中国人民大学
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

GTMGC: Using Graph Transformer to Predict Molecule's Ground-State Conformation

Guikun Xu, Yongquan Jiang, Pengchuan Lei, Yan Yang, Jim X. Chen

(ICLR 2024)

Fanmeng Wang

2024-4-25

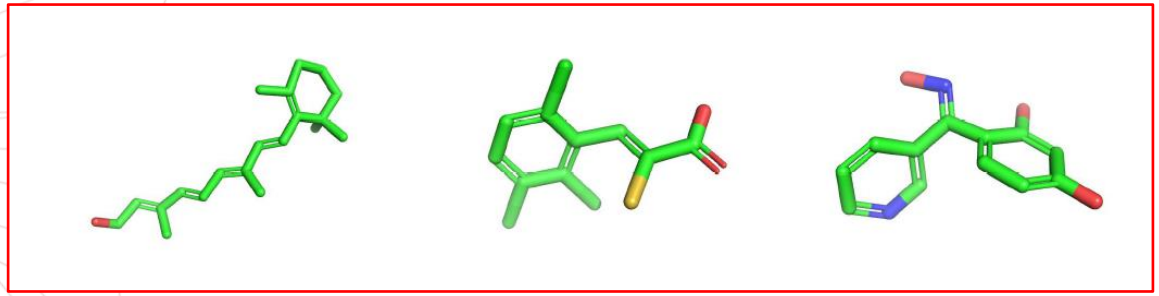
Outline

- Introduction
- Method
- Experiment
- Conclusion

- Introduction
- Method
- Experiment
- Conclusion

Introduction

- The **molecule's ground-state conformation** refers to **the lowest energy state** on its potential energy surface, which requires the least amount of energy to maintain.
  - It represents **the most stable 3D molecular structures**, and plays an important role in determining the physical, chemical, and biological properties of molecules.
  - However, **experimental or computational methods**, such as density functional theory, are time-consuming and labor-intensive for obtaining this conformation.
  - Therefore, finding an efficient way to obtain ground-state conformation has become an important research topic.



Introduction

- Current works:
  - Recently, **deep generative models** have been employed to generate **low-energy stable conformations** of molecules (i.e., **molecular conformation generation**).
  - However, these generative methods mainly focus on generating **many potential stable conformations**, rather than specifically targeting the **ground-state conformation.**
  - As a result, **additional screening steps** are required in real-world applications to find the best conformation among these generated conformations.

