Geometric Transformer for End-to-End Molecule Properties Prediction

Yoni Choukroun and Lior Wolf
School of Computer Science, Tel Aviv University
choukroun.yoni@gmail.com, wolf@cs.tau.ac.il

Abstract
Transformers have become methods of choice in many applications thanks to their ability to represent complex interactions between elements. However, extending the Transformer architecture to non-sequential data such as molecules and enabling its training on small datasets remains a challenge. In this work, we introduce a Transformer-based architecture for molecule property prediction, which is able to capture the geometry of the molecule. We modify the classical positional encoder by an initial encoding of the molecule geometry, as well as a learned gated self-attention mechanism. We further suggest an augmentation scheme for molecular data capable of avoiding the overfitting induced by the overparameterized architecture. The proposed framework outperforms the state-of-the-art methods while being based on pure machine learning solely, i.e. the method does not incorporate domain knowledge from quantum chemistry and does not use extended geometric inputs besides the pairwise atomic distances.

1 Introduction
Properties of chemical compounds can generally be estimated using methods such as density functional theory (DFT) or ab initio quantum chemistry [Jensen, 2017]. However, these can be computationally expensive and therefore have a limited applicability, especially for larger systems. In recent years, many approaches have started leveraging machine learning to reduce the computational complexity required for efficiently predicting molecular properties.

In this vein, many contributions have focused on the creation of handcrafted representations at the atomic or molecular level [Christensen and others, 2020; Huang and others, 2016] as input for various machine learning methods. Schrödinger’s equation indicates that the system variables that define the ground-state properties of a given molecule are a function of the inter-atomic distances and the nuclear charges solely. [Jensen, 2017]. Based on this observation, several recent methods predict molecular properties in an end-to-end fashion where the input is defined by the atoms’ type and spatial position. Such methods often incorporate quantum chemistry knowledge and rely on extensive hyper-parameter tuning.

Since atomic interactions are challenging to simulate, many recent works make use of graph neural networks as a natural tool to model molecules [Gilmer et al., 2017]. Related to graph neural networks, Transformers [Vaswani et al., 2017] have recently become extremely popular in numerous application domains.

In this paper we extend the ubiquitous Transformer to chemical compounds data in order to predict their ground-state properties. Our work does not employ extended domain knowledge, and is based solely on the simple distance relevance assumption, namely, that the bigger the distance between atomic elements, the lower the interaction.

Contrary to other works, the framework does not assume any extended input, e.g. quantum mechanical properties [Qiao and others, 2020], complex geometric constants such as bending or torsion angles [Klicpera et al., 2020b], additional solvers, e.g. fast DFT as residual solvers (i.e. delta learning) [Unke and others, 2019; Qiao and others, 2020], or even knowledge adaptation from quantum chemistry into the machine learning model design [Schütt et al., 2018a; Klicpera et al., 2020b].

The Transformer we design is endowed with an adapted positional encoder, and with learned inter-atomic geometric embedding at the different levels of the model, allowing increased representational power, while maintaining the molecule invariance to rigid transformations and permutation. For better regularization, we suggest to augment the training set by merging pairs of molecules positioned far apart.

The experimental results demonstrate the representational power of the end to end model, potentially allowing removal of undesirable inductive bias [Goyal and Bengio, 2020] such as handcrafted envelope functions. Also, by allowing the model to fit the data while minimizing the domain knowledge, scientific insights can emerge, such as effective atomic cut-off radius, or most relevant molecular interactions for a given property.

Related Work Classical molecule property prediction methods combined handcrafted features generally based on force field methods at the atom level [Christensen and others, 2020] or molecular level [Huang and others, 2016], integrated into various machine learning models such as Kernel methods [Christensen and others, 2020], Gaussian processes [Bartok

Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
2895
composed of coupled with the augmented self-attention module (b). The model is on the pairwise distance matrix (c), and the metric learning module. Figure 1: The proposed Transformer architecture (a). The main architecture modifications are the initial positional encoding based on the pairwise distance matrix (c), and the metric learning module coupled with the augmented self-attention module (b). The model is composed of $M$ encoding blocks and the output block is composed of a normalized feed-forward neural network followed by a per atom summation to accumulate the contribution of each atom.

et al., 2010) or Neural networks [Schütt et al., 2018a].

