Efficient SE(3)-Transformer for Learning from 3D Structures with Limited Data

Shikai Qiu [*1,2] **Hunter Nisonoff** [*3] **Akosua Busia** [1] **Derek Jones** [4] **Hyojin Kim** [4] **Garrett Stevenson** [4]
Jonathan Allen [4] **Jennifer Listgarten** [1,3]

Abstract

Numerous important problems in physics, chemistry, biology, and vision deal with 3D data. SE(3)-equivariant models have recently been developed to properly account for the rotational and translational symmetries that are inherent to many problems involving 3D data. Among these models, the SE(3)-Transformer implements one of the most general forms of SE(3)-equivariant message passing on 3D graph data. However, its applicability has been limited due to requiring a large number of parameters and large compute resources. We propose a variant of SE(3)-Transformer that is much more compute and parameter efficient, while still achieving SE(3)-equivariance and capable of capturing the important 3D geometry. We demonstrate that our model leads to a significant improvement on an important but data-limited task, binding affinity prediction, and achieves similar performance on QM9 compared to the original SE(3)-Transformer while using only 3% of the parameters.

1. Introduction

There is growing interest in leveraging neural networks to solve problems involving 3D point cloud and graph data, with protein structure prediction (Jumper et al., 2021) and molecular quantum mechanical property prediction (Ramakrishnan et al., 2014) being two examples. Meanwhile, there has been a flurry of progress in the development of novel architectures that capture the inherent geometrical and permutation symmetries in these problems (Fuchs et al., 2020; Anderson et al., 2019; Kondor et al., 2018; Weiler et al., 2018; Thomas et al., 2018; Kondor, 2018). Among these models, the SE(3)-Transformer provides one of the the most general forms of SE(3)-equivariant message passing based on irreducible representations of the group SE(3), and has demonstrated impressive results on a diverse set of tasks (Baek et al., 2021; Fuchs et al., 2021).

Unfortunately, the SE(3)-Transformer requires a large number of parameters and large compute resources, which limits its applicability in general settings. For example, the RoseTTAFold (Baek et al., 2021) architecture achieving accuracies approaching those of AlphaFold2 (Jumper et al., 2021) on CASP14 targets was limited to using multiple discontinuous crops of the protein sequence in order to utilize the SE(3)-Transformer. Moreover, numerous important problems in structural biology and drug-discovery are extremely data-limited and therefore require models capable of achieving SE(3)-equivariance with a small number of parameters. As a canonical example, there is significant interest in using neural networks to replace physics-based scoring functions in virtual screening applications (Jones et al., 2021; Stevenson et al., 2021; Feinberg et al., 2018; Zhang et al., 2019; Ragoza et al., 2017). This task involves utilizing a neural network that takes as input the 3D structure of a small-molecule bound to a protein and outputs the binding affinity. This challenging problem relies on available experimental data, which is currently limited to $O(10^4)$ training examples.

In this paper we present the Efficient SE(3)-Transformer, a new SE(3)-equivariant Transformer architecture that achieves significantly higher compute and parameter efficiency than the original SE(3)-Transformer, by generalizing depthwise separable convolution (Sifre & Mallat, 2014; Abadi et al., 2015; Chollet, 2016) to SE(3)-equivariant message passing, and replacing the dot product attention weights with those generated directly from learned edge features. We then analyze its performance on a binding affinity prediction task (Jones et al., 2021) and the QM9 small molecule property prediction task (Ramakrishnan et al., 2014). Our Efficient SE(3)-Transformer leads to a significant improvement on the former task, and achieves similar performance on the latter compared to the original SE(3)-Transformer while using only 3% of the parameters.

[*]Equal contribution [1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA [2]Department of Physics, University of California, Berkeley, CA, USA [3]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA [4]Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, United States. Correspondence to: Jennifer Listgarten Ph.D. <jennl@berkeley.edu>.

Preliminary work.

Our contributions are the following:

- We quantitatively characterize how the equivariant message passing employed in the SE(3)-Transformer leads to a substantial increase in its number of parameters compared to standard graph neural networks.
- We demonstrate how to generalize depthwise separable convolutions to SE(3)-equivariant message passing by factorizing its attention kernel.
- We devise an alternative method to efficiently generate SE(3)-invariant attention coefficients from learned edge features.
- We demonstrate our modifications to the SE(3)-Transformer significantly reduce the required compute and number of parameters while providing better performance in the low-data regime and comparable performance in the high-data regime.