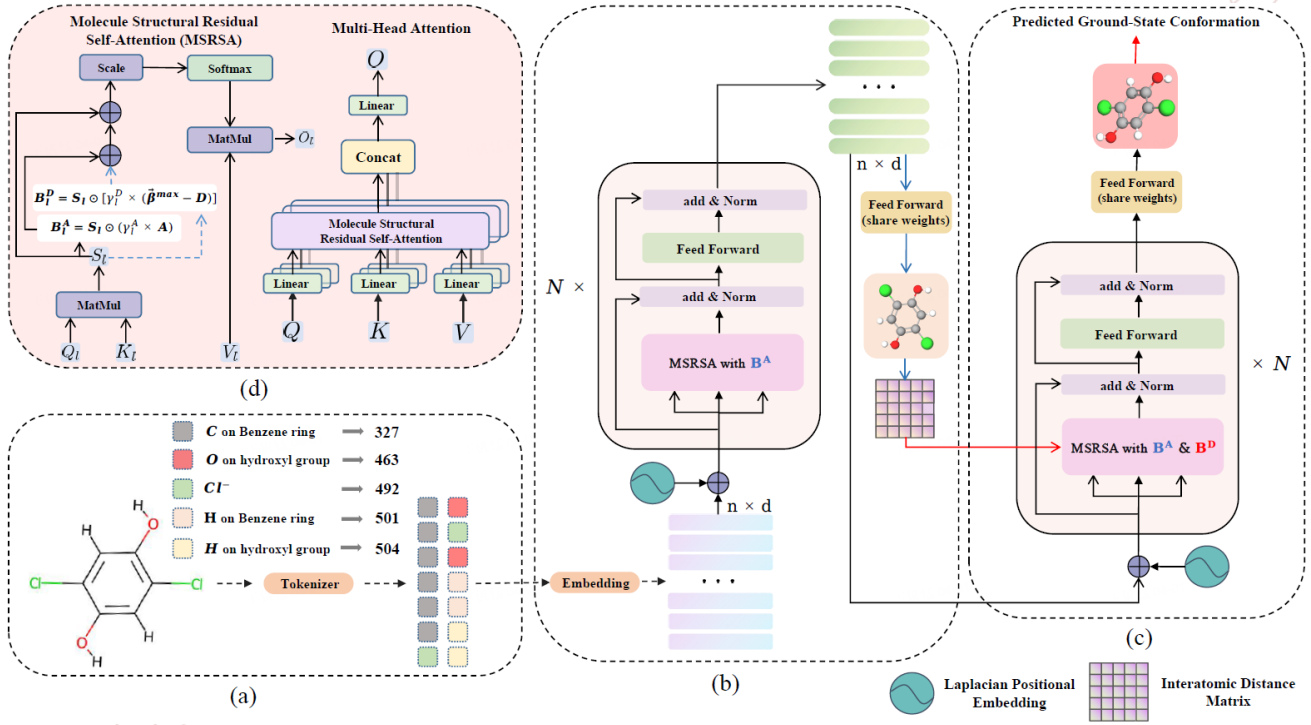
Molecular conformation generation

Molecular ground-state conformation prediction

Introduction

■ GTMGC

➢ A novel network based on Graph-Transformer that seamlessly predicts the **ground-state conformation** of molecules in 3D space from their **2D topological graph** in an end-to-end manner.

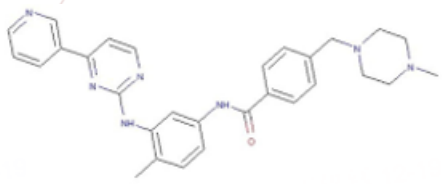


- Introduction
- **Method**
- Experiment
- Conclusion

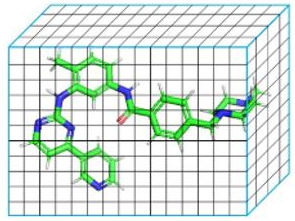
Method

■ Notations

- Each molecule is represented by a **graph** $\mathcal{G} = (V, E)$ where V is the set of vertices representing **atoms** and E is the set of edges representing inter-atomic **bonds**.
- In the context of a graph structure, its **adjacency matrix** is denoted by $A \in R^{n \times n}$, where $A_{ij} = 1$ signifies the presence of an edge connecting node v_i to v_j.
- The **ground-state conformation** of a molecule is denoted as $G \in R^{n \times 3}$, which represents a set of 3D Cartesian coordinates.
- Expanding on this, the **interatomic distance matrix** of a molecule is defined as $D \in R^{n \times n}$, where D_{ij} represents the Euclidean distance between atom v_i and v_j .



Graph

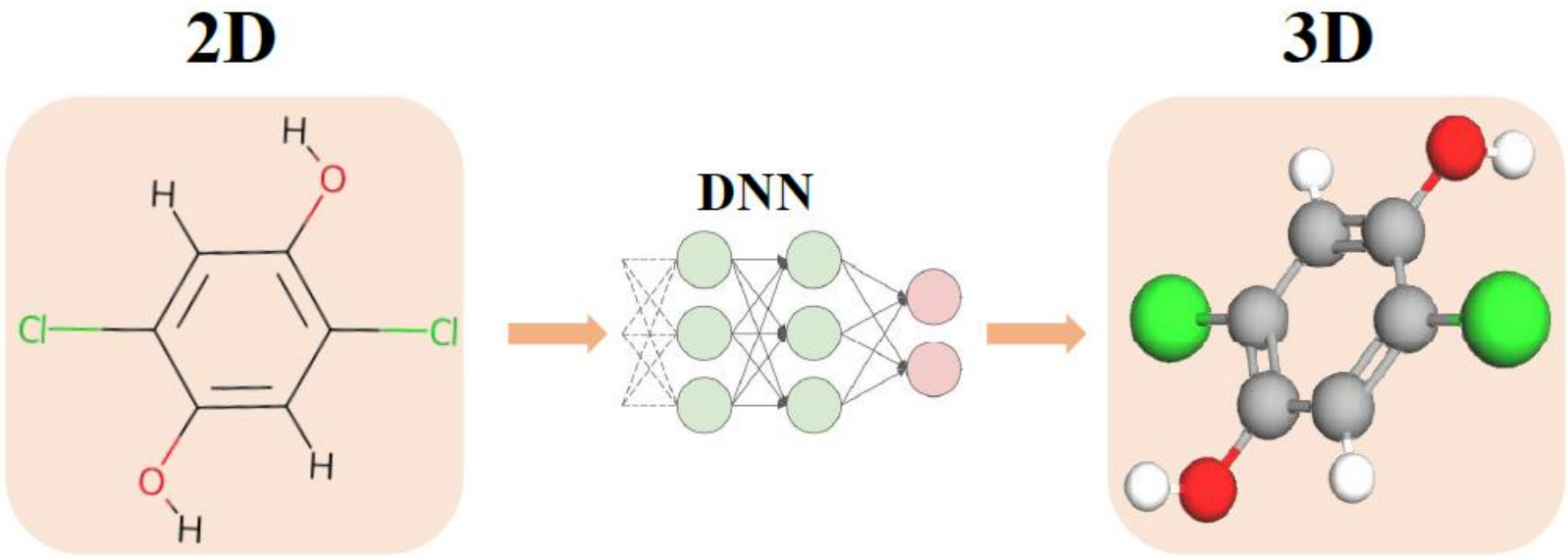


Conformation

Method

■ Problem Definition

➢ The **ground-state conformation prediction task** aims to predict the conformation of a molecule in its ground state, solely based on its 2D molecular structure.