These methods have recently been superseded by end-to-end neural networks, alleviating the need for handcrafted signatures. The most popular and powerful models are based on graph neural networks, which allow a natural representation of the molecular graph [Schütt et al., 2018a; Unke and others, 2019; Anderson et al., 2019; Klicpera et al., 2020b; Qiao and others, 2020]. These architectures are generally designed based on the message-passing mechanism, by aggregating features obtained from atom types, geometric invariants such as pairwise inter-atomic distances [Schütt et al., 2018a; Unke and others, 2019], bending or torsion angles [Klicpera et al., 2020b], or handcrafted atomic features derived from quantum mechanics [Qiao and others, 2020]. Existing methods generally involve a deep understanding and a cautious adaptation of the underlying physics in order to provide better pre-conditioning [Klicpera et al., 2020b; Qiao and others, 2020].

Transformer neural networks were originally introduced for machine translation [Vaswani et al., 2017] and they now dominate most applications in the field of Natural Language Processing. Transformer encoders primarily rely on the self-attention operation in conjunction with feed-forward layers, allowing manipulation of variable-size sequences and learning of long-range dependencies. Many works have augmented the self-attention mechanism using domain-specific knowledge [Chen et al., 2017; Bello et al., 2019].

Recently, a transformer architecture called MAT has been proposed for chemical molecules [Maziarka et al., 2020; Maziarka et al., 2021]. MAT modifies the self-attention module by summing the inverse exponent of the pairwise distance matrix to the self-attention tensor. As one contribution of our work, we show that the MAT self-attention architecture is sub-optimal in modelling interactions, and we propose a better self-attention module capable of capturing the connectivity of the graph. Our method also outperforms concurrent efforts [Wu et al., 2021; Kwak et al., 2021] which integrate mollifiers from the literature [Schütt et al., 2018a; Klicpera et al., 2020b] to the key element of the self-attention.

2 Method

A molecule is defined by the atomic numbers $z = \{z_1, \ldots, z_N\} \in \mathbb{Z}^+$, which serve to identify each type of atom, and the three-dimensional positions $X = \{x_1, \ldots, x_N\} \in \mathbb{R}^3$ of the $N$ atoms composing it. Molecular predictions must satisfy fundamental symmetries and invariance of physical laws such as invariance to rigid spatial transformation (rotation and translation) and permutation (atoms of the same type are indistinguishable). Therefore, the positional input is transformed to interatomic Euclidean distances $D = \{d_{ij}\}_{i,j=1}^N$ where $d_{ij} = \|x_i - x_j\|_2$ for rigid transform invariance, while permutation invariance is obtained via equal initial atomic representations of identical particles.

In this work, we design a parameterized deep neural network $f_\theta$ for scalar regression of properties $p \in \mathbb{R}$ such that $f_\theta : \{z, D\} \rightarrow \mathbb{R}$. We do not include any further auxiliary features in the input (e.g., adjacency, bond type, hybridization [Gilmer et al., 2017; Maziarka et al., 2020], DFT solver [Unke and others, 2019; Qiao and others, 2020]), neither any kind of knowledge adaptation from quantum physics/chemistry [Unke and others, 2019; Klicpera et al., 2020b; Klicpera et al., 2020a; Qiao and others, 2020]. Figure 1 depicts the proposed architecture. The positional encoding provides an initial geometry aware embedding of the atoms while the self-attention mechanism enables the accurate learning of the molecule geometry as well as the determination of the complex geometric interactions that are modeled in order to perform the regression task.

Transformer Transformer was introduced by [Vaswani et al., 2017] as a novel, attention-based building block for machine translation. The input sequence is first embedded into a high-dimensional space, coupled with positional embedding for each element. The embeddings are then propagated through multiple normalized self-attention and feed-forward blocks.

The self-attention mechanism introduced by Transformers is based on a trainable associative memory with (key, value) vector pairs where a query vector $q \in \mathbb{R}^d$ is matched against a set of $k$ key vectors using scaled inner products as follows

$$A(Q, K, V) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V, \quad (1)$$

where $Q \in \mathbb{R}^{N \times d}$, $K \in \mathbb{R}^{k \times d}$ and $V \in \mathbb{R}^{k \times d}$ represent the packed $N$ queries, $k$ keys and values tensors respectively. Keys, queries and values are obtained using linear transformations of the sequence’ elements. A multi-head self-attention layer is defined by extending the self-attention using $h$ attention heads, i.e., $h$ self-attention functions applied to the input, reprojected to values via a $dh \times D$ linear layer.

