Geometric Transformer for End-to-End Molecule Properties Prediction

Yoni Choukroun , **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 interaction between elements. However, extending the Transformer architecture to non-sequential data such as molecules and enabling its training on small datasets remain 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 beside 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 *et al.*, 2020; Huang and Von Lilienfeld, 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 *et al.*, 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 Meuwly, 2019; Qiao *et al.*, 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. Our paper's main contributions are:

- An initial positional encoding based on a learned mapping of the global embedding at the atomic level, which is able to encode the molecule geometry.

- A new self-attention module where the positional encoding is inserted as an invariant *gating*, enabling better realisation of the inter-atomic assumption.
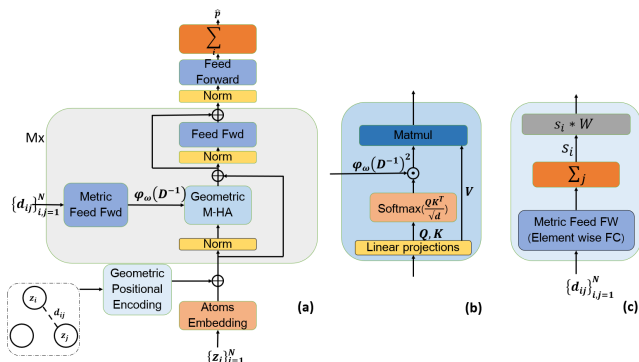
Figure 1: The proposed Transformer architecture (a). The main architectural modifications are the initial positional encoding based on the pairwise distance matrix (c), and the metric learning module coupled with the augmented self-attention module (b). The model is composed of M encoding blocks and the output block is composed of a normalized feed-forward neural network followed by a per atom i summation to accumulate the contribution of each atom.

- This self-attention mechanism is coupled with a learned inter-atomic pairwise *metric* in order to learn adaptively the molecule's soft adjacency matrix.

- An augmentation technique for molecular data based on rigid transformations. This regularization approach can be used for permutation-invariant models where other regularization techniques such as dropout cannot be applied in a straightforward manner.

- A novel end-to-end Transformer architecture setting state-of-the-art performance, outperforming all other Transformer based methods by a large margin. In contrast to existing methods, the proposed approach does not require any external information or knowledge, directly demonstrating the power and flexibility of suitably designed Transformers.

2 Related Work

Classical molecule property prediction methods combined handcrafted features generally based on force field methods at the atom level [Christensen *et al.*, 2020] or molecular level [Huang and Von Lilienfeld, 2016], integrated into various machine learning models such as Kernel methods [Christensen *et al.*, 2020] , Gaussian processes [Bartok *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 Meuwly, 2019; Anderson *et al.*, 2019; Klicpera *et al.*, 2020b; Qiao *et al.*, 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 Meuwly, 2019], bending or torsion angles [Klicpera *et al.*, 2020b], or handcrafted atomic features derived from quantum mechanics [Qiao *et al.*, 2020]. Existing methods generally involve a deep understanding

and a cautious adaptation of the underlying physics in order to provide better preconditioning [Klicpera *et al.*, 2020b; Qiao *et al.*, 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.

3 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_θ 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, bonds type, hybridization [Gilmer *et al.*, 2017; Maziarka *et al.*, 2020], DFT solver [Unke and Meuwly, 2019; Qiao *et al.*, 2020]), neither any kind of knowledge adaptation from quantum physics/chemistry [Unke and Meuwly, 2019; Klicpera *et al.*, 2020b; Klicpera *et al.*, 2020a; Qiao *et al.*, 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.