2. Background

2.1. CNN and GNN

Convolutional Neural Networks (CNNs) (Lecun & Bengio, 1995) and Graph Neural Networks (GNNs) (Scarselli et al., 2009; Gilmer et al., 2017; Battaglia et al., 2018) are the two most common approaches to learn from three-dimensional structures.

CNN In a CNN, convolution is used to transform an input feature map with c_{in} channels $f_{\text{in}} : \mathbb{Z}^3 \to \mathbb{R}^{c_{\text{in}}}$ to an output feature map with c_{out} channels $f_{\text{out}} : \mathbb{Z}^3 \to \mathbb{R}^{c_{\text{out}}}$ via

$$f_{\text{out}}(x) = \sum_{x' \in N(x)} K(x - x') f_{\text{in}}(x'), \tag{1}$$

with a learnable kernel $K : \mathbb{Z}^3 \to \mathbb{R}^{c_{\text{in}} \times c_{\text{out}}}$ whose argument is the relative position of the output voxel with respect to to the input voxel. Here $N(x)$ denotes the finite set of voxels neighbouring x for which $K(x - x') \neq 0$. Afterwards, a pointwise non-linearity is applied to the output feature map. CNNs of the general form in Equation 1 are guaranteed to be equivariant under translation but are not necessarily equivariant under rotation. A major disadvantage of applying CNNs to spatially sparse 3D structures such as molecules is the enormous computational overhead of convolving over mostly unoccupied voxels.

GNN GNNs provide a much more efficient alternative to process sparse 3D structures consisting of a discrete set of points $\{x_i \in \mathbb{R}^3\}$ and their associated feature vectors $\{f_{\text{in},i} \in \mathbb{R}^{c_{\text{in}}}\}$. As inputs to a GNN, such structures are commonly represented as graphs in which points are connected via edges based on a distance cutoff (Feinberg et al., 2018; Jones et al., 2021). In a GNN, convolution in the form of Equation 1 is replaced by message passing (Battaglia et al., 2018), one popular instantiation of which is through the attention mechanism (Vaswani et al., 2017), where

$$f_{\text{out},i} = \sum_{j \in N(i)} \alpha_{ij} v_i, \tag{2}$$

$$\alpha_{ij} = \frac{\exp\left(q_i^\top k_i\right)}{\sum_{j \in N(i)} \exp\left(q_i^\top k_j\right)}, \tag{3}$$

$$q_i = Q f_{\text{in},i}, \quad k_j = K f_{\text{in},j} \tag{4}$$

$$v_j = V f_{\text{in},j}. \tag{5}$$

Here $Q, K, V \in \mathbb{R}^{c_{\text{in}} \times c_{\text{out}}}$ are learned matrices, or attention kernels, and $N(i)$ denotes the set of points connected to i through an edge. Message passing of the form in Equation 2 resembles convolution with a kernel V without spatial dependence. Most GNNs are therefore less capable of utilizing structural information than CNNs. To incorporate additional structural information without breaking SE(3) symmetry, earlier works have attempted to indirectly specify the 3D structure of the graph by featurizing its edges with pairwise distances (Qasim et al., 2019; Schütt et al., 2017). Unlike CNNs, the feature vectors and outputs learned by these methods are invariant under global SE(3) transformations. However, pairwise distances are not sufficient to specify the full 3D structure unless the graph is a complete graph, which is computationally infeasible to train GNNs on for macromolecular structures consisting of $O(10^3)$ atoms.

2.2. SE(3)-Transformer

We briefly describe the architecture of the SE(3)-Transformer (Fuchs et al., 2020), and illustrates how the equivariant message passing employed in the SE(3)-Transformer results in a substantial increase in its required compute and number of parameters compared to standard graph neural networks.

