Journal of Cheminformatics

RESEARCH

Open Access

Tora3D: an autoregressive torsion angle prediction model for molecular 3D conformation generation

Zimei Zhang[1,2], Gang Wang[2,3], Rui Li[5,2], Lin Ni[4,2], RunZe Zhang[2,3], Kaiyang Cheng[4,2], Qun Ren[4,2], Xiangtai Kong[2,3], Shengkun Ni[2,3], Xiaochu Tong[2,3], Li Luo[7], Dingyan Wang[6], Xiaojie Lu[2,3], Mingyue Zheng[1,2,3,4*] and Xutong Li[2,3*]

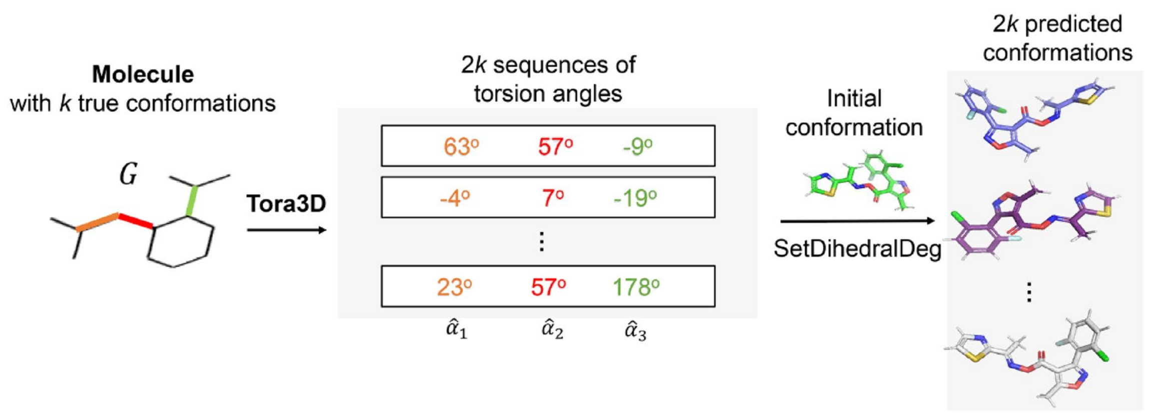
Abstract

Three-dimensional (3D) conformations of a small molecule profoundly affect its binding to the target of interest, the resulting biological effects, and its disposition in living organisms, but it is challenging to accurately characterize the conformational ensemble experimentally. Here, we proposed an autoregressive torsion angle prediction model Tora3D for molecular 3D conformer generation. Rather than directly predicting the conformations in an end-to-end way, Tora3D predicts a set of torsion angles of rotatable bonds by an interpretable autoregressive method and reconstructs the 3D conformations from them, which keeps structural validity during reconstruction. Another advancement of our method over other conformational generation methods is the ability to use energy to guide the conformation generation. In addition, we propose a new message-passing mechanism that applies the Transformer to the graph to solve the difficulty of remote message passing. Tora3D shows superior performance to prior computational models in the trade-off between accuracy and efficiency, and ensures conformational validity, accuracy, and diversity in an interpretable way. Overall, Tora3D can be used for the quick generation of diverse molecular conformations and 3D-based molecular representation, contributing to a wide range of downstream drug design tasks.

Keywords Conformations generation, Autoregressive, Transformer, Deep learning, Small molecules

*Correspondence:
Mingyue Zheng
myzheng@simm.ac.cn
Xutong Li
lixutong@simm.ac.cn
Full list of author information is available at the end of the article

Graphical Abstract


Molecule
with k true conformations
G
Tora3D
$2k$ sequences of torsion angles
63° 57° -9°
-4° 7° -19°
⋮
23° 57° 178°
$\hat{\alpha}_1$ $\hat{\alpha}_2$ $\hat{\alpha}_3$
Initial conformation
SetDihedralDeg
$2k$ predicted conformations
⋮

Introduction

Molecular conformation is important for determining a molecule's chemical and physical properties. Conformation generation is also important in applications such as quantitative structure–activity relationships (QSAR), docking, and virtual screening for drug development [1–5]. The intersection between deep learning and conformation generation has recently drawn attention to their accuracy and efficiency. Deep learning methods generate small molecular conformations set with high accuracy and efficiency, which can accelerate molecular docking and improve its accuracy. Deep learning-based models can also learn molecular representations incorporating 3D structural information, which provides a way forward to improve the predictive modeling of small molecule bioactivities and properties [6].