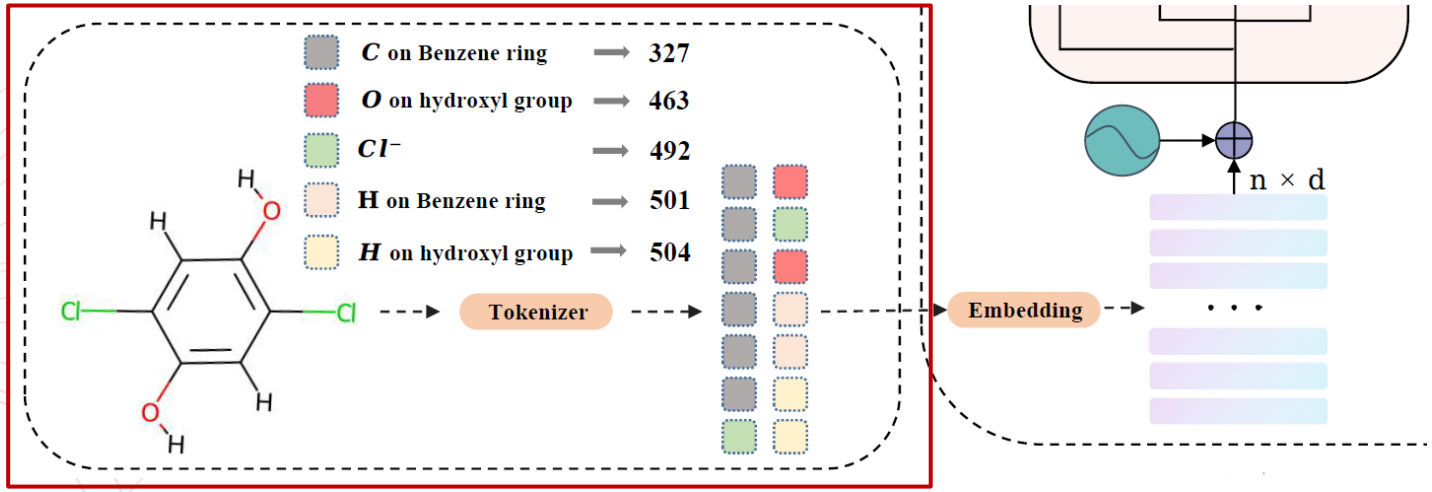


Method

■ Model Input

➢ **Input IDs**. We use the chemically meaningful IDs tokenized by the **MoleBERT Tokenizer** as input IDs for our model.

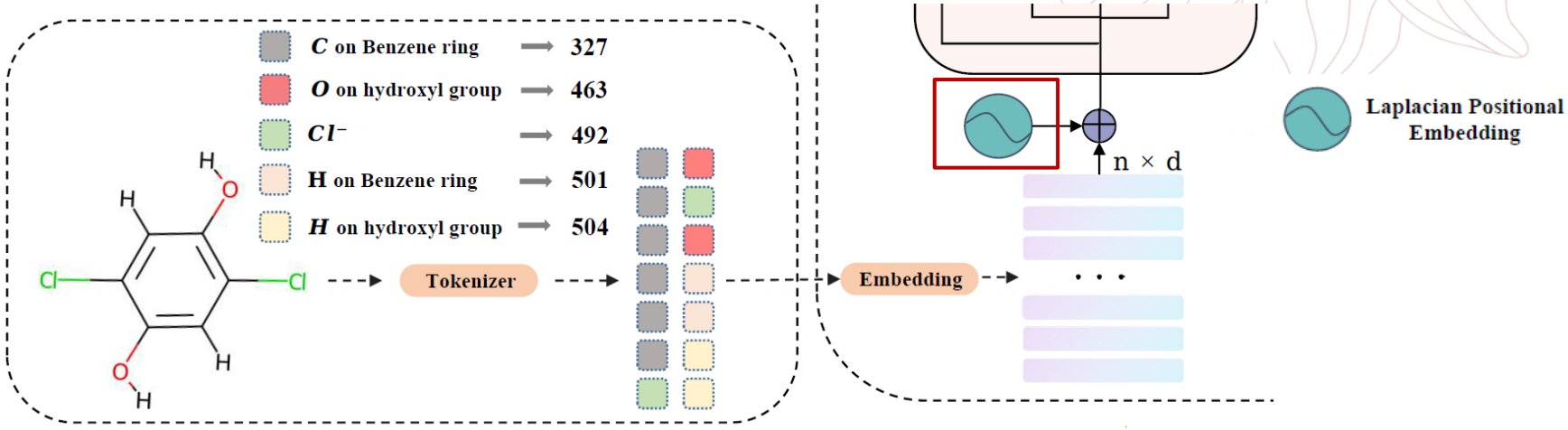
- The **same atoms** can be categorized into **distinct subclasses**, mitigating the quantitative divergence between prevalent and scarce atoms.
- Since these tokens already contain ample chemical information about atoms, we cease utilizing **edge features** to enhance the representation of molecular input features, thereby making the model more concise.



