ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

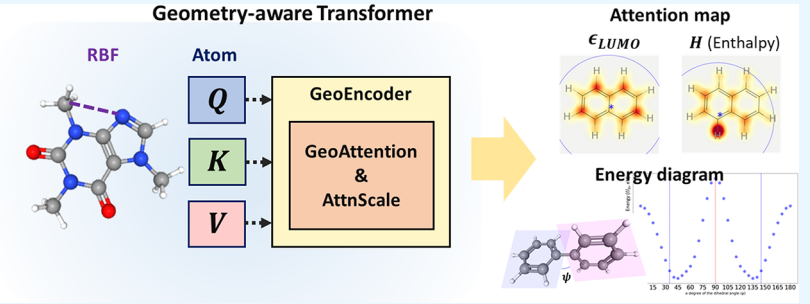
GeoT: A Geometry-Aware Transformer for Reliable Molecular Property Prediction and Chemically Interpretable Representation Learning

Bumju Kwak, Jiwon Park, Taewon Kang, Jeonghee Jo,* Byunghan Lee, and Sungroh Yoon*

Cite This: *ACS Omega* 2023, 8, 39759−39769

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: In recent years, molecular representation learning has emerged as a key area of focus in various chemical tasks. However, many existing models fail to fully consider the geometric information on molecular structures, resulting in less intuitive representations. Moreover, the widely used message passing mechanism is limited to providing the interpretation of experimental results from a chemical perspective. To address these challenges, we introduce a novel transformer-based framework for molecular representation learning, named the geometry-aware transformer (GeoT). The GeoT learns molecular graph structures through attention-based mechanisms specifically designed to offer reliable interpretability as well as molecular property prediction. Consequently, the GeoT can generate attention maps of the interatomic relationships associated with training objectives. In addition, the GeoT demonstrates performance comparable to that of MPNN-based models while achieving reduced computational complexity. Our comprehensive experiments, including an empirical simulation, reveal that the GeoT effectively learns chemical insights into molecular structures, bridging the gap between artificial intelligence and molecular sciences.

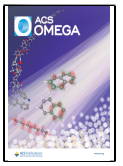
INTRODUCTION

Quantum mechanical calculations have been used in the development of chemicals in various fields such as drugs and catalysts. Density functional theory (DFT) is the most widely used computational methods of quantum mechanics (QM) modeling,[1] but it requires a great deal of computation to predict the properties of even a small molecule. For this reason, several machine learning-based techniques have been explored as cost-effective alternatives.[2−5] In particular, deep learning has been used to predict molecular properties including energy and forces.[6−12]

It is common to regard a molecule as a graph in which the atoms are nodes and the edges are bonds in a message-passing neural network (MPNN).[13] SchNet,[7] PhysNet,[8] and several other MPNNs constructed the localized messages that are centered on atoms by applying continuous filters based on interatomic distances. More recent networks such as DimeNet,[11] DimeNet++,[14] and GemNet[15] explicitly incorporate angle computation to represent molecular conformations. However, these previous models use a cutoff distance, which restricts the receptive field of atom-based convolutions, resulting in computationally intensive operations.

An MPNN's localized view has inherent limitations. From a chemical point of view, all atom−atom pairs should be considered as essential components in molecular property prediction models, regardless of their interatomic distances. However, the localized convolutions of MPNN conflict with the conformational behavior of molecules since the MPNN with a finite cutoff value assumes that messages are transferred within a restricted region. The use of a fixed cutoff distance fails to capture long-range interatomic relationships in which the actual interactions between all the pairs of atoms in a molecule are determined by charge and distance.

Received: August 6, 2023
Accepted: September 21, 2023
Published: October 9, 2023



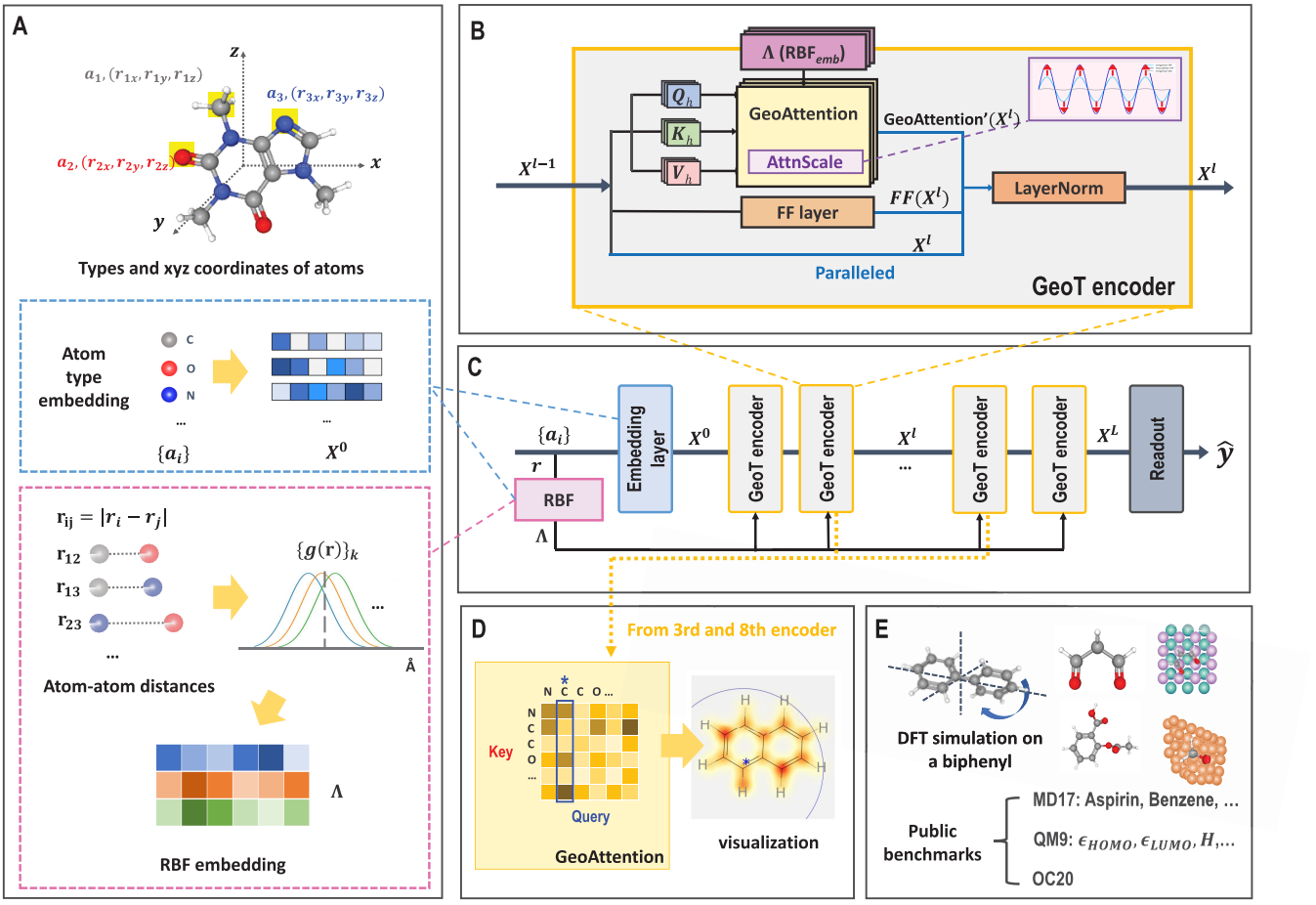


Figure 1. Overview of a geometry-aware transformer. (A) Embedding of a type of each atom and its position of a molecule as an input. A set of atom types $\{a_i\}$ are embedded as inputs X_0 of GeoT in the learnable embedding layer. On the other hand, a set of atom−atom distances are represented as k different RBFs (RBF_{emb}), to make the $\mathbf{\Lambda}$-embedding matrix. Different from the embedding layer for atom types, the distance embedding $\mathbf{\Lambda}$ from RBF_{emb} is used as an input of each GeoT encoder, which is a block of GeoT. (B) Internal structure of the GeoT encoder accompanied by GeoAttention. The query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ are obtained from the former layer X^{l-1}, and matrix $\mathbf{\Lambda}$ is the input of GeoAttention before applying the AttnScale. In addition, the components of each GeoT encoder are parallelized, which is different from the original self-attention. (C) Architecture of GeoT. It has one atom type-embedding layer at the front, L GeoT encoder blocks, and a readout layer at the end of the network. To extract attention feature maps for visualization, we used the third and eighth GeoT encoders. (D) Example of attention map visualization from GeoAttention. Given a query atom (denoted as a blue asterisk), associated attention weights of other atoms are shown as yellow-red shades. More red regions have higher attention weights associated with a given query atom. (E) Benchmarks used for model performance: MD17 and QM9 datasets consist of small-sized molecules, while the OC20 dataset is a pair of surface and substrate.

