Bombarelli, 2020) experiments. More specifically, GEOM-QM9 is an extension to QM9 dataset. It contains multiple conformations for most molecules, while the original QM9 only contains one.

We randomly draw 1000 different molecules from GEOM-QM9, each with 20 different conformations. Models are required to distinguish the molecular type given different conformations. We take a half of each molecular conformations as the training set and another half as the test split. Since it is a multi-class classification problem with 1000 classes, we compute the micro-average and macro-average ROC-AUC as well as the accuracy for numerical evaluations.

Results. Molformer achieves a micro-average and macro-average ROC-AUC of 1.0 and 1.0, and an accuracy of 0.999. This indicates strong robustness of our model against different spatial conformations of molecules.

B.2 AFPS VS. FPS.

To have a vivid understanding of the atom sampling algorithm, we conducted a case study on a random crystal (see Figure 3). Points selected by FPS are randomized and exclude vital atoms like the heavy metal Nickel (Ni). With the adoption of AFPS, sampled points include Ni and Nitrogen (N) besides that they keep remote distances from each other. Moreover, FPS integrates too many features of trivial atoms like Hydrogen (H) while misses out key atoms, which will significantly smooth the molecular representations and lead to poor predictions. This illustrative example firmly shows the effectiveness of our AFPS to offset disadvantages of the conventional FPS in 3D molecular representation.

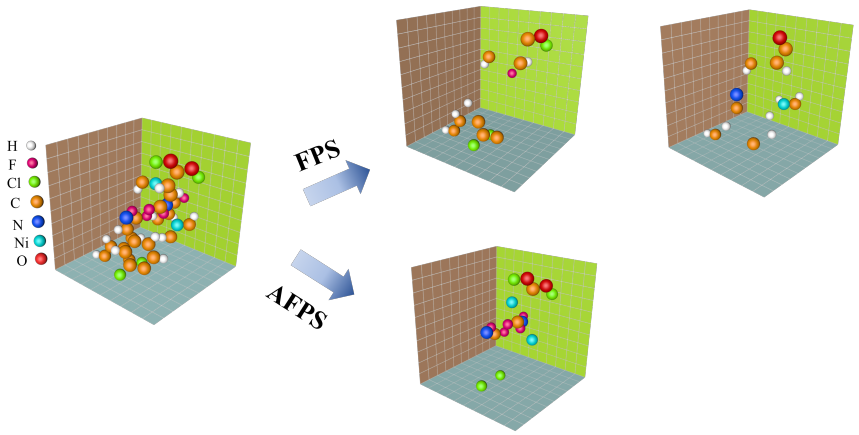

Figure 3: Sampled points using FPS and AFPS. We do not show dummy nodes there.