Method

Model Input

- **Positional Encoding**. We use the **eigenvectors of the graph laplacian matrix** to encode positional relationships between nodes.
  - It can be solely computed from every molecular graph's **adjacency matrix** without any complex or excessive prior knowledge.

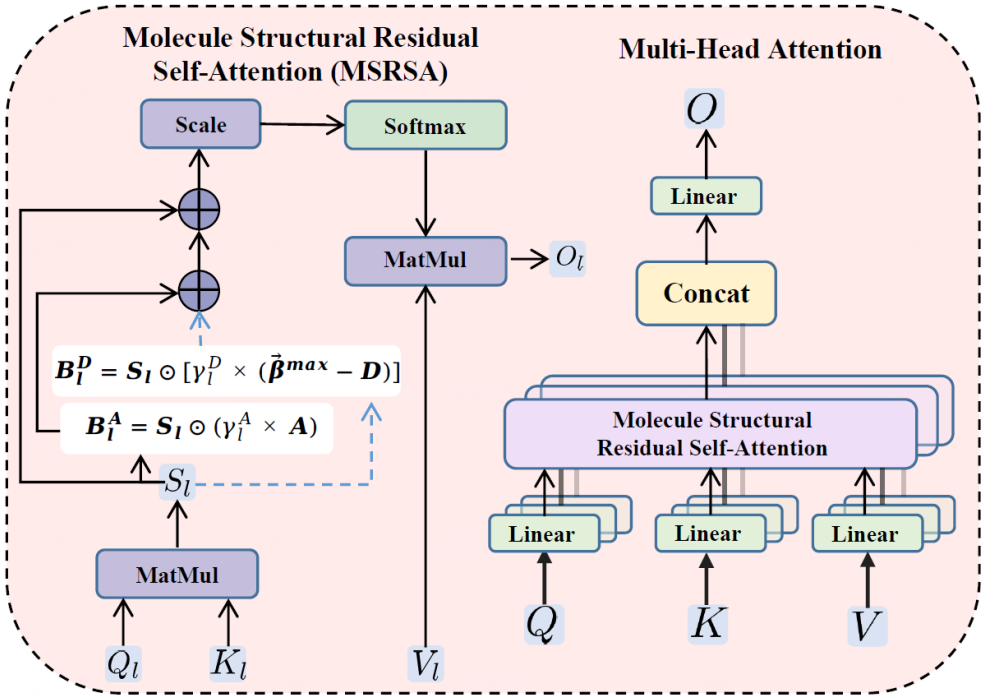


- **The final model input**, is solely obtained by adding the **laplacian positional encoding vectors**, to the **feature vectors**, embedded from the MoleBERT-tokenized input IDs

Method

■ Molecule Structural Residual Self-Attention (MSRSA)

- For molecules, **the local structure of atoms** (such as functional groups) and **the connectivity of edges** (whether they have chemical bonds or not) have a huge impact on their chemical and structural properties.
- Since the **original self-attention mechanism** is unable to effectively capture these information, we propose **an extended self-attention mechanism** based on molecular structure.



Method

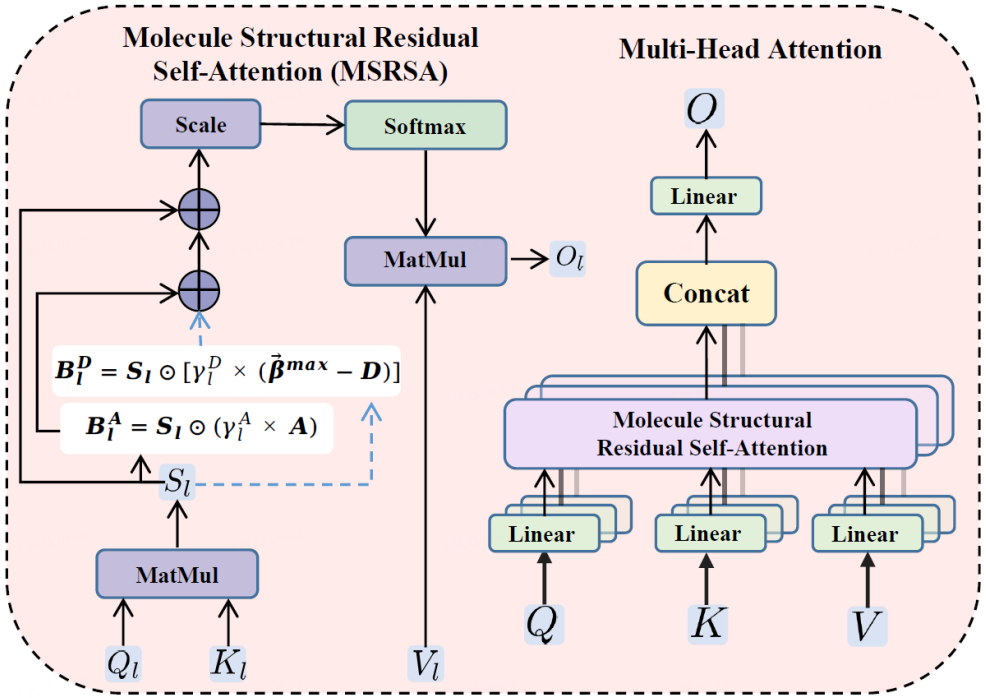
■ Molecule Structural Residual Self-Attention (MSRSA)