To overcome these limitations, we adopt transformers[16] as our model framework for molecular graphs. The transformer consists of self-attention blocks learning the relations between two components from the sequence. If a graph can be represented as a sequence without loss of its topological information, a transformer can be a viable alternative to the MPNN. The self-attention mechanism in transformers can learn relationships between entities regardless of their positions, whereas MPNNs are limited to localized neighbors due to their restricted receptive fields.

Since transformers were initially developed for processing sequential data with a specific order, they are generally not well-suited for handling order-invariant graph data. To overcome this problem, several previous transformer-based architectures[10,17−20] were proposed; however, these methods cannot fully consider geometric information, which is the key factor of describing the nature of molecules. For example,[10,18,19] it blindly encodes an atom−atom interaction as a categorized representation of the bond type. There are two problems with this categorization. First, categorization of bond types cannot consider its individual length information, which depends on the associated atom types. Second, it does not allow for the consideration of cases where atoms are not bonded but are closely located to each other and moreover[10,18] require excessive prior knowledge such as valence or aromaticity for molecular property prediction, which can be a limitation for exploring little-known molecules.

To address the above-mentioned issues, we aim to develop a model considering the nature of molecules represented by graphs with geometric information, named geometry-aware transformer (GeoT). The overview of the GeoT architecture is presented in Figure 1. To achieve the interpretable prediction results, we introduce several modifications on the self-attention mechanism in GeoT. By incorporating these concerns, we propose two strategies into self-attention: (1) introducing k radial basis function (RBF_{emb}) for embedding of geometric information on atom−atom distances and (2) replacing the softmax with the alternative scaling method to enhance more important atom features. We named our modified self-attention as GeoAttention.

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

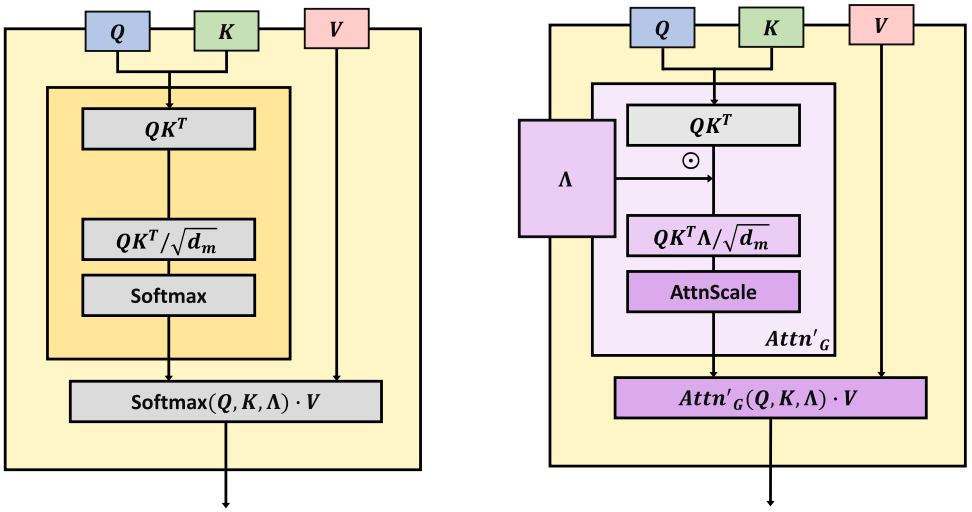


Figure 2. Comparison between the standard self-attention (left) and GeoAttention (right). The key differences of two architectures are that the GeoAttention 1) used the RBF_{emb} to represent distances of atom pairs and 2) removed the softmax function existing in the standard self-attention.

The modifications described above are based on intuition that different parts of a molecule contribute differently to various molecular properties. To encourage the model to attend to the most significant parts of a molecule, it is not desirable to use softmax, which indiscriminately adjusts all entities to the same scale.

We investigated the relationship between the GeoT's attention pattern on different molecules and the associated chemical insights of the training objective to verify our intuition. Specifically, we investigated two contrasting molecular properties, ϵ_{LUMO} and H cases, which are the energy of the lowest unoccupied molecular orbital (LUMO) and molecular enthalpy, respectively. Surprisingly, we found a salient pattern from GeoAttention, which is the modified self-attention in GeoT trained with ϵ_{LUMO}. Additionally, we conducted the DFT simulation in the case of five molecules including a biphenyl by predicting on its energy over the change of conformation and verified superior generalizability of the GeoT.

We also evaluated the prediction performance on the public benchmarks. First, we evaluated the prediction performance of the GeoT on three public benchmark datasets MD17,[21−23] QM9,[24,25] and OC20,[26] which are the most widely used molecular property prediction tasks in current days. Second, we did an ablation study of the proposed model refinements to verify the effectiveness of the GeoT. By comparing the results from those experiments, we concluded that GeoAttention has additional advantages. We summarize the contributions of GeoT below.

- Development of a geometry-aware transformer architecture for molecular graphs, which incorporates geometric information of molecules for intuitive and interpretable representation learning.
- Introduction of the model refinement strategies enabling molecular graph representations to improve model performance and robustness.
- Verification of better interpretability and generalizability produced by the GeoAttention from a chemical perspective.
- Evaluation of the model performance over a range of benchmarks and the empirical study for molecular property prediction.

RESULTS

We analyzed the performance of the GeoT in both a qualitative and quantitative manner. For qualitative evaluation of the interpretability of the GeoT, we visualize the trained attention score of GeoAttention in various cases. Visualization of the weights in self-attention has widely been used to provide the semantic relationships between different elements of data.[27,28] As a quantitative evaluation of the GeoT, the public benchmark for molecular property prediction MD17, QM9, and OC20 was used.

Interpretative Analysis of GeoAttention Visualization from the Chemical Perspective. First, we provide a concise overview of RBF_{emb} and AttnScale, which are the main components for understanding the mechanism of GeoAttention. The RBF_{emb} consists of multiple k Gaussian functions with different centers to embed an atom−atom distance as a k-dimensional vector. This vector $\{\boldsymbol{g}(\mathbf{r})\}_k$ is then added to the sum of two atom-embedding vectors, $Z(a_i) + Z(a_j)$, to make $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times k}$. Accordingly, the query atom vector $\boldsymbol{Q}$, the key atom vector $\boldsymbol{K}$, and $\boldsymbol{\Lambda}$ are multiplied together, to make $\mathrm{Attn}_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})$. Second, we introduced AttnScale to intensify the high-frequency (HF) signal A'_{HF} feature from $\mathrm{Attn}_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})$. The HF signal in a molecule can be viewed as the interactions between two atoms located close to each other. The result is $\mathrm{Attn}'_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda}) = \mathrm{Attn}_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda}) + A'_{HF}$. Finally, the formulas of GeoAttention of the GeoT are as follows. We illustrated it in Figure 2, and further details are described in the Methods section.