2.2.1. Architecture

The SE(3)-Transformer extends attention-based GNNs to explicitly make use of relative positions between points, similar to what's done in CNNs, while achieving SE(3)-equivariance. In the SE(3)-Transformer, each feature vector f_i is a direct sum over subvectors of different degrees $\ell \in \{0, 1, 2, ...\}$ and channels $c \in \{0, 1, 2, ...\}$,

$$f_i = \bigoplus_{\ell} \bigoplus_{c} f_i^{\ell c}, \tag{6}$$

where $\bigoplus$ denotes the direct sum, or vector concatenation. A subvector of degree ℓ has length $2\ell + 1$, and transforms under $g \in \text{SE}(3)$ via left multiplying by a square matrix $D_\ell(g)$, the ℓ-th Wigner D-matrix. Using multiple channels per degree allows more information to be captured in the

hidden layers, similar to the use of channels in a CNN. The SE(3)-equivariant attention is identical to the original attention except with the following definitions of the key, query, and value vectors:

$$v_{ij} = \bigoplus_{\ell,c} \sum_{\ell'} \sum_{c'} V^{\ell c}_{\ell' c'}(x_i - x_j) f^{\ell' c'}_{\text{in},j}, \tag{7}$$

$$k_{ij} = \bigoplus_{\ell,c} \sum_{\ell'} \sum_{c'} K^{\ell c}_{\ell' c'}(x_i - x_j) f^{\ell' c'}_{\text{in},j}, \tag{8}$$

$$q_i = \bigoplus_{\ell,c} \sum_{c'} Q^{\ell c}_{c'} f^{\ell c'}_{\text{in},i}, \tag{9}$$

where $V^{\ell c}_{\ell' c'}, K^{\ell c}_{\ell' c'} : \mathbb{R}^3 \to \mathbb{R}^{(2\ell+1)\times(2\ell'+1)}$ maps a relative position vector to a $(2\ell+1)\times(2\ell'+1)$ matrix, and $Q^{\ell c}_{c'} \in \mathbb{R}$. For comparison, we can rewrite the original attention by expanding the matrix multiplications as

$$v_j = \bigoplus_{c} \sum_{c'} V^{c}_{c'} f^{c'}_{\text{in},j}, \tag{10}$$

$$k_j = \bigoplus_{c} \sum_{c'} K^{c}_{c'} f^{c'}_{\text{in},j}, \tag{11}$$

$$q_i = \bigoplus_{c} \sum_{c'} Q^{c}_{c'} f^{c'}_{\text{in},i}, \tag{12}$$

where $V^{c}_{c'}, K^{c}_{c'}, Q^{c}_{c'} \in \mathbb{R}$. This shows that the original attention mechanism is a special case of the SE(3)-equivariant attention with $\ell = 0$ and attention kernels V, K which has no spatial dependence. In order for the kernels V, K to have spatial dependence while guaranteeing equivariance, they must be a linear combination of the allowed basis kernels,

$$K^{\ell c}_{\ell' c'}(x) = \sum_{J=|\ell-\ell'|}^{\ell+\ell'} \varphi_J^{\ell c \ell' c'}(\|x\|) W_J^{\ell\ell'}(x), \tag{13}$$

$$W_J^{\ell\ell'}(x) = \sum_{m=-J}^{J} Y_{Jm}\left(\frac{x}{\|x\|}\right)(Q_{\text{CG}})^{\ell\ell'}_{Jm} \tag{14}$$

where Y_{Jm} are spherical harmonics, Q_{CG} are Clebsch-Gordan matrices, and $\varphi_J^{\ell c \ell' c'}(\|x\|)$ is a function mapping radial separation to coefficients. In general, $\varphi_J^{\ell c \ell' c'}(\cdot)$ could take in additional quantities, such as edge features, as arguments so long as they are SE(3) invariant. It is implemented as a neural network $\phi^{\ell}_{\ell'}$ for each pairs of degrees (ℓ, ℓ'), which we refer to as the coefficient network, whose output is a $2 \times \min(\ell, \ell') \times c_{\text{in}} \times c_{\text{out}}$ dimensional vector of coefficients, i.e. $\phi^{\ell}_{\ell'}(\|x\|) = \bigoplus_{c,c',J} \varphi_J^{\ell c \ell' c'}(\|x\|)$. For a proof of equivariance and more details on the architecture, we refer the reader to the original SE(3)-Transformer paper (Fuchs et al., 2020).