- Global: we use the **original self-attention mechanism** to capture global information by focusing on all atoms in the molecule.

$$\mathbf{S}_l = \mathbf{X}\mathbf{W}_l^Q(\mathbf{X}\mathbf{W}_l^K)^T = \mathbf{Q}_l\mathbf{K}_l^T$$

- Nearby: we use the **adjacency matrix** A to capture something useful about the molecule's local structure.

$$\mathbf{B}_l^A = \mathbf{S}_l \odot (\gamma_l^A \times \mathbf{A})$$



- Spatial: we use the **interatomic distance matrix** D to capture spatial structure of molecules. (only used in Decoder, based on coordinates predicted by Encoder.)

$$\mathbf{D}_{row\text{-}sub} = \vec{\beta}^{\max} - \mathbf{D},\ \vec{\beta}^{\max} = \max_{i\in[1,n]} \mathbf{D} \qquad \mathbf{B}_l^D = \mathbf{S}_l \odot (\gamma_l^D \times \mathbf{D}_{row\text{-}sub})$$

Method

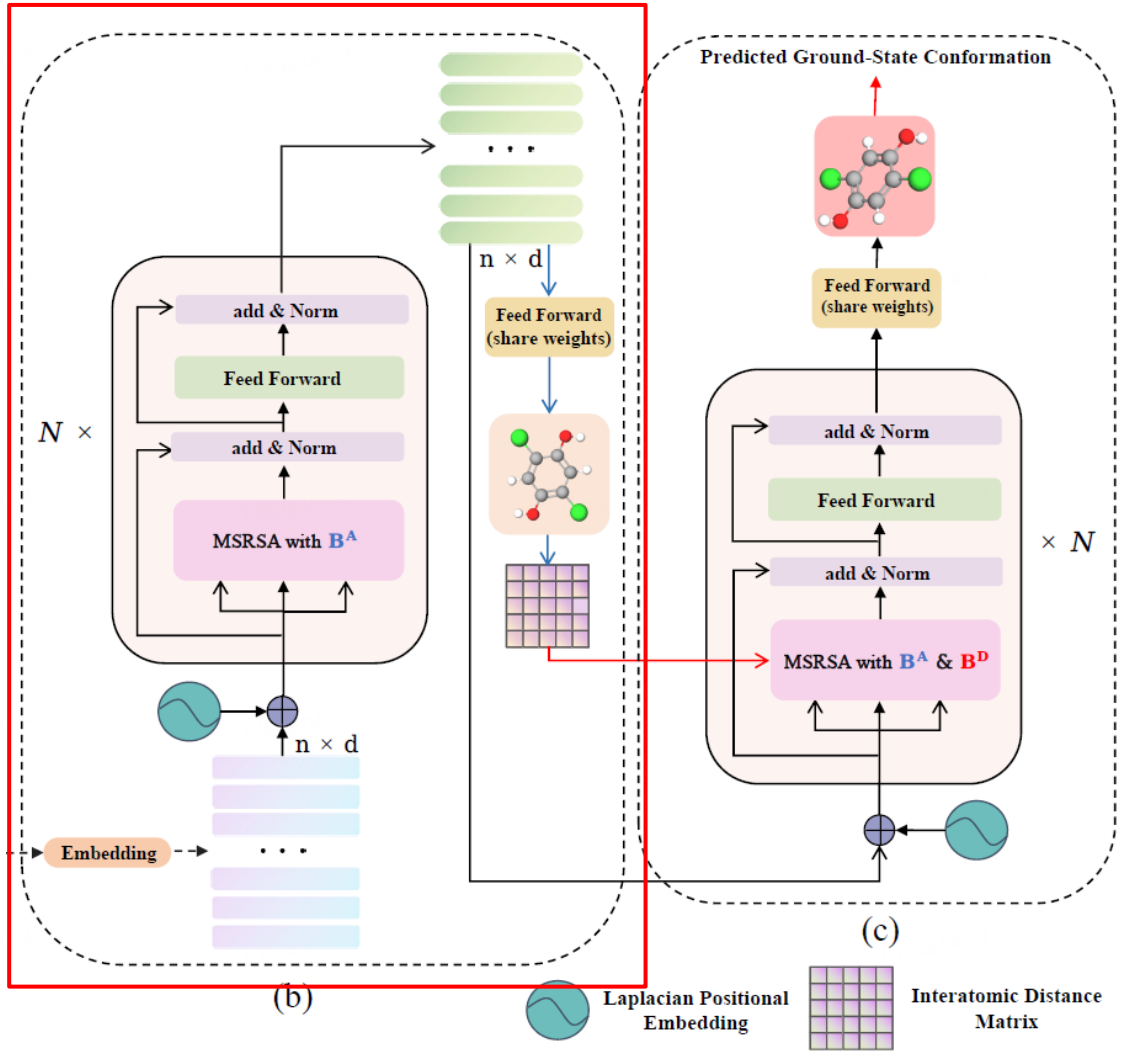
■ Model Architecture

➢ Encoder

- It takes the model input (feature vectors + laplacian positional encoding vectors) as inputs.
- It only use the **adjacency matrix** residual bias term to enhance the original self-attention mechanism.

$$\mathbf{S}_l' = \mathbf{S}_l + \mathbf{B}_l^A$$

- The output is passed through a FFN head to obtain a relatively **rough conformation prediction result** G_{cache}, and further calculate the **interatomic distance**.



$$\mathbf{B}_l^A = \mathbf{S}_l \odot (\gamma_l^A \times \mathbf{A}) \quad \mathbf{B}_l^D = \mathbf{S}_l \odot (\gamma_l^D \times \mathbf{D}_{row\text{-}sub})$$