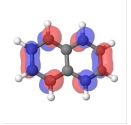
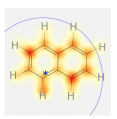
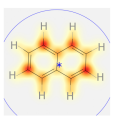
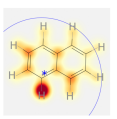
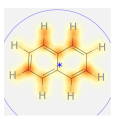
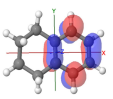
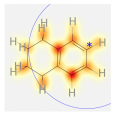
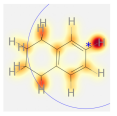
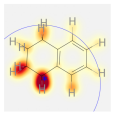
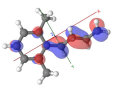
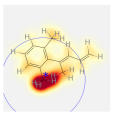
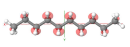
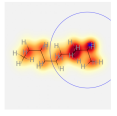
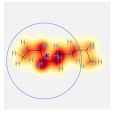
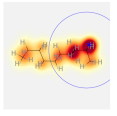
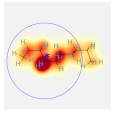
$$\mathrm{GeoAttention}'(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{\Lambda}) = \mathrm{Attn}'_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})\cdot\boldsymbol{V} \quad (1)$$

We selected six molecules having different molecular geometries (ring or linear) and conjugation patterns. For these molecules, we report each pattern of attention distributions produced by GeoT trained on two types of label ϵ_{LUMO} and H of the QM9 dataset. For the sake of brevity, we will refer to "GeoT trained on $\boldsymbol{\epsilon}_{LUMO}[\mathbf{H}]$ in the qm9 dataset" as $\mathbf{GeoT}_{LUMO}[\mathbf{GeoT}_H]$.

Comparative Analysis on the Pattern of Attention Weights by Two Target Objectives: Molecular Orbital and Enthalpy. As a preliminary step, we introduce a brief overview of the chemical theory to provide a context for our analysis. According

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

Table 1. Comparison between Attention Distributions of GeoT$_{LUMO}$ and GeoT$_{H}$ in the qm9 Dataset[a,b]

Molecule	LUMO from DFT simulation (reference)	the attention distribution produced by GeoT$_{LUMO}$		the attention distribution produced by GeoT$_{H}$	
(a)					
(b)					
(c)					
(d)					

[a]In each molecule, two individual query atoms are selected (marked by blue asterisk). The attention weights are described by color gradients: more red/yellow shades represent stronger/weaker attention weights. [b]The represented molecules are (a) naphthalene, (b) tetralin, (c) 1,3-dimethyl-2-(1,3-butadienyl)benzene, and (d) decane.

to thermodynamic theory, H is a sum of various energies of intramolecular interactions and chemical bondings, which is mainly determined by σ-bondings. On the other hand, in terms of molecular orbital (MO) theory, ϵ_{LUMO} strongly depends on the specific factor: the conjugated π-bondings included in conjugated molecules. In general, π-bondings are weaker than σ-bondings. The resonance energy from π-bonding also contributes to the overall H, but its energy scale is less significant than that of σ-bonding.[29]

The shapes of electron density distributions associated with σ- and π-bondings are also different. The electron density of the σ-bonding always localizes between the two atoms involved in the bonding. On the other hand, consecutive π-orbitals are conjugated to construct the LUMO of a molecule. Accordingly, the electron density delocalizes (spread) over the molecular scaffold. More details are provided in Appendix S1.

Motivated by the contrasting characteristics between the two chemical concepts, we visualized the attention maps produced by GeoT$_{LUMO}$ and GeoT$_{H}$ in Table 1. Then, we compared the observed patterns in terms of the theoretical expectations of two targets (delocalization for ϵ_{LUMO} and localization for H). To this end, we selected four molecules of different conjugation patterns: naphthalene, 1,2,3,4-tetrahydronaphthalene (tetralin), 1,3-dimethyl-2-(1,3-butadienyl)benzene, and n-decane.

Completely Conjugated Ring. Naphthalene is a fully conjugated molecule composed of two fused hexagonal rings, represented in the first row (a). In this case, we observed that the attention maps produced by GeoT$_{LUMO}$ spread out all over the molecular scaffold alongside the conjugated double bonds. In sharp contrast, the attention maps produced by GeoT$_{H}$ are more localized around the C−H bond in which the query atom is involved. These contrasting trends are consistently found across various types of molecules and query atoms. It indicates that the distribution of GeoAttention weight properly reflects the chemistry theory of MO, rather than depending on the individual query atoms.

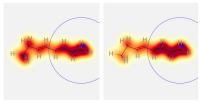
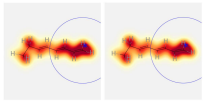
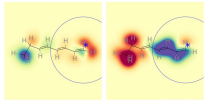
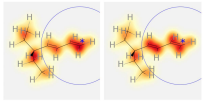
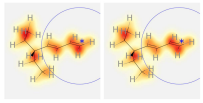
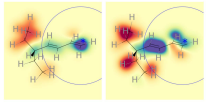
Partially Conjugated Ring. Tetralin has the same scaffold with naphthalene, but one of the hexagonal rings is not involved in the conjugation, represented in the second row (b). The attention maps of tetralin highlight that the conjugated scaffold is critical for the spread of attention weights, which is analogous to the delocalization phenomenon in π-bondings. When the query atom is selected from the conjugated part of the molecule, the corresponding attention weights produced by GeoT$_{LUMO}$ were mainly distributed only inside the conjugated part, which is similar to delocalization of the molecular orbital. In contrast, when the query was selected from the nonconjugated part, the corresponding attention weights are localized around the query atom. Similarly, the attention maps from GeoT$_{H}$ could not spread out from where the query atoms were selected. We emphasize the result because it is strong evidence that GeoT can differentiate aromatic and nonaromatic rings, which have almost similar shape to each other.

Conjugated Compound. 1,3-Dimethyl-2-(1,3-butadienyl)-benzene has an aromatic ring attached with conjugated butadiene and nonconjugated dimethyl, represented in the third row (c). Similar to the tetralin case, the corresponding attention weights spread out in the conjugated region following the LUMO of a molecule if a conjugated atom is selected as the query. However, if a nonconjugated atom is selected, the attention weights are localized and do not attend the LUMO. It is another example that the GeoT can understand the behavior of the LUMO and distinguish conjugated atoms from a molecule.

Nonconjugated Linear Compound. For the last, we picked a decane as a contrasting example in the last row (d). Decane is not conjugated, and accordingly, there is no π-bonding. In this case, the distribution patterns from GeoT$_{LUMO}$ and GeoT$_{H}$ are

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

Table 2. Comparative Analysis on the Effect of AttnScale of GeoAttention[a,b]

Molecule	the attention distribution produced by GeoT$_H$	the attention distribution produced by GeoT$_H$-base	The difference of GeoT$_H$ − GeoT$_H$-base, (+:blue, -:red)
(a)			
(b)			

[a]Overall, the AttnScale makes the attention weights more concentrated around the query. The scale of the AttnScale effect is larger in the 8th GeoAttention, rather than in the 3rd one. [b]The represented molecules are (a) 1,3,5-octatriene and (b) 7,7-dimethyl-1,3,5-octatriene.