2.2.2. Model complexity

The main advantage of the SE(3)-Transformer is its ability to learn spatially dependent attention kernels to perform message passing, allowing the model to capture non-trivial 3D geometric patterns in the data. In contrast, ordinary Transformers and other GNNs do not have kernels with non-trivial spatial dependence. However, one problem with using spatially dependent kernels is that it leads to a significant increase in the number of parameters. For a fixed pair of degrees (ℓ, ℓ'), the number of parameters in a single attention kernel for the SE(3)-Transformer is equal to that in the corresponding coefficient network $\phi^{\ell}_{\ell'}$. Assume a minimal setting where $\phi^{\ell}_{\ell'}$ is a single player MLP. Since $\phi^{\ell}_{\ell'}$ outputs $2 \times \min(\ell, \ell') \times c_{\text{in}} \times c_{\text{out}}$ coefficients, if its hidden layer size is H, it will have at least $2 \times \min(\ell, \ell') \times c_{\text{in}} \times c_{\text{out}} \times H$ parameters. The spatial dependence of the kernel therefore multiplies its parameter count by at least factor of H (usually ≥ 10).

2.2.3. Computational challenge

As noted in the original paper (Fuchs et al., 2020), the SE(3)-Transformer brings about computational challenge in terms of both computation time and memory usage, a major cause of which is the need to evaluate the attention kernels during the forward pass, which entails forward passes in the coefficient networks $\{\phi^{\ell}_{\ell'}\}_{\ell,\ell'}$ as well as expensive evaluations of spherical harmonics. For small molecule tasks such as QM9, these challenges do not prevent the use of the SE(3)-Transformer because the number of nodes in the graph is sufficiently small($O(10)$). But in tasks involving macro molecules, such as proteins, the cost of using the SE(3)-Transformer can become prohibitive for researchers without access to abundant compute resources (Baek et al., 2021).

3. Method

We propose two following modifications to the original SE(3)-Transformer: 1) Depthwise separable SE(3)-equivariant message passing, which reduces the complexity in computing equivariant messages between points; 2) Edge feature generated attention weights, which reduces the complexity in computing the pairwise attention weights.

3.1. Depth-wise separable SE(3)-equivariant message passing

Recall the SE(3)-Transformer computes pairwise value vectors through

$$v_{ij} = \bigoplus_{\ell,c} \sum_{\ell'} \sum_{c'} V^{\ell c}_{\ell' c'}(x_i - x_j) f^{\ell' c'}_{\text{in},j}. \tag{15}$$

The complexity of this operation is manifested in the two sums over ℓ' and c', which play different roles. The sum over ℓ' is essential because it enables correlations among feature vectors of different degrees. Without it, scalar features ($\ell = 0$) including the scalar outputs are learnt independently from

features of higher degrees ($\ell > 0$), which carry geometric information. On the other hand, the sum over channels c' is purely to increase the model capacity by allowing each output channel to receive information from each input channel. To reduce the model's complexity while retaining correlations among feature vectors of different degrees, we factorize the attention kernel V to be the product of a degree-wise mixing kernel $\tilde{V}$ and a channel-wise mixing matrix U,

$$V^{\ell c}_{\ell' c'}(x) = U^c_{c'} \tilde{V}^{\ell}_{\ell'}(x). \tag{16}$$

This simplifies the message passing to be

$$v_{ij} = \bigoplus_{\ell,c} \sum_{\ell'} \sum_{c'} U^c_{c'} \tilde{V}^{\ell}_{\ell'}(x_i - x_j) f^{\ell' c'}_{\text{in},j}, \tag{17}$$

$$= \bigoplus_{\ell,c} \sum_{c'} U^c_{c'} \underbrace{\sum_{\ell'} \tilde{V}^{\ell}_{\ell'}(x_i - x_j) f^{\ell' c'}_{\text{in},j}}_{\text{mixing among degrees}}, \tag{18}$$

$$= \bigoplus_{\ell,c} \underbrace{\sum_{c'} U^c_{c'} \tilde{v}^{\ell c'}_{ij}}_{\text{mixing among channels}} . \tag{19}$$

$$\tag{20}$$

The channel-wise mixing matrix U can additionally depend on the output degree, which is necessary if the number of channels differ for different input or output degrees as U would need to take different shapes. One can interpret this decomposition as the generalization of the depthwise separable convolution (Sifre & Mallat, 2014; Abadi et al., 2015; Chollet, 2016) to the SE(3)-equivariant message passing, by rewriting the message passing in convolution form