Method

■ Model Architecture

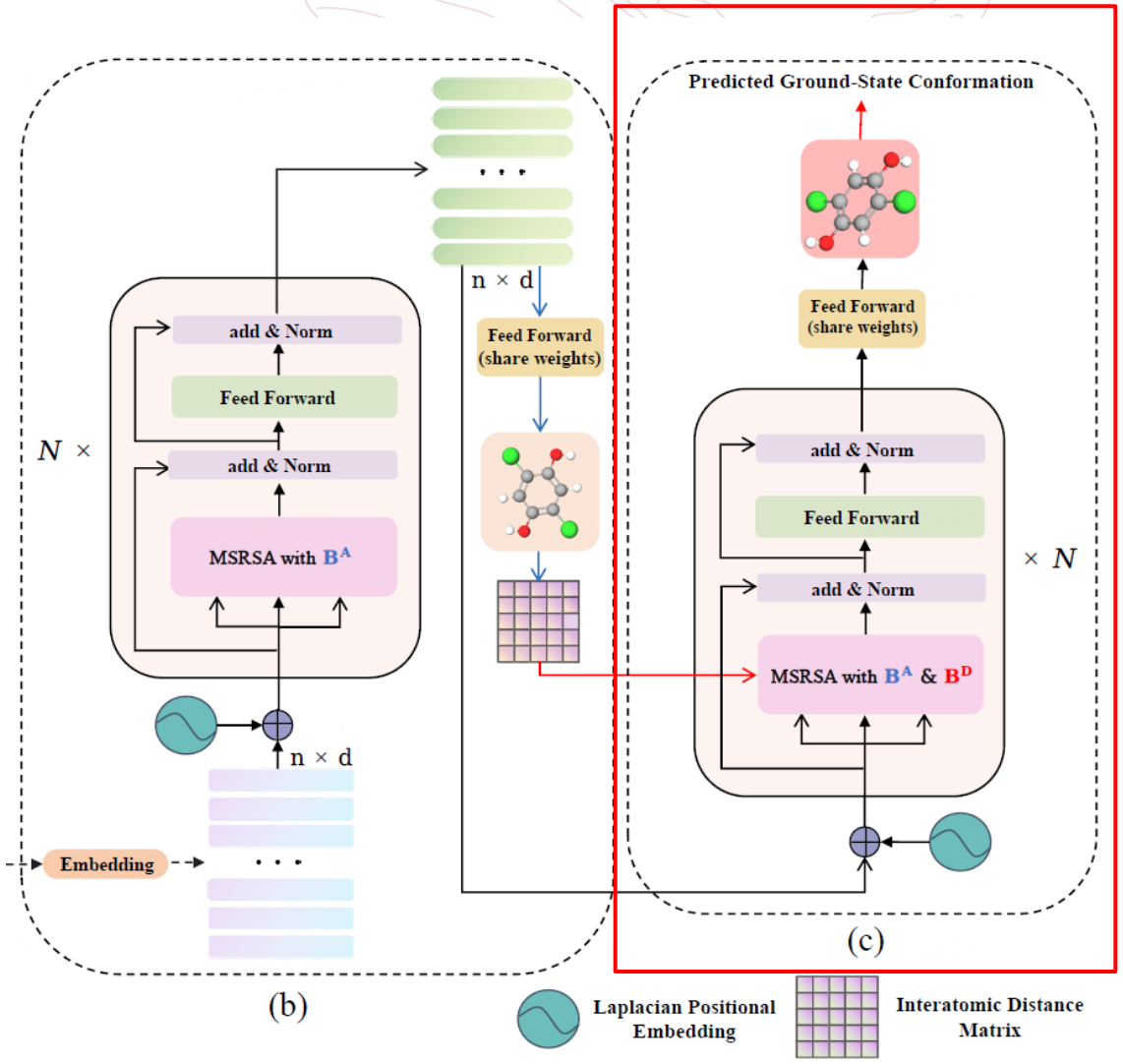
➢ Decoder

- It takes the Encoder output as inputs.
- It use both the **adjacency matrix** residual bias term and **interatomic distance matrix** D_{cache} residual bias term to enhanced the original self-attention mechanism.

$$\mathbf{S}_l' = \mathbf{S}_l + \mathbf{B}_l^A + \mathbf{B}_l^D$$

- The output is also passed through a FFN head to obtain the final conformation G^*.

$$\mathcal{L} = \mathrm{MAE}(\mathbf{D}, \mathbf{D}^*) + \mathrm{MAE}(\mathbf{D}, \mathbf{D}_{cache})$$



- Introduction
- Method
- Experiment
- Conclusion

Experiment

■ Basic Settings

➢ Datasets

- **Molecule3D:** The first benchmark introduced by [1], comprising approximately 4 million molecules.