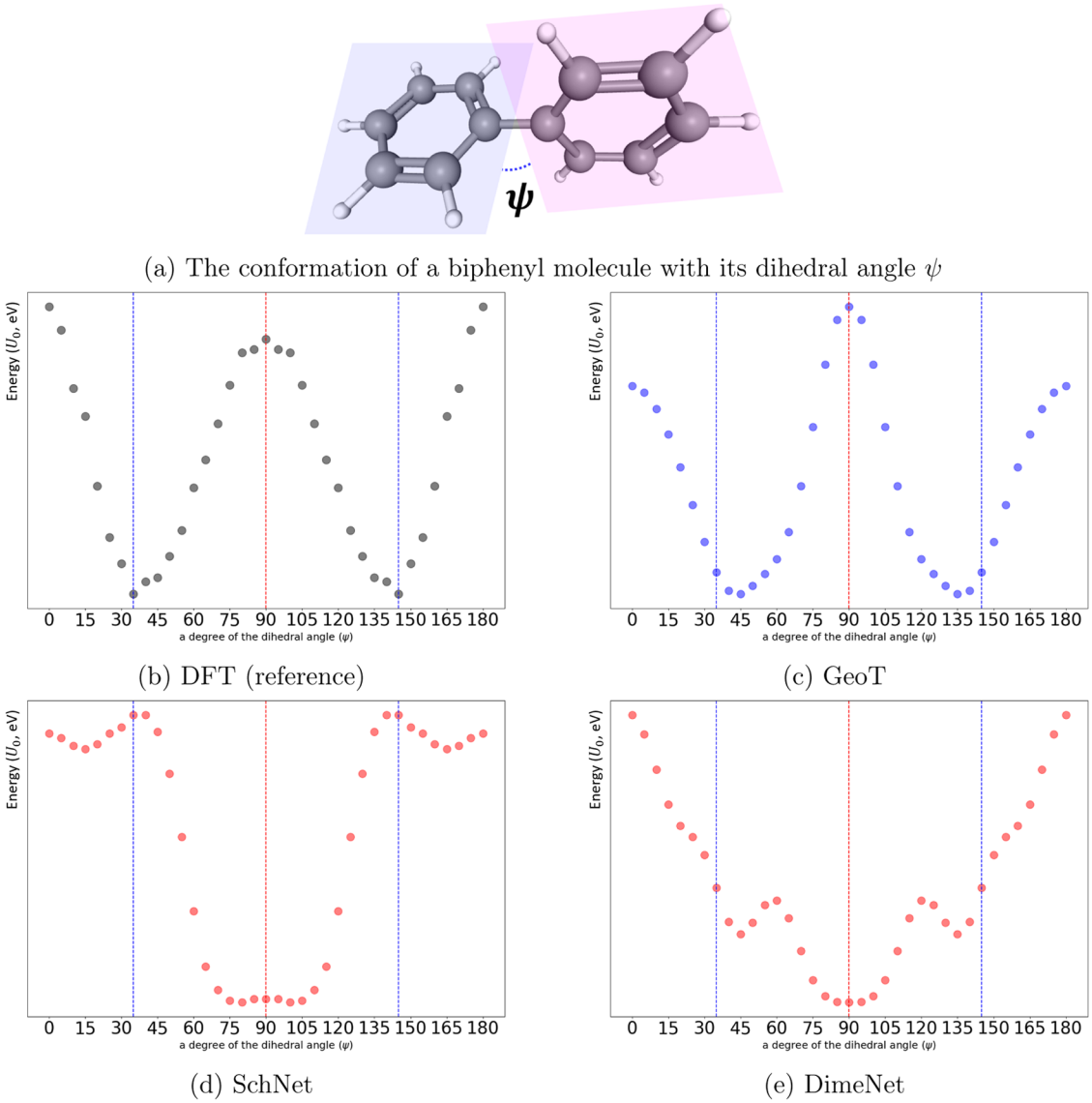


Figure 3. Plot of internal energy (U_0, eV) of a biphenyl molecule as a function of the dihedral angle ($0° \le \psi \le 180°$, $\Delta\psi = 5.0°$) obtained by different methods. (a) Illustration of a biphenyl molecule, (b) computed by DFT simulation, (c−e), predicted by GeoT, SchNet, and DimeNet. The red and blue dashed lines presented on the each graph highlight the ψ-coordinate where DF-computed global maximum ($\psi = 90°$) and minimum (about $\psi = 35$ and 145°) are located, as the ground truth.

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

Table 3. Mean Absolute Errors (MAE) of the Force Prediction of MD17[a]

molecule dependence on angles	sGDML Yes	DimeNet Yes	GemNet-T Yes	GeoT	GeoT+A	GeoT+B	GeoT+C	GeoT+B +C	GeoT+A+B +C
aspirin	0.6803	0.4981	0.2191	0.7081	0.7147	0.7657	0.606	0.741	0.721
benzene[9]		0.1868	0.1453		0.10879	0.109	0.1132	0.1206	0.0978
ethanol	0.3298	0.2306	0.0853	0.09494	0.093508	0.08967	0.0952	0.07741	0.08599
malonaldehyde	0.4105	0.3828	0.1545	0.13985		0.1408	0.14215	0.1476	0.151
naphthalene	0.1107	0.2145	0.0553	0.2198	0.2264	0.1947	0.2321	0.1949	0.1924
salicylic acid	0.2790	0.3736	0.1268	0.2869	0.289	0.2849	0.2758	0.2755	0.2667
toluene	0.1407	0.2168	0.0600	0.1541	0.1538	0.1307	0.1642	0.1403	0.147
uracil	0.2398	0.3021	0.0969	0.1471	0.1499	0.1341	0.1631	0.13097	0.1285

[a]"A", "B", and "C" mean the use of RBF_k, the parallelization of GeoT encoder, and AttnScale, respectively. The measure is kcal/mol/A.

similar to each other. It is also clear that the GeoT can understand the close relation between the conjugation and LUMO.

These results show that the GeoT has the ability to consider "the significance of atomic pairs" at the appropriate region of a molecular scaffold in predicting the target molecular properties, as if GeoT understood the related chemical theory.

Effect of AttnScale on GeoAttention. Next, we analyzed the effect of the modifications on the distribution of the attention scores. Specifically, we assumed that the introduction of AttnScale can indeed improve the interpretability of GeoT by emphasizing more significantly related atoms to the target objective. We picked H as the target objective because this distinguishing effect of AttnScale can be meaningful to interpret the attention weights of $GeoT_H$, which showed more localized attention weights in Table 1. We also compared the strengths of attention weights from the third and eighth GeoAttention blocks. Note that we did not analyze the effect of RBF_{emb} on the attention map because the training loss of GeoT was not converged without RBF_{emb} regardless of the target type.

Table 2 shows two different conjugated molecule cases that originated from $GeoT_H$. We extracted the attention map from the third and eighth GeoAttention blocks in both GeoT with AttnScale ($\mathbf{GeoT_H}$) and one without AttnScale ($\mathbf{GeoT_H}$**-base**).

In both cases, the attention distribution is more concentrated on the query atom with $GeoT_H$. On the contrary, the distribution is spread over the whole molecule in the case of the $GeoT_H$-base. Considering two aspects that (1) the bonding energy is the most significant factor of determining H of a molecule and (2) most bonds exist between neighboring atoms, we conclude that the AttnScale can increase the attention scales between more closely located query and key atoms. These results also accord with the motivation of AttnScale,[30] which is the boosting of the feature with HF signals because it can help recognize the localized features from a query.

In addition, we found that the strength of AttnScale effects depends on the location of the GeoAttention block in GeoT. Specifically, AttnScale effects are more salient in the eighth GeoAttention block rather than in the third block, in both of molecules. This result also supports the conclusion of previous study[30] that AttnScale can prevent transformer-based model from performance drop (oversmoothing problem), especially with deep architecture. More cases are represented in Appendix S1−S3.

Case Study: The Prediction of the Energy Profile as a Function of the Dihedral Angle of a Molecule. Conformational changes of a molecule structure alter geometric attributes, including dihedral angles, while preserving the atomic connectivity graph. Because GeoAttention can explicitly learn long-range interatomic relationships, we expected that GeoT shows superior performance for predicting energy profiles of conformational transitions.