$$f_{\text{out},i} = \sum_{j \in N(i)} \alpha_{ij} v_{ij}, \tag{21}$$

$$= \bigoplus_{\ell,c} \sum_{j \in N(i)} \alpha_{ij} \left(\sum_{\ell',c'} V^{\ell c}_{\ell' c'}(x_i - x_j) f^{\ell' c'}_{\text{in},j} \right), \tag{22}$$

where the convolution over neighbouring voxels is replaced by the sum over neighbours. The decomposition then allows one to write

$$f_{\text{out},i} = \bigoplus_{\ell,c} \sum_{j \in N(i)} \alpha_{ij} \left(\sum_{c'} U^c_{c'} \sum_{\ell'} \tilde{V}^{\ell}_{\ell'}(x_i - x_j) f^{\ell' c'}_{\text{in},j} \right), \tag{23}$$

$$= \bigoplus_{\ell,c} \sum_{c'} U^c_{c'} \underbrace{\sum_{j \in N(i)} \alpha_{ij} \left(\sum_{\ell'} \tilde{V}^{\ell}_{\ell'}(x_i - x_j) f^{\ell' c'}_{\text{in},j} \right)}_{\text{depthwise convolution}}, \tag{24}$$

$$= \bigoplus_{\ell,c} \underbrace{\sum_{c'} U^c_{c'} \tilde{f}^{\ell c'}_{\text{in},i}}_{1 \times 1 \text{ convolution}} . \tag{25}$$

The SE(3)-equivariant attention is therefore factorized into a depth-wise convolution performed independently for each channel, mixing only features from different points and different degrees, followed by a 1×1 convolution performed independently for each point, mixing features from different channels but not across degrees. This reduces the parameter count in any attention layer by a factor of $c_{\text{in}} \times c_{\text{out}}$, which is usually $O(10^2)$.

3.2. Edge feature generated attention weights

We further reduce the model complexity by simplifying the calculation of the attention coefficients α_{ij}. In the SE(3)-Transformer, the keys k_{ij} required to evaluate the attention coefficients are generated using equivariant kernels $\{K^{\ell c}_{\ell' c'}\}_{\ell,\ell',c,c'}$. Since the resulting attention coefficients are only used as weights for the value vectors, we hypothesis we can drastically simplify their calculation without significantly limiting the expressivity of the model. Instead of using the traditional dot product attention, we associate every edge $(i,j), j \in N(i)$ with a feature vector $e_{ij} \in \mathbb{R}^D$ constructed to be SE(3)-invariant. In each attention layer, a learned linear transformation $A \in \mathbb{R}^{H \times D}$ maps each edge feature vector to a vector containing the unnormalized attention weights for each attention head, replacing the dot products in the original attention mechanism,

$$\alpha_{ij} = \frac{\exp(A e_{ij})}{\sum_{j \in N(i)} \exp(A e_{ij})} \in \mathbb{R}^H. \tag{26}$$

After each attention layer, an MLP updates the edge feature vectors based on the degree-0 features of the two points connected by the edge,

$$e_{\text{out},ij} = e_{\text{in},ij} + \text{MLP}(e_{\text{in},ij}, \bigoplus_c f^{0c}_{\text{in},i}, \bigoplus_c f^{0c}_{\text{in},j}). \tag{27}$$

As the complexity of an MLP is negligible compared to an equivariant kernel, adopting the edge feature generated

attention weights enables us to halve the number parameters in the attention layers and significantly speed up the computation. The updates to the edge features allow the attention weights to non-trivially evolve over different layers in the network.

4. Results

4.1. Binding affinity prediction

We evaluate the model on the task of structure-based binding affinity prediction. We use the PDBBind dataset (Cheng et al., 2009), a comprehensive collection of experimentally measured binding affinity data for the protein-ligand complexes deposited in the Protein Data Bank. It is commonly used to evaluate the performance of computational scoring functions. To compare with previously published work, we use PDBBind2016 general set for training and validation. For testing, we use the structure-based holdout set proposed by (Jones et al., 2021), which consists of 222 complexes that are structurally novel with respect to the training set. We use the same input featurization and training-validation split as (Jones et al., 2021), with 11,693 complexes for training and 377 complexes for validation. We use degrees $\ell \in \{0, 1, 2\}$, 16 channels, and 3 message passing layers. We show the Pearson correlation on the test set achieved by our model and previous methods in Table 1. Our model shows significant improvement over previous methods on this task.