- **QM9:** It comprises approximately 130,000 organic molecules with 9 heavy atoms.

➢ Metrics

- Based on **interatomic distance**

$$\text{D-MAE}(\{d_i\}_{i=1}^N, \{d_i^*\}_{i=1}^N) = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^*|$$

$$\text{D-RMSE}(\{d_i\}_{i=1}^N, \{d_i^*\}_{i=1}^N) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - d_i^*)^2}$$

- Based on **molecular conformation**

$$\text{C-RMSD}(\mathbf{G}, \hat{\mathbf{G}}^*) = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\|\mathbf{g}_i - \hat{\mathbf{g}}_i^*\|_2^2}$$

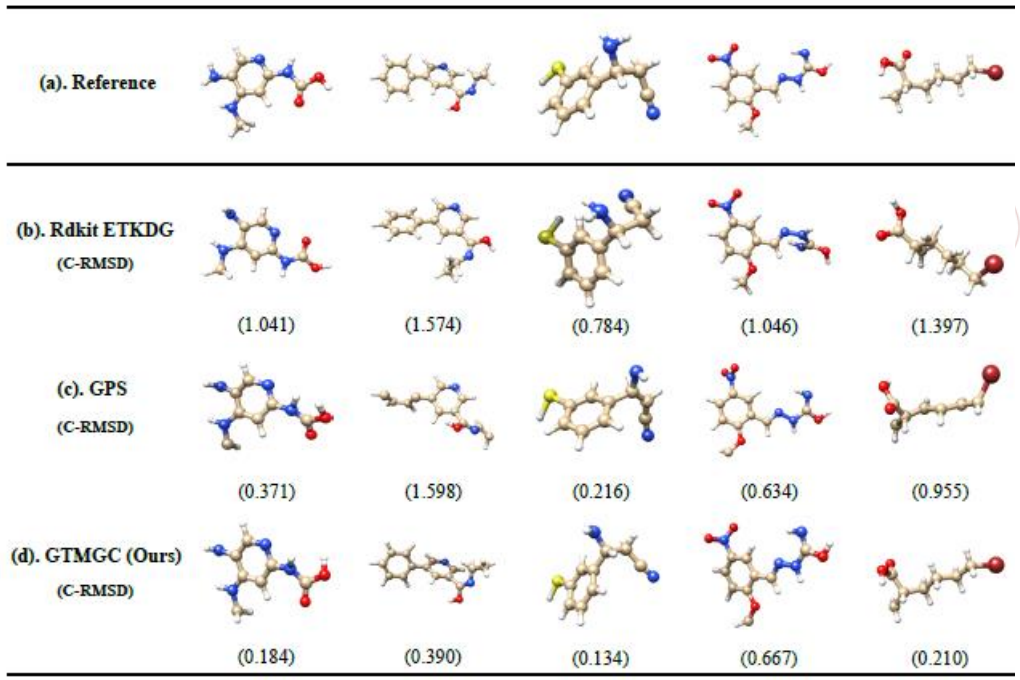
[1] Xu Z, Luo Y, Zhang X, et al. Molecule3d: A benchmark for predicting 3d geometries from molecular graphs[J]. arXiv preprint arXiv:2110.01717, 2021.