3.1 Transformer

Transformer was introduced by [Vaswani *et al.*, 2017] as a novel, attention-based building block for machine translation.
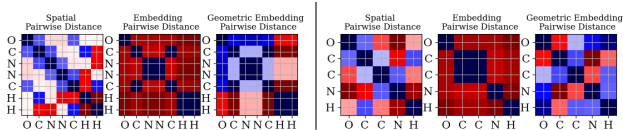
Figure 2: The impact of the initial positional encoder on the embedding for two different molecules (left and right). We show the pairwise distance matrix D (left), pairwise distance of the initial embedding $Emb(z)$ (middle), and the pairwise distance of the final geometric embedding (right). The marks on the map represent the type of the atoms z_i. Cold and warm colors represent low and high values respectively.
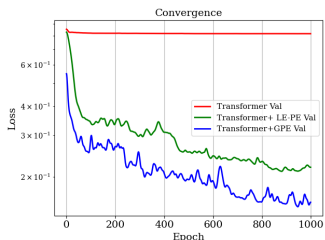

Figure 3: The validation L_1 loss (U0 property) of a regular transformer without any geometric input (red), the same transformer with our Laplacian extension (green), and with the proposed initial geometric positional encoding (GPE).

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.

3.2 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 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 positional information to each atom. For each atom we first embed 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 $Emb(z_i) : \mathbb{Z}_+ \to \mathbb{R}^d$ and the pairwise distance matrix $D \in \mathbb{R}^{+^{n \times n}}$ such that $(D)_{ij} := (D_i)_j = d_{ij}$, we have

$$y_i = Emb(z_i) + W \sum_j f_{\text{pos}}(d_{ij}). \quad (2)$$

Here, y_i denotes the obtained initial positioned embedding of atom z_i, $f_{\text{pos}} : \mathbb{R}^+ \to \mathbb{R}$ 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 positional encoding brings information about the geometry. The convergence plot in Figure 3 demonstrates that the proposed positional encoding allows the transformer to learn the molecular 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. Our approach allows to encode the global geometry of the molecule with respect to each atom, while the permutation-invariant aggregation part in Eq. (2) maintains the symmetry invariance of the embedding. This preconditioning enables faster convergence and slightly better performance.

3.3 Geometric Self-Attention

In order to maintain the invariant properties of molecules, we propose to augment the initial positional 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 similarity of the atoms' representations via the normalized inner product. The self-attention layer (Eq. 1) is then extended to

$$\left(\tilde{A}(Q, K, V, D)\right)_i = \varphi_\omega(Q_i, K, D_i)V$$
$$= \sum_{j=1}^{k} \varphi_\omega(Q_i, K_j, D_{ij})V_j, \quad (3)$$

where D denote the pairwise distance matrix, and φ_ω a parameterized, potentially learned, similarity function. Several formulations can be conceived, however every extension should satisfy the *distance relevance assumption* such that φ_ω is a vanishing mapping with respect to distance D_{ij}, such that for a given positive δ 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 distance (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 $\psi_\omega(D)$ is given and applied element-wise, several approaches have been proposed in order to extend the self-attention mechanism. One popular extension is performed inside the softmax function as in [Wang *et al.*, 2020] such that

$$\varphi_\omega(Q, K_j, D) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + \psi_\omega(D)\right), \quad (4)$$

This extension has the ability to fulfill the *relevance assumption* requirement if the function $\psi_\omega(D)$, because of the softmax function, assigns negative values to distant atoms (at the
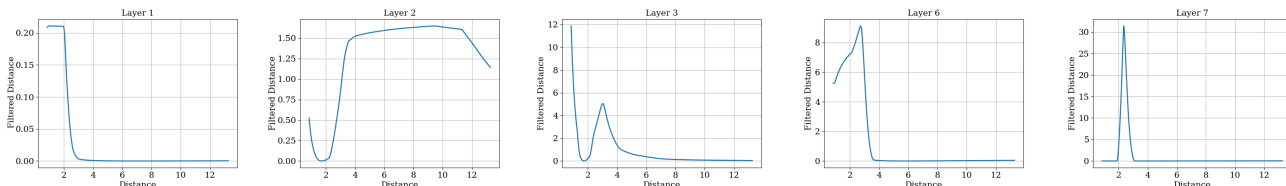
Figure 4: 1D cuts of the learned spherical metric at different blocks of the proposed Transformer. The distance is in Å. One can observe different low-, band- and high-pass filters, which demonstrates the model's ability to learn the connectivity that is relevant for each level of the network. Notice the different dynamic ranges of the filtered values, which lead to different levels of impact on the self-similarity matrix.