We chose a biphenyl, a butane, a 4-amino-4′-cyanobiphenyl, a 4-(*N*,*N*-dimethylamino)-4′-cyanobiphenyl, and a 2-phenylpyridine as test cases. First, a biphenyl is composed of two benzene rings connected by a single bond, and its total energy is fairly sensitive to the change in the dihedral angle (ψ) between two rings (in Figure 3). This molecule has been actively studied because its rotational characteristic of a dihedral angle significantly affects the physical and chemical properties of a molecule.[31−33] In theory, when the two rings are coplanar to each other (ψ = 0 or 180°), the repulsion between hydrogen atoms accounts for the energy increase of a biphenyl. When the two rings become perpendicular to each other (ψ approaches toward 90°), conjugation over the two aromatic systems would be broken and the total energy will be (locally) maximized. The output from DFT simulation clearly supports the above-mentioned theory as the reference (Figure 3b), and we compared the predictive performance for this energy profile by GeoT, SchNet, and DimeNet (Figure 3c−3e).

Obviously, GeoT reproduced the overall shape of the energy profile with better accuracy than other methods do. Especially, GeoT successfully predicted the angle (ψ = 90°) of which the energy is maximized. GeoT also predicts the minimum points (ψ = 35 and 145°) with better accuracy than the other methods. Clearly, this task is a representative case that all carbon atoms of the molecule are involved, and thus, it can be achieved only by considering long-range interatomic relationships.

Second, we evaluated GeoT on butane (Figure S4). We also evaluated SchNet and DimeNet as comparison targets, the same with the case of biphenyl. In the case of butane, GeoT and other two models successfully predicted the energy profile as a function of dihedral angle. Additionally, all the models correctly pointed out both of the maximum (0°) and minimum (180°) degree.

We also evaluated our GeoT on biphenyl derivatives: 4-amino-4′-cyanobiphenyl and 4-(*N*,*N*-dimethylamino)-4′-cyanobiphenyl (Figures S5 and S6). In these cases, either the maximum or minimum prediction by GeoT was not accurate, whereas the overall shape of a function is similar to that of DFT (reference). However, in both cases, SchNet and DimeNet failed to predict the energy profiles accurately. The predictions by two models were significantly different from the reference labels; both of the models incorrectly identified the maximum point completely opposite and the overall shapes deviated from the references.

We chose 2-phenylpyridine (Figure S7) as the also selected target molecule. In terms of the minimum point (0°), all models,

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

including GeoT, provide accurate predictions. However, GeoT, as well as SchNet, failed to accurately predict the maximum point (90°). Only DimeNet was able to make the correct prediction for the maximum point.

In summary, GeoT showed good prediction performance on biphenyl and butane. However, it shows relatively less accurate predictions in biphenyl derivatives and 2-phenylpyridine.

Model Performance on Public Benchmarks for Molecular Property Prediction. Table 3 shows MAE values on the MD17 dataset, which comprises the energy prediction tasks of various conformations of each molecule. We compared the result of GeoT with three previous results sGDML,[22] DimeNet,[11] and GemNet-T[15] in the left side of the table. In the right side of the table, the effects of using RBF_k introduced in eq 7 and other two model refinement strategies parallel with MLP and AttnScale introduced in the Enhancing the HF Term in GeoAttention section are shown. All strategies contribute to improving the model performance in six out of eight types of molecules, except for aspirin and malonaldehyde. GeoT outperformed three molecules: benzene, ethanol, and malonaldehyde, whereas GemNet-T achieved the best performance on the remaining five molecules.

Table S1 composes the MAE on QM9 datasets of GeoT and five previous studies for comparison. SchNet,[7] Cormorant,[9] PhysNet,[8] and DimeNet++[14] are message passing-based methods, and GRAT[10] is the only transformer-based model. Note that Cormorant[9] and DimeNet++[14] use angles as well as distances between atoms, whereas other three models[7,8,10] only use distances between atoms for prediction tasks, as GeoT does. Our model outperformed these three previous models that use distance information only. However, the performance of our model does not exceed that of DimeNet++.[14] GeoT achieved performances comparable to those of DimeNet++ on six targets and outperformed in one case.

Table S2 presents the performance results in the OC20 IS2RE (10k) dataset, compared with those of five previous models[7,11,14,34] reported.[26] Six types of ablation studies were conducted to validate three types of model refinement strategies, the same as mentioned above. We should mention that SchNet and DimeNet used periodic boundary conditions (PBC) to represent repeated structures of surfaces. The three types of model refinements were not effective for the 10k task. Table S3 presents the performance results of various versions of GeoT and competing OC20 IS2RE tasks with a size of 10,000 and full datasets.

Ablation Study. We performed ablation studies on simple linear and Gaussian basis functions which are used in SchNet,[7] as well as radial bases. Bessel basis functions, which are used in DimeNet,[11,14] correspond to the expressions given in Table S4. The results in Table S5 show that the Gaussian basis functions produced the best performance with GeoT.

DISCUSSION

We developed GeoT, a novel transformer-based model for molecular property prediction based on the molecular conformation. Each GeoT encoder block has a GeoAttention module, which learns the relationship between all pairs of atoms with their type information on a molecule graph. GeoAttention is the modified self-attention block that incorporates atom−atom distance information and enhances the high-frequency signal from heterogeneous key atoms. To the best of our knowledge, it is the first interpretative analysis of the attention pattern derived in molecular graphs with geometric features from the chemical perspective.

GeoT has three advantages over previous studies. First, GeoT can visualize the contributions of all atom−atom relationships for determining properties of a given molecule. Surprisingly, the attention map obtained from trained GeoAttention shows clearly distinctive patterns depending on the type of targets. The attention pattern from $GeoT_{LUMO}$ follows the resonance structure of π-bonding regardless of molecular shape, whereas those from $GeoT_H$ are highly correlated with the σ-bonding.

From those observations, we conclude that attention can differentiate the scale of various types of energies and more attend to features with larger contributions to the target objective. Extensive study on various shaped molecules with other target types is needed in the future study.

Second, we conducted a case study on five molecules to evaluate the performance of GeoT in predicting the energy distribution across conformational changes, including the identification of the maximum and minimum energy geometries. GeoT predicted the maximum and minimum accurately in two cases: biphenyl and butane. When a biphenyl molecule forms a planar structure ($\psi = 0°$), it becomes unstable because the steric hindrance effect between the 2 and 2′-carbon grows larger. When a biphenyl molecule forms a perpendicular conformation ($\psi = 90°$), it also becomes unstable because the π-conjugation between two benzene rings weakens. Due to these two opposing effects, $\psi = 35°$ represents the global energy minimum in the energy profile of a biphenyl molecule. The same situation also applies in the case of biphenyl derivatives: 4-amino-4′-cyanobiphenyl and 4-(N,N-dimethylamino)-4′-cyanobiphenyl.

In the case of butane, the energy is determined by the rotation of the two middle carbons and the resulting relative positions of the two terminal methyl groups (syn, gauche, and anti conformation).[35] When two methyl groups are in the syn conformation ($\psi = 0°$), steric hindrance occurs, leading to a molecule becoming unstable. GeoT accurately captured this high-energy state in this case. One of the recent studies[36] reported that the relative energy of other conformations (gauche and anti conformation) is primarily determined by σ-hyperconjugation. It means that the conformation energy of alkanes consisting only σ-bonds without any π-bonds is significantly affected by these effects. GeoT showed good predictions on overall points in spite of some underestimation of the energy discrepancy between the gauche ($\psi = 60°$) and the anticlinal ($\psi = 120°$).