Geometric Positional Encoding The initial embedding of the molecule is based solely on the atoms’ type and thus is unable to differentiate similar atoms since the molecule geometry is omitted. The original Transformer’s positional encoding module aims to transfer a measure of proximity of
Figure 2: The impact of the initial positional encoder on the embedding for two different molecules (left and right). We show the pairwise distance matrix \( D \) (left), pairwise distance of the initial embedding \( \text{Emb}(z) \) (middle), and the pairwise distance of the final geometric embedding (right). The marks on the map represent the type of the atoms \( z_i \). Cold and warm colors represent low and high values respectively.

Figure 3: The validation \( L_1 \) loss (U0 property) of a regular transformer without any geometric input (red), the same transformer with our Laplacian extension (green), and with the proposed initial geometric positional encoding (GPE).


de notsional information to each atom. For each atom we first em-

g the input is defined as a set rather than a sequence, the
metric positional encoding (GPE).

Figure 3: The validation \( L_1 \) loss (U0 property) of a regular trans-
former without any geometric input (red), the same transformer with
our Laplacian extension (green), and with the proposed initial geo-
metric positional encoding (GPE).

the sequence elements to the initial embedding. In our case,
since the input is defined as a set rather than a sequence, the
positional encoder needs to be adapted in order to provide a
geometry-aware initial embedding. Here we propose to use
the pairwise inter-atomic distance matrix in order to bring po-
sitional information to each atom. For each atom we first em-
bed its pairwise distance vector to a single scalar in order to
keep the module invariant to the size of the molecule, and then
project it to the initial embedding space. Formally, denoting
the atom embedding \( \text{Emb}(z_i) : \mathbb{Z} \to \mathbb{R}^d \) and the pairwise
distance matrix \( D \in \mathbb{R}^{n \times n} \) such that \( (D)_{ij} := (D_i)_{j} = d_{ij} \),
we have

\[
y_i = \text{Emb}(z_i) + W \sum_j f_{pos}(d_{ij}).
\]

Here, \( y_i \) denotes the obtained initial positioned embedding of
atom \( z_i \), \( f_{pos} : \mathbb{R}^+ \to \mathbb{R} \) denotes the mapping of the pairwise
distance parameterized as a shallow neural network, and \( W \in \mathbb{R}^d \) denotes the projection matrix of the atomic one-
dimensional embedding onto the embedding space. Figure 2
depicts the impact of the positional encoder on the initial
embedding. As can be observed, the initial embedding does
not differentiate between atoms of the same type, while posi-
tional encoding brings information about the geometry. The
convergence plot in Figure 3 demonstrates that the proposed
positional encoding allows the transformer to learn the mol-
ecular geometry and thus be able to predict properties of
the molecule. We also compare our method with the positional
encoder of [Dwivedi and Bresson, 2020] where, since in our
setting no graph is given, we propose to compute the 15 (half
the biggest molecule size) first Laplacian eigenmaps [Belkin
and Niyogi, 2003] instead of the combinatorial Laplacian.

\[
\text{Emb}(z) = \text{L}(z) + \text{GPE} + \text{GPE}\text{Emb}(z)
\]

\[
\text{Emb}(z) = \text{L}(z) + \text{GPE}(z) + \text{GPE}\text{Emb}(z)
\]

Geometric Self-Attention In order to maintain the invariant
properties of molecules, we propose to augment the initial posi-
tional encoding layer and import the pairwise information
directly into the self-attention layer. This is a natural choice
since the self-attention layer already computes pairwise sim-
ilarity of the atoms’ representations via the normalized inner
product. The self-attention layer (Eq. 1) is then extended to

\[
\tilde{A}(Q, K, V, D) = \varphi(\omega(Q, K, D)V)
\]

\[
= \sum_{j=1}^{k} \varphi(\omega(Q_i, K_j, D_{ij})V_j)
\]

where \( D \) denote the pairwise distance matrix, and \( \varphi \) a
parameterized, potentially learned, similarity function. Sev-
eral formulations can be conceived, however every extension
should satisfy the distance relevance assumption such that \( \varphi \)
is a vanishing mapping with respect to distance \( D_{ij} \), such that
for a given positive \( \delta \) scalar for all \( D_{ij} > \delta \) we would have
\( \varphi(\omega(Q_i, K_j, D_{ij}) \to 0 \).