limit $\lim_{d\to\infty} \psi_\omega(d) = -\infty$). 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

$$\varphi_\omega(Q, K_j, D) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) + \psi_\omega(D). \quad (5)$$

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 $\psi_\omega(D) = \omega\exp(-D)$, 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

$$\varphi_\omega(Q, K_j, D) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \odot \psi_\omega(D), \quad (6)$$

with $\odot$ denoting the Hadamard product. This way, the interatomic distance has a direct *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 ψ with the similarity tensor values in order to ensure the desired values of the softmax function, i.e., in that case $\phi_\omega(Q_i, K_j, D_{ij}) \to 0$ for large D_{ij} implies that $\psi_\omega(D_{ij}) \to \pm\infty$ and that $Q_i^T K_j$ and $\psi_\omega(D_{ij})$ are of opposite signs.)

3.4 Learning the Graph Geometry

Many methods have struggled to model the interaction function ψ_ω. 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 *et al.*, 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 *et al.*, 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 *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

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

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

We parameterize ψ_ω 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 ψ_ω transforms an already vanishing function (i.e. the multiplicative inverse function).

In contrast with existing methods that use envelopes [Unke and Meuwly, 2019; Klicpera *et al.*, 2020b] or distance mollifiers [Schütt *et al.*, 2018a], as well as cut-off parameters, all inducing a *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 Meuwly, 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 ($2 - 6$Å). The diversified filter bank demonstrates the *dynamic* graph connectivity the network learns via the induced masking of the similarity map.

3.5 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
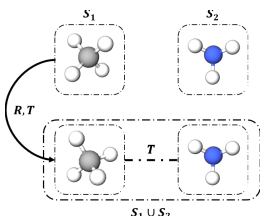
Figure 5: The proposed augmentation scheme. Given two molecular systems S_1 and S_2, we create a new system $S_1 \cup S_2$ by performing random rigid transformation on the system S_1.

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 *et al.*, 2017] intend to create new data samples from pairs (or more) of data. Here, following our initial *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.

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

$$\tilde{X}^j := \{R \cdot x_1 + T, \dots, R \cdot x_{N_j} + T\}$$
$$M_{ij} = M_i \cup M_j := \left\{ z = \{z^i, z^j\}, X = \{X^i, \tilde{X}^j\} \right\}, \quad (8)$$

with a rotation matrix $R \in \mathrm{SO}(3)$, and $T = t \cdot \mathbb{1}, t \in \mathbb{R}$ a spatial translation vector ensuring large enough distance between the molecules so that the interaction is null or negligible (e.g. $t > 10^3$Å). 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

$$p(M_{ij}) = p(M_i \cup M_j) = p(M_i) + p(M_j). \quad (9)$$

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.

4 Experiments

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

4.1 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 $U0, U, H$, and G.

In Table 1 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 Meuwly, 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 hyper-parameters 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.

4.2 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

$$\mathcal{L} = ||f_\theta(z, D) - p||_1 + \frac{\rho}{3}|| - \partial_X f_\theta(z, X) - F||_1, \quad (10)$$

where F denotes the nuclear three-dimensional Cartesian forces to be predicted and ρ is the forces' loss coefficient. In our experiments ρ 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 2 our framework

Table 1: MAE on QM9. Best in bold and second underlined. Transformer based architectures are to the right of the vertical line.