In the case of 2-phenylpyridine, the situation becomes more complex. When $\psi = 0°$, there exists an additional stabilizing effect caused by the interaction between C−H and the nearby nitrogen atom, in addition to all the factors that determine conformation energy of a biphenyl molecule. As a result, it can be a more challenging case than biphenyl, in conformation analysis for neural networks. GeoT successfully predicted the minimum point ($\psi = 0°$), and the overall predicted shape is similar. However, the prediction of the maximum point ($\psi = 90°$) deviated from the reference. This result may be caused by the complicated effects from multiple factors, which is a more challenging task for GeoT.

From these above-mentioned observations, we argue that GeoT can be generalized to predict the molecular energy in the suboptimal state as well as those in the optimal state in the case of the molecules with relatively simple effects. Notably, GeoT also achieved remarkable performance on three public benchmarks: MD17, QM9, and OC20. In particular, GeoT showed

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

comparable performances with other MPNN-based models, which require additional computations for angle values. Indeed, GeoT outperformed the previous MPNN models and other transformer-based models without the use of angle values.

Third, GeoT is more computationally efficient than the MPNNs that use angle values between three atoms. Unlike these models, our approach does not require a cutoff value and can learn long-range features without any restriction, providing an advantage in terms of computational efficiency over previous MPNN-based studies.

One limitation of GeoT is that the prediction performances on the public benchmarks are not consistently superior to those of the previous models, although GeoT has a lower computational cost than other models requiring angle computation and extra embedding spaces. Another point is that GeoT is capable of providing interpretability from a relative perspective with respect to an arbitrarily selected query atom but not from a global perspective. This is due to the fact that GeoT is based on the transformer architecture, which is designed to consider the relative relationships between components of data.

To overcome these limitations of GeoT, we will seek to improve prediction performance on molecule graphs and advanced representation strategies of attention weights for gaining deeper chemical insights.

CONCLUSIONS

We developed GeoT, a geometry-aware transformer-based model for molecular graphs, which is the first interpretative study of the attention pattern with geometric features from a chemical perspective. GeoT validated the applicability and effectiveness of its self-attention mechanism in predicting properties based on the molecular structure. While GeoT has some limitations in the prediction performance, it can provide interpretative visualization of the molecular property. In the future, we plan to incorporate a wider range of molecular datasets including real experimental values to improve the performance and applicability of GeoT.

METHODS

In this section, we provide the preliminaries and descriptions of the GeoT architecture and the experimental details and training strategies.

Preliminary: Transformer Encoder. We provide the basics of the original self-attention mechanism and transformer encoder proposed elsewhere.[16] We do not describe decoders because our methods are based only on the encoder part of the original transformer.

Self-Attention. The self-attention mechanism allows the transformer to focus on different elements of the input depending on the task. A set of input vectors $\{\boldsymbol{X}_i^l\}_{i\in\{1,\cdots,N\}}$ is split into three types of tensors: query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. In practice, each matrix is obtained by the multiplication of individual weights W_Q, W_K, and W_V on the same input X.

$$\boldsymbol{Q} = W_Q\boldsymbol{X},\ \boldsymbol{K} = W_K\boldsymbol{X},\ \boldsymbol{V} = W_V\boldsymbol{X} \tag{2}$$

After that, the dot product is executed between $\boldsymbol{Q}$ and $\boldsymbol{K}$ and scaled by the vector dimension d_m.

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_m})\cdot\boldsymbol{V} \tag{3}$$

Multihead Self-Attention. It is common to implement multiple individual self-attention blocks in the transformer-based architecture because it can be advantageous for tasks depending on the complex interactions between different parts of the data, such as natural language processing. Each matrix $\{\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\}$ is split into h matrices $\{\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\}_{i\in\{1,\cdots,h\}}$ with dimension d_m/h. Self-attention is then applied to each $\{\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\}_{i\in\{1,\cdots,h\}}$, followed by concatenation.

$$\boldsymbol{H}_i = \text{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) i \in \{1, ..., h\}$$
$$\text{MSA}(\boldsymbol{X}) = \text{Concat}(\boldsymbol{H}_1, ..., \boldsymbol{H}_h) \tag{4}$$

Transformer Encoder Layer. The transformer encoder layer is constructed by the layer normalization (LayerNorm)[37] and a feedforward (FF) layer with skip connections. First, the output from $\text{MSA}(\boldsymbol{X}^l)$ was added to the original input $\mathbf{X}^l$ by the skip connection with a LayerNorm. After that, the last output $\tilde{\mathbf{X}}^l$ is fed into the FF layer, consisting of two linear transformations with an ELU[38] activation. Finally, another LayerNorm is applied to the output from the FF layer added by $\tilde{\mathbf{X}}^l$. This can be formulated as

$$\tilde{\mathbf{X}}^l = \text{LayerNorm}(\text{MSA}(\boldsymbol{X}^l) + \boldsymbol{X}^l)$$
$$\boldsymbol{X}^{l+1} = \text{LayerNorm}(\text{FF}(\tilde{\mathbf{X}}^l) + \tilde{\mathbf{X}}^l) \tag{5}$$

Details of GeoT and GeoAttention. We denote a molecule as a set of N atoms $\{a_i\}_{i\in\{1,\cdots,N\}}$ with their coordinates $\{\mathbf{r}_i\} \in \mathbb{R}^3$. We calculate the Euclidean distance $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$ between the two atoms a_i and a_j. Our model is a stack of L layers, which transforms input $\boldsymbol{X}^l$ into output $\boldsymbol{X}^{l+1}$, especially when the first layer input $\boldsymbol{X}^0$ is given from the atom type-embedding layer.

Combining RBF into GeoAttention. A single RBF $\mathbf{g}(r_{ij})$ used in GeoT is a Gaussian function, which is defined by $\mathbf{g}(r_{ij})$: $\mathbb{R} \to \mathbb{R}^{d_m}$, where an input r_{ij} is the Euclidean distance between two atoms a_i and a_j and d_m is the dimension of distance embedding. RBF is analogous to an embedding layer for a continuous-valued input set.

The k multiple RBFs are $\mathbf{g}_k(r)$ of the different centers of the Gaussian distribution δk. We selected multiple different Gaussian basis functions for constructing $\mathbf{g}_k(r)$, as proposed in SchNet,[7] where $\gamma = 10$ and $\delta = 0.1$ Å are predefined constants.

$$\mathbf{g}_k(r_{ij}) = \exp(-\gamma(\|r_{ij} - \delta k\|)^2) \tag{6}$$

With $\mathbf{g}_k(r_{ij})$, we created the RBF embedding matrix $\boldsymbol{\Lambda}$, depending distance r_{ij} with two feedforward layers f_θ as follows. We also embed atom types a_i and a_j into f_θ because the atomic interaction depends on both the distance between two atoms and their atom types. To achieve this, we combined atom type-embedding vector $Z(a_i)$ and $Z(a_j)$ of two atom i and j to $\mathbf{g}_k(r_{ij})$. The $\mathbf{g}_k(r_{ij})$ is defined at the start of the network, and the output $\boldsymbol{\Lambda}$ is individually provided in each GeoAttention block (detailed in Figure 1). The matrix form of constructing $\boldsymbol{\Lambda}$ is defined below, where $\mathbf{r}_{ij}$ and $\mathbf{a}$ are the matrix representations of given atom indices $\{r_{i,j}\}_{i,j=1,\cdots,N}$ and $a_{i=1,\cdots,N}$, respectively.

$$\boldsymbol{\Lambda} = f_\theta(\mathbf{g}_k(\mathbf{r}_{ij}) \oplus (Z(\mathbf{a}_i) + Z(\mathbf{a}_j))) \tag{7}$$

Accordingly, the base form of GeoAttention is defined as below.

$$\text{Attn}_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda}) = (\boldsymbol{Q}\boldsymbol{K}^T\boldsymbol{\Lambda})/\sqrt{d_m}$$
$$\text{GeoAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{\Lambda}) = \text{Attn}_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})\cdot\boldsymbol{V} \tag{8}$$

Enhancing the HF Term in GeoAttention. AttnScale[30] is a type of modification for self-attention, which was originally

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

developed to amplify the HF signal of self-attention weights in the deep transformer architecture. The original paper[30] considers the direct current (DC) component of self-attention $\boldsymbol{A}$ as $\boldsymbol{A}_{DC} = \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T$ and the other components are defined as HF terms $\boldsymbol{A}_{HF}$. As described above, AttnScale rescales the HF term by $(1 + w_a)$, preventing from the oversmoothing effect, which preserves only indistinguishable DC components in feature maps especially in deep the transformer architecture.