Table 1. Pearson correlation between predicted binding affinity and ground truth on the structure-based holdout set. Mean and standard deviation of our method are computed over 3 runs. SG-CNN (Jones et al., 2021) is a GNN model with pair-wise distance information, (SG-CNN + 3D-CNN) Fusion (Jones et al., 2021) is a fusion model between a CNN and a GNN, K_{DEEP} (Jiménez et al., 2018) and Pafnucy (Stepniewska-Dziubinska et al., 2018) are both CNN models.

Model	Correlation
SG-CNN	0.515
(SG-CNN + 3D-CNN) Fusion	0.545
K_{DEEP}	0.487
Pafnucy	0.528
Efficient SE(3)-Transformer	**0.608±0.02**

We compare the proposed Efficient SE(3)-Transformer with several of its variants. The scaler only variant only uses features of degree $\ell = 0$ and learns SE(3)-invariant features only. The dot product attention variant does not use the more efficient edge feature generated attention weights but keeps their original definition via dot products between keys and queries defined in Equation 8 and 9. Finally the original SE(3)-Transformer does not use either depthwise separable SE(3)-equivariant message passing or edge feature generated attention weights. All models use 16 channels and degrees $\ell \in \{0, 1, 2\}$, except the scalar only variant which only use $\ell = 0$. Table 2 shows the performance and resource usage of these models. Compared to the original SE(3)-Transformer, our model reduces the total parameter count and peak memory usage by over a factor of 10, and reduces the forward time by a factor of 2.7. Our model outperforms the scalar only variant, showing the advantage of using equivariant feature vectors to capture geometric information in this task. It also outperforms the dot product attention variant, which demonstrates the benefit of using edge feature generated attention weights to reduce model complexity.

Table 2. Comparing the correlation, parameter count, forward time, and peak memory usage among variants of the Efficient SE(3)-Transformer. The forward time and peak memory usage is measured by running each model on a single fixed complex. We did not produced correlation for the original SE(3)-Transformer on this task due to insufficient GPU memory.

Model	Corr.	$\frac{\#\text{param}}{10^3}$	$\frac{\text{Time}}{\text{ms}}$	$\frac{\text{Memory}}{\text{MB}}$
Eff. SE(3)-Tr. scalar only	0.56	5.9	12	4.4
Eff. SE(3)-Tr.	0.61	9.4	25	18.6
Eff. SE(3)-Tr. dot product attn.	0.53	13	35	20.5
Original SE(3)-Tr.	-	120	67	689.6

4.2. Small molecule property prediction

The QM9 dataset (Ramakrishnan et al., 2014) contains the results of quantum-chemical calculations for 134,000 small molecules consisting of at most nine non-hydrogen atoms. Each atom is represented by its spatial coordinates and a one-hot encoding of C, O, N, F, and H, while bond types are provided as edge features. We compare our model with the original SE(3)-Transformer several tasks from this benchmark in Table 3. Our model achieves comparable performance with only 3% of the parameters.

Table 3. The mean absolute error achieved by our model and the original SE(3)-Transformer, along with their parameter count in the last column. Our model achieves comparable performance with only 3% of the parameters.

	$\frac{\varepsilon_{\text{HOMO}}}{\text{meV}}$	$\frac{\varepsilon_{\text{LUMO}}}{\text{meV}}$	$\frac{\alpha}{\text{bohr}^3}$	$\frac{\Delta\varepsilon}{\text{meV}}$	$\frac{\#\text{param}}{10^6}$
SE(3)-Tr	35	33	.142	53	9.8
Eff. SE(3)-Tr.	38	35	.140	54	0.29

5. Conclusion

We have presented a new SE(3)-equivariant Transformer architecture that achieves significantly higher compute and parameter efficiency than the SE(3)-Transformer, by

generalizing depthwise separable convolution to SE(3)-equivariant message passing, and efficiently generating attention weights from learned edge features. Our Efficient SE(3)-Transformer leads to a significant improvement an important but data-limited task, binding affinity prediction, and achieves similar performance on QM9 compared to the original SE(3)-Transformer while using only 3% of the parameters.