Target	Unit	Schnet	Physnet	MGCN	Cormorant	DimeNet	SE(3)-T	R-MAT	3D-T	GeoT	Our
μ	D	0.0330	0.0529	0.0560	0.038	<u>0.0286</u>	0.051	0.110	0.045	0.0297*	**0.0264**
α	a_0^3	0.235	0.0615	0.085	0.0681	**0.0469**	0.142	0.082	0.086	0.052	<u>0.051</u>
ϵ_{HOMO}	meV	41.0	32.9	42.1	34	27.8	35	31	**21**	<u>25*</u>	27.5
ϵ_{LUMO}	meV	34.0	24.7	57.4	38	**19.7**	33	29	26	20.2*	<u>20.4</u>
Δ_ϵ	meV	63.0	42.5	64.2	38	**34.8**	53	48	39	43	<u>36.1</u>
$\langle R^2 \rangle$	a_o^3	**0.073**	0.765	0.110	0.961	0.331	-	0.676	-	0.30	<u>0.157</u>
ZPVE	meV	1.70	1.39	1.12	2.03	<u>1.29</u>	-	2.23	-	1.7*	**1.24**
U_0	meV	14.0	8.15	12.9	22	<u>8.02</u>	-	12	-	11.1	**7.35**
U	meV	19.0	8.34	14.4	21	<u>7.89</u>	-	10	-	11.7	**7.55**
H	meV	14.0	8.42	14.6	21	<u>8.11</u>	-	10	-	11.3	**7.73**
G	meV	14.0	9.40	16.2	20	<u>8.98</u>	-	10	-	11.7	**8.21**
c_v	$\frac{cal}{mol\,K}$	0.0330	0.0280	0.0380	<u>0.026</u>	**0.0249**	0.054	0.036	-	0.0276	0.0280

Table 2: MAE on MD17 forces using 1000 training samples.

Target	Schnet	DimeNet	GeoT	Our
Aspirin	1.35	<u>0.499</u>	0.85	**0.451**
Benzene	0.31	<u>0.187</u>	**0.135**	0.28
Ethanol	0.39	0.230	<u>0.225</u>	**0.212**
Malonaldehyde	0.66	<u>0.383</u>	0.402	**0.369**
Naphthalene	0.58	**0.215**	-	<u>0.44</u>
Salicylic acid	0.85	**0.374**	-	0.372
Toluene	0.57	**0.216**	0.328	<u>0.24</u>
Uracil	0.56	**0.301**	-	0.301

sets or reaches SOTA performances even when the training size remains extremely small, a challenging setting for Transformer models. Also, it demonstrates our method's flexibility and its ability to generalize to other tasks and datasets.

4.3 Discussion

In this section, we study the contributions of our method.

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.
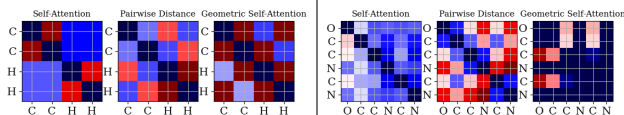

Figure 6: Regular self-attention similarity map (left), the pairwise distance matrix (center), and the resulting geometric self-attention similarity map (right) of two different molecules (left and right) at layers 2 and 9 respectively. The similarity maps are averaged over the dimensions of the self-attention heads.

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 [Maziarka *et al.*, 2020] from Eq. (5), the MAT method with a learned metric ψ_ω, and the sum self-attention from Eq. (4) [Wang *et al.*, 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 $U0$ and μ. 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 μ is 175% in terms of testing MAE and 75% for $U0$.
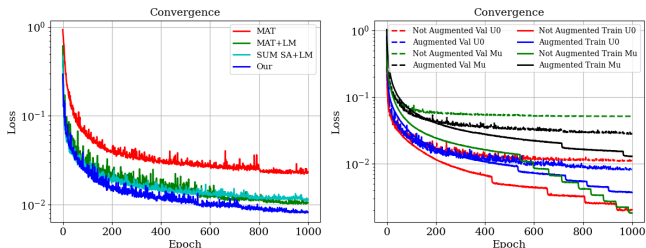

Figure 7: Left: Comparison of validation MAE losses between MAT, MAT with the proposed learned metric (MAT+LM), the summed self-attention mechanism with learned metric (SUM SA +LM), and our method. Right: The impact of data augmentation on the MAE convergence and generalization of the network (val=validation) for $U0$ and μ. Continuous lines are consistently below their corresponding dashed ones.