Following from the original paper, we enhanced $\boldsymbol{A}_{HF}$ of GeoAttention, which can naturally correspond to interatomic relations between closely located atom pairs in our tasks. The enhanced HF signal is $\boldsymbol{A}'_{HF} = (1 + w_a)\boldsymbol{A}_{HF}$, where $w_a \geq 0$. Consequently, the detailed definition of Attn'_G is defined as below.

$$\begin{aligned}\mathrm{Attn}'_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda}) &= \boldsymbol{A}_{DC} + \boldsymbol{A}'_{HF} \\ &= \boldsymbol{A}_{DC} + (1 + w_a)\boldsymbol{A}_{HF} \\ &= \frac{1}{N}\sum_k \boldsymbol{A}_k + (1 + w_a)\boldsymbol{A}_{HF}\end{aligned} \tag{9}$$

Construction of GeoAttention as Multiheads. GeoAttention is defined by the multiplication of $\boldsymbol{V}$ by the above-mentioned term $\mathrm{Attn}'_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})$. We implemented GeoAttention as H heads to achieve multihead GeoAttention (denoted as $\{\mathrm{GeoAttention}\}_H$) by splitting each $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ into H components, applying individual GeoAttention to each component set and concatenating them, where $h \in \{1, \cdots, H\}$ means a head index.

$$\begin{aligned}\boldsymbol{G}_h &= \mathrm{GeoAttention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h, \boldsymbol{\Lambda}) \\ \{\mathrm{GeoAttention}\}_H &= \mathrm{Concat}(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_H)\end{aligned} \tag{10}$$

For the sake of brevity, we omitted the H term: all GeoAttention described in this paper means the multihead term $\{\mathrm{GeoAttention}\}_H$.

Parallelization in the GeoT Encoder. For ensuring stable convergence of GeoT during training, we further modified the structure of the GeoT encoder. Inspired by the previous work analyzing self-attention,[39] we implemented only one layer normalization, which takes (multihead) GeoAttention, FF layers, and skip connection simultaneously. A block of GeoT encoder is formulated as below, where $l \in \{1, \cdots, L\}$ means an encoder index.

$$\begin{aligned}\boldsymbol{X}^{l+1} = \mathrm{LayerNorm}(&\mathrm{GeoAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{\Lambda}) + \mathrm{FF}(\boldsymbol{X}^l) \\ &+ \boldsymbol{X}^l)\end{aligned} \tag{11}$$

Readout Layer of GeoT for Prediction Output. We introduced a sum-pooling layer at the final step to obtain scalar-valued molecular property by aggregating atom features. This is a common strategy in MPNNs,[7,11,15] which is named as the "readout" layer.

$$\hat{y} = W_L\boldsymbol{X}^L + b \tag{12}$$

Dataset. We describe three benchmarks: MD17,[21−23] QM9,[24,25] and OC20.[26] The original MD17 dataset[21] comprises 10 small molecules. For each molecule, the dataset contains energy and forces for different geometries with more than 10,000 structures, based on DFT calculations. Instead of using the original dataset, we used a dataset comprising 1000 training samples[23] with more precise calculations than the original MD17 dataset. The task of the datasets is to predict the forces and energy of a given conformation from other conformations of specific molecules.

QM9 comprises 134,000 small organic molecules consisting of carbon, hydrogen, nitrogen, oxygen, and fluorine in an equilibrium state. All molecular conformations and their corresponding properties were created through computational simulations based on DFT calculation. The datasets contain stable three-dimensional coordinates of each molecule and their 12 scalar quantum chemical properties, including geometric, energetic, electronic, and thermodynamic properties, of the given molecular structure. The task of the datasets is to predict the properties from the molecular conformation. As the datasets contain comprehensive chemicals with high consistence,[25] many molecular property prediction tasks were evaluated on the datasets.

Open Catalyst 2020 (OC20)[26] is an open-source for learning catalysis dynamical properties with chemical configurations. Similar to the benchmarks mentioned above, OC20 data comprise three-dimensional configurations with atomic numbers of each adsorbate onto surfaces in the initial and DFT relaxation states. Nevertheless, OC20 is a more advanced task, as the relaxed state between adsorbates and surfaces must predict the initial state thereof. OC20 includes three different types of tasks: from the structure to predict energy and forces (S2EF), from the initial structure to predict the relaxed structure (IS2RS), and from the initial structure to predict the relaxed energy (IS2RE). Each task is subdivided into several tasks according to the dataset size. We focused on the IS2RE tasks on 10k and full-sized datasets, which require relatively less computations.

Visualization details. We extracted $\|\mathrm{Attn}'_G(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{\Lambda})\|_2$ of several molecules from two types of GeoT trained on QM9: $\mathrm{GeoT}_{\mathrm{LUMO}}$ and GeoT_H. Subsequently, an atom i was chosen for each molecule as a query, and its GeoAttention values were realized using the other atoms $\{j\}\backslash i$. Query atoms are marked with blue asterisks, and attention norms of key atom j from query atom i are colored with red shades. The shades are gradually colored by interpolations for visibility and easy comparison with MO. The blue circle around the query atom has a radius of 5 Å.

Prediction of the Energy Profile as a Function of the Dihedral Angle of a Molecule. GAMESS, which is the open source for computational chemistry analysis,[40,41] was used to produce DFT simulation of the internal energy of biphenyl, butane, 2-phenylpyridine, and other two biphenyl derivatives. The 6-31G* was used as the basis set, and the DFT functional was B3LYP. We used the pretrained SchNet and DimeNet provided by Pytorch geometry[42] version 2.2, as the comparative study.

Training. In this study, the depth of GeoT was set to $L = 8$, 16 to perform QM9 prediction, whereas it was four ($L = 4$) each in the case of the MD17 and OC20 datasets. We used 300 ($n_{\mathrm{basis}} = 300$) different Gaussian functions with $\mu_k = 0.1 \cdot k$ to embed atom−atom distances. The dimension of GeoT was defined as (256, 512, 1024) with (4, 16, 64) multihead GeoAttention ($H =$ 4, 16, 64). The Swish[43] and ELU[38] activation functions were used for the RBF and feedforward layers, respectively.

For QM9 training, we removed 3k molecules, which were previously reported to have an unstable conformation, following from the previous studies.[7−10,14] The mean absolute error (MAE) was used to perform evaluations according to the guidelines, and each label was trained individually. Adam[44] was used as the optimizer with MAE loss, and the batch size was 32 in

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769

all experiments. The learning rate was set to 0.0002 at the initial step and decreased by 0.95 for every 200k steps with a linear warmup of 3000 steps. We applied the early stopping method by evaluating every 10k step of training. The maximum number of training epochs was up to 300.