These assumptions are certainly reminiscent of cut-off dis-
tance (radii) and of many inter-atomic formulations such as
the Van der Waals force or the Axilrod-Teller-Muto potential
[Axilrod and Teller, 1943; Muto, 1943].

Assuming a vanishing mapping \( \omega(D) \) is given and ap-
plied element-wise, several approaches have been proposed
in order to extend the self-attention mechanism. One pop-
ular extension is performed inside the softmax function as in
[Wang and others, 2020] such that

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right),
\]

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

This extension has the ability to fulfill the relevance assump-
tion requirement if the function \( \varphi(\omega(D) \), because of the soft-
max function, assigns negative values to distant atoms (at the
limit \( \lim_{d \to \infty} \omega(D) = -\infty \)). However, such a requirement
can be hard to design or even to learn with parameterized
representation. Another option is to extend the transformer
outside of the normalization such that

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

In this vein and concurrent to our work, [Maziarka et al.,
2020] suggested MAT, a transformer architecture where the
self-attention mechanism is defined as \( \omega(D) = \omega \exp(-D) \),
with hyper-parameter \( \omega \in \mathbb{R} \). This approach clearly fails in
fulfilling the assumption presented above since distant atoms
still impact the self-attention via the softmax similarity.

We propose to directly multiply the distance relation by the
similarity tensor as follows

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]

\[
\varphi(\omega(Q, K_j, D) = \text{Softmax} \left( \frac{QK^T}{\sqrt{d}} + \omega(D) \right)
\]
with \(\odot\) denoting the Hadamard product. This way, the interatomic distance has a direct \textit{gating} impact on the pairwise atomic contribution obtained from the embedding. Such representation is especially important for the augmentation scheme we present later.

Other multiplicative alternatives are possible, such as transferring the Hadamard product inside the softmax [Wu et al., 2021; Fuchs et al., 2020; Kwak et al., 2021]. However there is a computationally demanding and hard to train need to couple the input of \(\psi\) with the similarity tensor values in order to ensure the desired values of the softmax function, i.e., in that case \(\phi_\psi(Q_1, K_1, D_{ij}) \to 0\) for large \(D_{ij}\) implies that \(\psi_\omega(D_{ij}) \to \pm \infty\) and that \(Q_{ij}^T K_j\) and \(\psi_\omega(D_{ij})\) are of opposite signs.

\section*{Learning the Graph Geometry}

Many methods have struggled to model the interaction function \(\psi_\omega\). From force field methods to the recent learning-based approach, there is a need to empirically redefine the Euclidean pairwise distance in order to satisfy physical experimentation and/or performance.

Many handcrafted methods adopt molecular mechanics approximations of un/bonded interactions (e.g. stretch, bending, electrostatic or Van der Waals energies [Christensen and others, 2020]) where molecular properties are obtained via the modification of the interatomic distance, generally using exponential mapping of the distance.

Neural network-based methods adopt a similar approach, where atomic embedding is obtained by modifying the pairwise distance using various learned handcrafted mollifiers, Gaussian radial basis functions, or complex basis in the corresponding function space [Schütt et al., 2018a; Klicpera et al., 2020b; Qiao and others, 2020].

One of the most critical hyper-parameters present in every method is the cut-off distance parameter, connecting only atoms which lie within the cut-off sphere. This hyper-parameter may also change according to the property to be predicted [Schutt et al., 2018b].

Here we propose to learn the pairwise metric \textit{and} the molecule connectivity at each level of the Transformer. The transformation of the Euclidean distance coupled with the self-attention mechanism suggested above allows us to directly optimize the inter-atomic representation as well as the cut-off distance according to the prediction objective, removing cumbersome hyper-parameterization, and allowing soft and differentiable construction of the adjacency matrix learned in a self-adaptive fashion. The similarity function is now simply given by

\[
\varphi_\omega(Q, K, D) = \text{Softmax} \left( \frac{Q K^T}{\sqrt{d}} \right) \odot \psi_\omega(D^{-1})^2, \tag{7}
\]

where \(\psi_\omega : \mathbb{R} \to \mathbb{R}\) is an element-wise learnable function.