5 Conclusion

We introduce a new Transformer architecture and a training scheme for molecular predictions. The proposed model allows effective representation of interactions based solely on pairwise distances. The graph geometry and connectivity can be learned in a soft fashion by the network via the geometric self-attention module. We further propose a molecular data augmentation procedure based on mixing strategies, which leads to a clear improvement in generalization for the proposed scheme. Our results indicate that our method is the first Transformer, as far as we can ascertain, that is able to model molecular data successfully without requiring any assumptions on the underlying physical model or involving complex geometric priors. We believe the advent of new datasets and new transformer architectures will allow the development of more efficient models also less prone to overfitting, making it a tool of predilection for the analysis of molecular data.

Acknowledgments

References

[Anderson *et al.*, 2019] B Anderson, T-S Hy, and R Kondor. 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.

[Chen *et al.*, 2017] Q Chen, X Zhu, Z Ling, et al. Neural natural language inference models enhanced with external knowledge. *arXiv:1711.04289*, 2017.

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

[Chmiela *et al.*, 2018] S Chmiela, H Sauceda, K-R Müller, et al. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature*, 2018.

[Christensen *et al.*, 2020] A Christensen, L Bratholm, F Faber, et al. Fchl revisited: 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.

[Gilmer *et al.*, 2017] J Gilmer, S Schoenholz, P Riley, O Vinyals, and G Dahl. Neural message passing for quantum chemistry. *arXiv:1704.01212*, 2017.

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

[Huang and Von Lilienfeld, 2016] B Huang and O A Von Lilienfeld. Understanding molecular representations in machine learning: The role of uniqueness and target similarity, 2016.

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

[Kingma and Ba, 2014] D Kingma and J Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[Klein *et al.*, 2017] G Klein, Y Kim, et al. Open-source toolkit for neural machine translation. In *ACL*, 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 Günnemann. 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.

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

[Maziarka *et al.*, 2021] L Maziarka, 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 *et al.*, 2020] Z Qiao, M Welborn, Anandkumar, et al. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal 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.

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

[Shazeer, 2020] Noam Shazeer. Glu variants improve transformer. *arXiv:2002.05202*, 2020.

[Unke and Meuwly, 2019] O Unke and M Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of 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 *et al.*, 2020] H Wang, Y Zhu, B Green, 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.

[Xiong *et al.*, 2020] R Xiong, Y Yang, D He, et al. On layer normalization in the transformer architecture. *arXiv:2002.04745*, 2020.

[Zhang *et al.*, 2017] H Zhang, M Cisse, et al. mixup: Beyond empirical risk minimization. *arXiv:1710.09412*, 2017.

A Experimental Setup

We used ten encoding blocks with an embedding size of $d = 512$ and with a dimension of 2048 for the inner layer of the feedforward network. The feedforward network is composed of GEGLU layers [Shazeer, 2020] and layer normalization is set in pre-layer norm setting as in [Klein *et al.*, 2017; Xiong *et al.*, 2020]. The contribution of the atom itself (i.e. $\varphi_\omega(Q_i, K_i, D_{ii})$) is omitted (masked) in the self-attention mechanism. The metric feed-forward module is a fully connected neural network with a 50-dimensional hidden layer and ReLU non-linearities in order to simulate distance thresholding, expanded to all the heads of the self-attention module. The geometric positional encoder is a fully connected network with one 1024 dimensional layer coupled with GELU non-linearity. The output module is a linear layer. As other methods, we train our model once for every target.

The Adam optimizer [Kingma and Ba, 2014] is used with 32 molecules per mini-batch for the QM9 dataset and 4 molecules only for the MD17 dataset. Half of each shuffled batch is augmented using the procedure presented in the previous section with translation scalar t set to 10^4. We initialized the learning rate to 10^{-4} coupled with a plateau region decay scheduler with ratio 0.8 down to a 10^{-6} threshold. No warmup has been employed [Xiong *et al.*, 2020] and the L_1 loss is used as the objective metric (Mean Absolute Error). The model has been implemented upon the Schnetpack framework [Schutt *et al.*, 2018b].