Experiment

■ Conformation Prediction

	Validation			Test		
	D-MAE↓	D-RMSE↓	C-RMSD↓	D-MAE↓	D-RMSE↓	C-RMSD↓
(a) Molcule3D Random Split						
RDKit DG	0.581	0.930	1.054	0.582	0.932	1.055
RDKit ETKDG	0.575	0.941	0.998	0.576	0.942	0.999
DeeperGCN-DAGNN (Xu et al., 2021d)	0.509	0.849	*	0.571	0.961	*
GINE (Hu et al., 2019)	0.590	1.014	1.116	0.592	1.018	1.116
GATv2 (Brody et al., 2021)	0.563	0.983	1.082	0.564	0.986	1.083
GPS (Rampášek et al., 2022)	0.528	0.909	1.036	0.529	0.911	1.038
GTMGC (**Ours**)	**0.432**	**0.719**	**0.712**	**0.433**	**0.721**	**0.713**
(b) Molcule3D Scaffold Split						
RDKit DG	0.542	0.872	1.001	0.524	0.857	0.973
RDKit ETKDG	0.531	0.874	0.928	0.511	0.859	0.898
DeeperGCN-DAGNN (Xu et al., 2021d)	0.617	0.930	*	0.763	1.176	*
GINE (Hu et al., 2019)	0.883	1.517	1.407	1.400	2.224	1.960
GATv2 (Brody et al., 2021)	0.778	1.385	1.254	1.238	2.069	1.752
GPS (Rampášek et al., 2022)	0.538	0.885	1.031	0.657	1.091	1.136
GTMGC (**Ours**)	**0.406**	**0.675**	**0.678**	**0.400**	**0.679**	**0.693**
(c) QM9						
RDKit DG	0.358	0.616	0.722	0.358	0.615	0.722
RDKit ETKDG	0.355	0.621	0.691	0.355	0.621	0.689
GINE (Hu et al., 2019)	0.357	0.673	0.685	0.357	0.669	0.693
GATv2 (Brody et al., 2021)	0.339	0.663	0.661	0.339	0.659	0.666
GPS (Rampášek et al., 2022)	0.326	0.644	0.662	0.326	0.640	0.666
GTMGC (**Ours**)	**0.262**	**0.468**	**0.362**	**0.264**	**0.470**	**0.367**

*The asterisk (*) indicates that the result for this metric was not reported in (Xu et al., 2021d).



GTMGC achieves **state-of-the-art performance** on both Molecule 3D and QM9 datasets

Experiment

■ Ablation Studies

➢ The impact of our proposed input format

Table 2: Ablation study on input format (Å).

	D-MAE↓	D-RMSE↓	⋆ C-RMSD↓
Ogb-style embeddings (Hu et al., 2021)	$.4299_{\pm.0014}$	$.7162_{\pm.0003}$	$.7431_{\pm.0198}$
Atom Type IDs	$.4338_{\pm.0002}$	$.7195_{\pm.0001}$	$.7217_{\pm.0002}$
MoleBERT Tokenized IDs (**Ours**)	$.4330_{\pm.0004}$	$.7213_{\pm.0004}$	$\mathbf{.7139_{\pm.0010}}$

⋆ means the indicator we mainly focus on.

➢ The influence of individual components within the MSRSA module

Table 3: Ablation study of MSRSA module for ground-state conformation prediction on Molecule3D random split (Å).

Index (ΔParam)	Methods						D-MAE↓	D-RMSE↓	⋆ C-RMSD↓
	LPE	MHSA	$B^{A}_{encoder}$	$B^{A}_{decoder}$	$B^{D(original)}_{decoder}$	$B^{D(row\text{-}sub)}_{decoder}$			
1 (+0)		✓					$.5464_{\pm.0026}$	$.9049_{\pm.0091}$	$.9724_{\pm.0121}$
2 (+0)	✓	✓					$.4395_{\pm.0004}$	$.7237_{\pm.0003}$	$.7388_{\pm.0069}$
3 (+48)	✓	✓	✓				$.4353_{\pm.0002}$	$.7217_{\pm.0004}$	$.7213_{\pm.0043}$
4 (+96)	✓	✓	✓	✓			$.4330_{\pm.0004}$	$.7216_{\pm.0006}$	$.7299_{\pm.0111}$
5 (+144)	✓	✓	✓	✓	✓		$\mathbf{.4325_{\pm.0002}}$	$.7214_{\pm.0006}$	$.7202_{\pm.0057}$
6 (+144)	✓	✓	✓	✓		✓	$.4330_{\pm.0004}$	$\mathbf{.7213_{\pm.0004}}$	$\mathbf{.7139_{\pm.0010}}$

⋆ means the indicator we mainly focus on.

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- We propose **a novel Transformer-based method**, GTMGC, for end-to-end prediction of 3D ground-state conformations of molecules from their 2D topological structures.
- Moreover, we introduce **a novel and simple self-attention mechanism** for molecular structure modeling, namely Molecule Structural Residual Self-Attention (MSRSA), to effectively model the molecular structure.
- Experiments show that our method achieves **significant performance improvement** over the previous methods, reaching the state-of-the-art level.

Thank You for listening!

Fanmeng Wang

2024-4-25