We parameterize \(\psi_\omega\) as a shallow, fully connected neural network, and we further enforce the positiveness of the new similarity map by squaring the filtered distances. Transforming the (element-wise) inverse of the distance \(D^{-1}\) instead of \(D\) speeds up the training since \(\psi_\omega\) transforms an already vanishing function (i.e. the multiplicative inverse function).

In contrast with existing methods that use envelopes [Unke and others, 2019; Klicpera et al., 2020b] or distance mollifiers [Schütt et al., 2018a], as well as cut-off parameters, all inducing a \textit{handcrafted} graph connectivity, we optimally unify the learning of the pairwise metric and of the graph adjacency.

We present the learned metrics for several of the Transformer blocks in Figure 4. As can be seen, the metric obtained is both more complex and more abstract than monotonic or Gaussian functions used in previous works [Schütt et al., 2018a; Unke and others, 2019; Maziarka et al., 2020]. It is interesting to notice that some of the obtained cut-off distances lie around values empirically set in other works \((2 - 6)\AA\). The diversified filter bank demonstrates the \textit{dynamic} graph connectivity the network learns via the induced masking of the similarity map.

\section*{Regularization via Molecule Augmentation}

Transformers are generally extremely large and over-parameterized models. Dropout layers commonly used in Transformers in order to avoid overfitting cannot be used in our setting, because of the permutation-invariance requirement as described in [Lee et al., 2019]. One of the most efficient techniques for reducing overfitting is data augmentation. However it is not straightforward to augment molecular data (elements of a set), especially not for regression tasks since modification of one atom type or its spatial positions has unpredictable effects on molecular properties.

While many augmentation methods modify each datum, Mixup strategies [Zhang and others, 2017] intend to create new data samples from pairs (or more) of data. Here, following our initial \textit{distance relevance assumption}, we propose to extend the Mixup idea to molecules where a new data sample is obtained by creating a new system of two molecules positioned far apart.

![Figure 4: 1D cuts of the learned spherical metric at different blocks of the proposed Transformer. The distance is in Å. One can observe different low-, band- and high-pass filters, which demonstrates the model’s ability to learn the connectivity that is relevant for each level of the network. Notice the different dynamic ranges of the filtered values, which lead to different levels of impact on the self-similarity matrix.](image)
Figure 5: The proposed augmentation scheme. Given two molecular systems \( S_1 \) and \( S_2 \), we create a new system \( S_1 \cup S_2 \) by performing random rigid transformation on the system \( S_1 \).

In our mixup scheme, we constrain the property of the new system to be the sum of the properties of the two sub-molecules, even if the predicted property is intensive, i.e., unlike extensive properties it is not physically additive. The sum of the contribution of each atom in the output module allows the model to learn to disentangle the two distanced sub-systems and reduce overfitting.

Formally, for a given property \( p \) and given two centered molecules \( M_i, M_j \) such that \( M_k = \{z^k, X^k\} \) we have

\[
\begin{align*}
\tilde{X}^i &:= \{R \cdot x_1 + T, \ldots, R \cdot x_{N_j} + T\} \\
M_{ij} &:= M_i \cup M_j := \{z = \{z^i, z^j\}, X = \{X^i, \tilde{X}^j\}\},
\end{align*}
\]

with a rotation matrix \( R \in \text{SO}(3) \), and \( T = t \cdot 1, t \in \mathbb{R} \) a spatial translation vector ensuring large enough distance between the molecules so that the interaction is null or negligible (e.g. \( t > 10^3 \AA \)). Thus, based on inter-atomic distances, we want our model to be able to differentiate the two systems such that the target properties \( p(M_i) \) and \( p(M_j) \) sum, thus

\[
p(M_{ij}) = p(M_i \cup M_j) = p(M_i) + p(M_j).
\]

An illustration of the augmentation scheme is given in Figure 5. We note that methods relying on cut-off distances cannot use this augmentation strategy in a straightforward manner since the cut-off parameters would automatically separate the two molecules. Also, this approach can be extended to more than just pairs of atoms. The main drawback of this method is an increase in training time since the created molecules can be up to twice as large as the largest molecule of the dataset. However, this augmentation method enables a significant improvement of generalization, even on relatively small datasets, as demonstrated in the Discussion section.

3 Experiments

The experimental setup including the architecture details and the training procedure is provided in the Appendix.