We hope that the developments made in this paper allow researchers working with smaller datasets and/or smaller compute resources to better leverage SE(3)-equivariant architectures.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Anderson, B., Hy, T.-S., and Kondor, R. Cormorant: Covariant molecular neural networks. *arXiv preprint arXiv:1906.04015*, 2019.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. doi: 10.1126/science.abj8754. URL https://doi.org/10.1126/science.abj8754.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks, 2018. URL http://arxiv.org/abs/1806.01261. cite arxiv:1806.01261.

Cheng, T., Li, X., Li, Y., Liu, Z., and Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49(4):1079–1093, 2009. URL http://dblp.uni-trier.de/db/journals/jcisd/jcisd49.html#ChengLLLW09.

Chollet, F. Xception: Deep learning with depthwise separable convolutions, 2016. URL http://arxiv.org/abs/1610.02357. cite arxiv:1610.02357.

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B., and Pande, V. S. Potentialnet for molecular property prediction. *ACS central science*, 4(11):1520–1530, 2018.

Fuchs, F., Worrall, D., Fischer, V., and Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Fuchs, F. B., Wagstaff, E., Dauparas, J., and Posner, I. Iterative SE(3)-transformers. In *Lecture Notes in Computer Science*, pp. 585–595. Springer International Publishing, 2021. doi: 10.1007/978-3-030-80209-7_63. URL https://doi.org/10.1007/978-3-030-80209-7_63.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W. (eds.), *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL http://dblp.uni-trier.de/db/conf/icml/icml2017.html#GilmerSRVD17.

Jiménez, J., Škalič, M., Martínez-Rosell, G., and Fabritiis, G. D. KDEEP: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, January 2018. doi: 10.1021/acs.jcim.7b00650. URL https://doi.org/10.1021/acs.jcim.7b00650.

Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W. D., Kirshner, D., Wong, S. E., Lightstone, F. C., and Allen, J. E. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of Chemical Information and Modeling*, 61 (4):1583–1592, 2021.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kondor, R. N-body networks: a covariant hierarchical neural network architecture for learning atomic potentials. *arXiv preprint arXiv:1803.01588*, 2018.

Kondor, R., Lin, Z., and Trivedi, S. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31: 10117–10126, 2018.

Lecun, Y. and Bengio, Y. *Convolutional Networks for Images, Speech and Time Series*, pp. 255–258. The MIT Press, 1995.

Qasim, S. R., Kieseler, J., Iiyama, Y., and Pierini, M. Learning representations of irregular particle-detector geometry with distance-weighted graph networks. *The European Physical Journal C*, 79(7), July 2019. doi: 10.1140/epjc/s10052-019-7113-9. URL https://doi.org/10.1140/epjc/s10052-019-7113-9.

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

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

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks (TNN)*, 20(1): 61–80, 2009. ISSN 1045-9227. doi: 10.1109/TNN.2008.2005605. URL http://ieeexplore.ieee.org/document/4700287/.

Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *NIPS*, pp. 991–1001, 2017. URL http://dblp.uni-trier.de/db/conf/nips/nips2017.html#SchuttKFCTM17.

Sifre, L. and Mallat, P. S. Ecole polytechnique, cmap phd thesis rigid-motion scattering for image classification author, 2014.

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, May 2018. doi: 10.1093/bioinformatics/bty374. URL https://doi.org/10.1093/bioinformatics/bty374.

Stevenson, G. A., Jones, D., Kim, H., Bennett, W., Bennion, B. J., Borucki, M., Bourguet, F., Epstein, A., Franco, M., Harmon, B., et al. High-throughput virtual screening of small molecule inhibitors for sars-cov-2 protein targets with deep fusion models. *arXiv preprint arXiv:2104.04547*, 2021.

Thomas, N., Smidt, T., Kearnes, S. M., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *CoRR*, abs/1802.08219, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1802.html#abs-1802-08219.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. S. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/488e4104520c6aab692863cc1dba45af-Paper.pdf.

Zhang, H., Liao, L., Saravanan, K. M., Yin, P., and Wei, Y. Deepbindrg: a deep learning based method for estimating effective protein–ligand affinity. *PeerJ*, 7:e7362, 2019.