To calculate forces, we assumed the predicted energy as a function of positions $E(\{\mathbf{r}_i\})$. Based on the relation between the force and potential energy of atoms, we calculated forces by differentiating energies with atom positions following from the previous studies.[7,22]

$$F_j = \frac{\partial}{\partial \mathbf{r}_j} E(\{\mathbf{r}_i\}) \tag{13}$$

In the MD17 dataset, both the energy and forces of molecular conformations are provided. To utilize both features with eq 13, we trained our model with the modified loss function with additional force terms as given by eq 14.[7,22]

$$L(E, \hat{E}) = |E - \hat{E}| + c \times \sum_{j=i}^{n} |F_j - \frac{\partial}{\partial \mathbf{r}_j} \hat{E}(\{\mathbf{r}_i\})| \tag{14}$$

where c is the weight coefficient for the loss on forces. We set c to 1000 in our experiments.

For training the OC20 dataset, the trained model was evaluated on the in-distribution subset for validation, following the official guideline.[26] Moreover, the periodic boundary condition (PBC) trick was not implemented here, whereas all of the other methods implemented it. PBC is an approximation technique for analysis in large systems with small repeated patterns. With the PBC method, multiple replicas of a small unique pattern are arranged to represent a regular system. Consequently, broad regular surface molecules can be represented with a much smaller unique cell. Thus, it significantly lowers the number of computations and ensures consistent representations of different atoms on the same surface.

■ ASSOCIATED CONTENT

Data Availability Statement

The authors utilized the public open web sources (http://quantum-machine.org/, http://www.sgdml.org/, and https://opencatalystproject.org/) to obtain three public benchmark datasets QM9, MD17, and OC20, respectively. The authors provide the generated input coordinates of biphenyl. The code used for this study is available on GitHub https://github.com/oleneyl/geometry-aware-transformer.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05753.

Biphenyl coordinates (ZIP)

Additional visualization of GeoAttention, additional results of the model performances, and theoretical background of chemistry (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jeonghee Jo − *Institute of New Media and Communications, Seoul National University, Seoul 08826, Republic of Korea;* orcid.org/0000-0003-1154-1761; Email: page1024@snu.ac.kr

Sungroh Yoon − *Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea; Department of Electrical and Computer Engineering and Artificial Intelligence Institute, Seoul National University, Seoul 08826, Republic of Korea;* orcid.org/0000-0002-2367-197X; Email: sryoon@snu.ac.kr

Authors

Bumju Kwak − *Recommendation Team, Kakao Corporation, Gyeonggi 13529, Republic of Korea;* orcid.org/0000-0002-6517-1711

Jiwon Park − *LG Chem, Seoul 07795, Republic of Korea; Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea*

Taewon Kang − *Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea*

Byunghan Lee − *Department of Electronic Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea;* orcid.org/0000-0002-6727-0975

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05753

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by Institute of Information and communications Technology Planning & Evaluation (IITP) grant (No. 2021-0-01343, Artificial Intelligence Graduate School Program in Seoul National University; and No. 2021-0-02068) and also supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A6A3A01087603, 2022R1A3B1077720, 2022M3C1A3081366). All supports were funded by the Korea government (MSIT).

■ REFERENCES

(1) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.

(2) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(3) Lorenz, S.; Groß, A.; Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.* **2004**, *395*, 210−215.

(4) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 146401.

(5) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(6) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.

(7) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

(8) Unke, O. T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.

(9) Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant molecular neural networks. *Adv. Neural Inf. Process. Syst.*. *32***2019**,.

(10) Yoo, S.; Kim, Y.-S.; Lee, K. H.; Jeong, K.; Choi, J.; Lee, H.; Choi, Y. S. Graph-Aware Transformer: Is Attention All Graphs Need?. 2020, arXiv:2006.05213. arXiv.org e-Print archive. https://arxiv.org/abs/2006.05213.
(11) Gasteiger, J.; Groß, J.; Günnemann, S. Directional message passing for molecular graphs. 2020, arXiv:2003.03123. arXiv.org e-Print archive. https://arxiv.org/abs/2003.03123.
(12) Choukroun, Y.; Wolf, L. In *Geometric Transformer for End-to-End Molecule Properties Prediction*, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2021.
(13) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., *Neural message passing for quantum chemistry*, International Conference on Machine Learning, 2017; pp 1263−1272.
(14) Gasteiger, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules. 2020, arXiv:2011.14115. arXiv.org e-Print archive. https://arxiv.org/abs/2011.14115.
(15) Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 6790−6802.
(16) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
(17) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. 2020. arXiv:2002.08264. arXiv.org e-Print archive. https://arxiv.org/abs/2002.08264.
(18) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12559−12571.
(19) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28877−28888.
(20) Chen, D.; O'Bray, L.; Borgwardt, K. In *Structure-Aware Transformer for Graph Representation Learning*, International Conference on Machine Learning, 2022; pp 3469−3489.
(21) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(22) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **2019**, *240*, 38−45.
(23) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, No. 3887.
(24) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(25) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, No. 140022.
(26) Chanussot, L.; Das, A.; Goyal, S.; et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **2021**, *11*, 6059−6072.
(27) Rao, A.; Park, J.; Woo, S.; Lee, J.-Y.; Aalami, O. In *Studying the Effects of Self-attention for Medical Image Analysis*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021; pp 3416−3425.
(28) Chefer, H.; Gur, S.; Wolf, L. In *Transformer interpretability beyond attention visualization*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; pp 782−791.
(29) Reusch, W. H. *An Introduction to Organic Chemistry*; John Wiley & Sons, 1977.
(30) Wang, P.; Zheng, W.; Chen, T.; Wang, Z. In *Anti-oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice*, International Conference on Learning Representations, 2022.
(31) Grein, F. Twist angles and rotational energy barriers of biphenyl and substituted biphenyls. *J. Phys. Chem. A* **2002**, *106*, 3823−3827.
(32) Johansson, M. P.; Olsen, J. Torsional barriers and equilibrium angle of biphenyl: reconciling theory with experiment. *J. Chem. Theory Comput.* **2008**, *4*, 1460−1471.
(33) Jain, Z. J.; Gide, P. S.; Kankate, R. S. Biphenyls and their derivatives as synthetically and pharmacologically important aromatic structural moieties. *Arabian J. Chem.* **2017**, *10*, S2051−S2066.
(34) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, No. 145301.
(35) Vollhardt, K.; Schore, N. *Organic Chemistry*; W. H. Freeman, 2010.
(36) Wu, J. I.-C.; Wang, C.; McKee, W. C.; von Ragué Schleyer, P.; Wu, W.; Mo, Y. On the large σ-hyperconjugation in alkanes and alkenes. *J. Mol. Model.* **2014**, *20*, No. 2228.
(37) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer normalization. 2016, arXiv:1607.06450. arXiv.org e-Print archive. https://arxiv.org/abs/1607.06450.
(38) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. In *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUS)*, International Conference on Learning Representations, 2016.
(39) Park, N.; Kim, S. In *How Do Vision Transformers Work?*, International Conference on Learning Representations, 2022.
(40) Perri, M.; Weber, S. Web-based job submission interface for the GAMESS computational chemistry program. *J. Chem. Educ.* **2014**, 2206−2208.
(41) Barca, G. M. J.; Bertoni, C.; Carrington, L.; Datta, D.; De Silva, N.; Deustua, J. E.; Fedorov, D. G.; Gour, J. R.; Gunina, A. O.; Guidez, E.; et al. Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **2020**, *152*, No. 154102.
(42) Fey, M.; Lenssen, J. E. In *Fast Graph Representation Learning with PyTorch Geometric*, ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019.
(43) Ramachandran, P.; Zoph, B.; Le, Q. V. Searching for activation functions. 2017, arXiv:1710.05941. arXiv.org e-Print archive. https://arxiv.org/abs/1710.05941.
(44) Kingma, D. P.; Ba, J. In *Adam: A Method for Stochastic Optimization*, International Conference on Learning Representations, 2015.

https://doi.org/10.1021/acsomega.3c05753
ACS Omega 2023, 8, 39759−39769