QM9 The popular QM9 dataset [Ramakrishnan et al., 2014] contains 130, 831 molecules with up to 9 atoms of the type C,N,O, and F saturated with hydrogen atoms in their equilibrium geometries, with chemical properties computed with DFT solvers. Following previous work, we split the dataset to 110, 000, 10, 000 and 10, 831 molecules for the training, validation and testing sets respectively. We use the atomization energy for \( U_0, U, H, \) and \( G \).

In Table 1 (left) we report the mean absolute error (MAE) on all QM9 targets and compare it to the state-of-the-art models SchNet [Schütt et al., 2018a], PhysNet [Unke and others, 2019], MGCN [Lu et al., 2019], Cormorant [Anderson et al., 2019], and DimeNet [Klicpera et al., 2020b]. We also compare with concurrent molecular data Transformers, R-MAT [Maziarka et al., 2021], 3D-T [Wu et al., 2021], SE(3)-T [Fuchs et al., 2020] and GeoT [Kwak et al., 2021].

The method outperforms or is similar to state-of-the-art methods for most of the properties and surpasses DimeNet by 5.22% average performance ratio, and other Transformers methods by large margins. In contrast to other works [Schutt et al., 2018b; Klicpera et al., 2020a], the same model and training procedure were applied for all properties. The proposed framework does not require tuning of physical hyperparameters or chemical approximations, making it a true end-to-end framework.

The recent SOTA work DimeNet++ [Klicpera et al., 2020a] substantially outperforms DimeNet (by 9% in average) with careful initialization and architectural modifications, and outperforms our framework by 0.78% (average performance ratio). We believe that a similar thorough architecture search may have similar impact on the proposed approach, depending on available computational resources since Transformers are computationally intensive. It is also important to notice that the proposed method does not take into account computationally heavy bond angles between triplets of atoms (i.e. potentially inducing cubic complexity) as in many recent frameworks.

MD17 We use MD17 [Chmiela et al., 2017] to test model performance in molecular dynamics simulations. The goal of this benchmark is to predict, for eight small organic molecules, the Cartesian atomic forces acting on each atom due to the overall potential energy. A separate model is to be trained for each molecule, in order to provide accurate individual predictions. We test our model in the challenging 1000 training samples setting [Chmiela et al., 2018]. The original training objective is extended to molecular dynamics predictions by backpropagating to the atom coordinates \( X \) as follows

\[
\mathcal{L} = ||f_\theta(z, D) - p||_1 + \frac{\rho}{3} ||- \partial_X f_\theta(z, X) - F||_1,
\]

where \( F \) denotes the nuclear three-dimensional Cartesian forces to be predicted and \( \rho \) is the forces’ loss coefficient. In our experiments \( \rho \) is set to \( 10^3 \), the augmentation scheme is extended straightforwardly to forces (i.e. concatenation) due to their translation invariance, and ReLU activations are replaced with GELU non-linearities to enforce twice continuous differentiability. As shown in Table 1 (right) our framework sets or reaches SOTA performances even when the training size remains extremely small, a challenging setting for Transformer models. It demonstrates our method’s flexibility and its ability to generalize to other tasks and datasets.
3.1 Discussion

Geometric Self-attention Analysis Figure 6, presents typical impact of the proposed self-attention mechanism on the similarity map. As can be seen, different filters applied to the pairwise distance have a major impact on the similarity matrix and redefine the adjacency matrix at each level of the network. One can observe the drastic impact (color flipping) of the learned metric on the similarity map.

Comparison and Ablation Studies Our ablation study compares typical impact of the different self-attention modules. We present the convergence curves of our method from Eq. (7), the concurrent Transformers based MAT method [Maziarik et al., 2020] from Eq. (5), the MAT method with a learned metric $\psi_{\text{learned}}$ and the sum self-attention from Eq. (4) [Wang and others, 2020] with learned metric: we refer to the last method as SUM SA. The compared networks and the training procedure are exactly the same, except for the aforementioned self-attention equation itself and the distance mapping module. The results are presented in Figure 7 (left). As can be seen, the MAT architecture presents the worse convergence, while the metric learning module significantly improves the performance. The proposed self-attention mechanism greatly surpasses all other architectures.

Finally, we present the typical effect of the data augmentation procedure on the generalization of the network during training. Figure 7 (right) presents the impact of the augmentation on the MAE convergence of the model for both an extensive and intensive property, namely $U_0$ and $\mu$. As can be seen, in both cases, when applying augmentation, the generalization gap is extremely reduced while the validation loss is much lower. The gain for the property $\mu$ is $175\%$ in terms of testing MAE and $75\%$ for $U_0$.
References

[Anderson et al., 2019] B Anderson, T Son Hy, and R Kon- dor. Cormorant: Covariant molecular neural networks. arXiv:1906.04015, 2019.

[Axilrod and Teller, 1943] BM Axilrod and Ei Teller. Interaction of the van der waals type between three atoms. The Journal of Chemical Physics, 1943.

[Bartok et al., 2010] A Bartok, M Payne, et al. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical review letters, 2010.

[Belkin and Niyogi, 2003] M Belkin and P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 2003.

[Bello et al., 2019] I Bello, B Zoph, A Vaswani, et al. Attention augmented convolutional networks. In ICCV, 2019.

[Chmiela et al., 2017] S Chmiela, A Tkatchenko, et al. Machine learning of accurate energy-conserving molecular force fields. Science, 2017.

[Christensen and others, 2020] A Christensen et al. Fchl re-visited: Faster and more accurate quantum machine learning. The Journal of Chemical Physics, 2020.

[Dwivedi and Bresson, 2020] VP Dwivedi and X Bresson. A generalization of transformer networks to graphs. arXiv:2012.09699, 2020.

[Fuchs et al., 2020] F B Fuchs, D E Worrall, V Fischer, and M Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. arXiv:2006.10503, 2020.

[Goyal and Bengio, 2020] A Goyal and Y Bengio. Induc- tive biases for deep learning of higher-level cognition. arXiv:2011.15091, 2020.

[Huang and others, 2016] Bing Huang et al. Communica- tion: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. J. of Chemical Physics, 2016.

[Jensen, 2017] Frank Jensen. Introduction to computational chemistry. John wiley & sons, 2017.

[Klicpera et al., 2020a] J Klicpera, S Giri, JT Margraf, et al. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. arXiv:2011.14115, 2020.

[Klicpera et al., 2020b] J Klicpera, J Groß, and S Günemann. Directional message passing for molecular graphs. arXiv:2003.03123, 2020.

[Kwak et al., 2021] B Kwak, J Jo, B Lee, and S Yoon. Geometry-aware transformer for molecular property prediction. arXiv:2106.15516, 2021.

[Lee et al., 2019] J Lee, Y Lee, J Kim, et al. Set transformer: A framework for attention-based permutation-invariant neural networks. In ICML, 2019.

[Lu et al., 2019] C Lu, Q Liu, C Wang, et al. Molecular property prediction: A multilevel quantum interactions modeling perspective. In AAAI, 2019.

[Mazziarka et al., 2020] L Mazziarka, T Danel, S Mucha, et al. Molecule attention transformer. arXiv:2002.08264, 2020.

[Mazziarka et al., 2021] L Mazziarka, D Majchrowski, T Danel, et al. Relative molecule self-attention transformer. arXiv:2110.05841, 2021.

[Muto, 1943] Yoshio Muto. Force between nonpolar molecules. J. Phys. Math. Soc. Japan, 1943.

[Qiao and others, 2020] Z Qiao et al. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. J. of Chemical Physics, 2020.

[Ramakrishnan et al., 2014] R Ramakrishnan, P Dral, M Rupp, et al. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 2014.

[Schütt et al., 2018a] K Schütt, H Sauceda, P-J Kindermans, et al. Schnet--a deep learning architecture for molecules and materials. The Journal of Chemical Physics, 2018.

[Schütt et al., 2018b] KT Schutt, P Kessel, et al. Schnetpack: A deep learning toolbox for atomistic systems. Journal of chemical theory and computation, 2018.

[Unke and others, 2019] O Unke et al. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. J. chemical theory and computation, 2019.

[Vaswani et al., 2017] A Vaswani, N Shazeer, N Parmar, et al. Attention is all you need. In NeurIPS, 2017.

[Wang and others, 2020] H Wang et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In ECCV, 2020.

[Wu et al., 2021] F Wu, Q Zhang, D Radev, et al. 3d-transformer: Molecular representation with transformer in 3d space. arXiv:2110.01191, 2021.

[Zhang and others, 2017] H Zhang et al. Mixup: Beyond empirical risk minimization. arXiv:1710.09412, 2017.