Over the past decades, generating an accurate 3D structure for a small chemical compound is not trivial. Molecular conformation can be physically determined using X-ray crystallography, but it is prohibitively costly for industry-scale tasks [7]. Ab initio methods can accurately predict molecular geometry, such as density functional theory (DFT) [8], but these approaches usually take up to several hours per small molecule [9]. To handle large-scale molecules, people start turning to classical force fields methods, like UFF [10] or MMFF [11], to estimate conformations, which is efficient but extremely inaccurate [12]. In addition, there are some classical methods to generate low-energy conformations by iteratively enumerating all possible conformations. Systematic search methods such as Monte Carlo simulation (MC) [13], and Distance geometry (DG) [14] are effective in exploring the conformational space, but they can converge to a local minimum rather than the global minimum. Stochastic methods such as Genetic Algorithm (GA) [15] randomly modify the structural parameters of the molecules to increase the probability of finding a global minimum, but the associated computational cost is an important limitation. In systematic search methods, rule-based fast conformational search algorithms such as Omega [3] and Conformator [2] are preferred for sampling large molecular libraries to generate representative conformation ensembles.

Recent deep learning developments hold promise for improving the prediction of the conformation ensembles of small molecules. Generative deep learning can produce structural candidates by predicting possible valid coordinates or distance matrices of a molecule. Since directly generating the 3D coordinates of atoms from the molecular graph like CVGAE [7] faces the problem of SE-(3) invariance, many researchers go for the prediction of the atomic pairwise distances, i.e. distance matrices which are invariant to rotation and translation. GraphDG [16] proposes to model the distribution of inter-atomic distances, while CGCF [17] and ConfVAE [18] take the distribution of distances as intermediate variables to generate conformations. Recently, Ganea et al. [19] further proposed GeoMol to solve the SE-(3) invariance by generating local 3D structures and torsion angles. There are also deep learning models that take an iterative approach to find low-energy conformations. ConfGF [20] directly estimates the gradient field of the log density of the atomic coordinates. GeoDiff [21] uses an SE-(3) equivariant score model to reverse a diffusion process that adds independent Gaussian noise to each atomic coordinate in Euclidean space. These methods can generate a conformation accurately by denoising a point cloud where atoms are in random initial positions but are much more time-consuming. GeoDiff [21] takes about 5000 denoising steps, which costs 9–10 min to generate conformations for a molecule on average.

Although deep learning models have been explored for molecular conformation generation in the hope of combining high accuracy with fast sampling, they typically have the drawback of generating invalid conformations. Most graph neural network (GNN)-based methods fail to learn long-range interactions in graphs, and thus cannot accurately capture dependencies among dihedral angles, which would lead to conflicts among local structures. In addition, it is difficult for distance geometry-based methods to enforce geometric graph constraint [22], hence the accumulated errors in bond angle and length would lead to invalid local structures. To address the problem, systematic search methods assume that bond lengths and some local structures in molecules are essentially constant, and promote slight variations in rotatable bonds to gradually change the conformation of the molecules [23] while avoiding the conflict between the local structures. Recently, some studies have also proposed that rotatable bonds play a crucial role in determining the conformation of molecules, such as Torsion Library [24] and TorsionNET [25].

Here, we build a deep learning model, namely Tora3D, to predict the torsion angles combinations of all rotatable single bonds in a molecule from a 2D molecular graph, to obtain the set of predicted conformations. Like systematic approaches, our methodology follows a basic assumption that the conformational space mainly originates from the rotation of single bonds in the molecule, while keeping bond lengths and angles [23] constant. We replace the time-consuming and compute-intensive iterative process of rotatable bonds in a systematic method with an autoregressive deep learning model. The combination of deep learning and prior knowledge guarantees the accuracy, speed and validity of conformation generation while avoiding the disadvantages of the systematic method. The framework of Tora3D is designed to address the problems inherent in previous methods: (1) An autoregression neural network with an attention mechanism can guarantee the overall structural validity of the molecular conformation. The autoregressive neural network predicts the torsion angles of rotatable bonds in a molecule one by one. Hence, Tora3D could consider the dependencies among each dihedral angle to avoid clashes among local structures, and the attention mechanism can explain the dependencies and ensure spatial rationality. (2) Reconstructing the conformation by a two-stage generation procedure can guarantee the local structural validity in molecular conformation. Tora3D utilizes predicted torsion angles to assemble valid local structures that were constructed of bond lengths and angles determined by standard cheminformatics tools. Compared with directly generating conformations in an end-to-end way, the two-stage generation procedure of Tora3D can significantly reduce the dimensionality of the sample space and avoid local structural invalidity caused by wrong bond lengths and angles. (3) Tora3D could generate a set of relatively low-energy molecular conformations quickly by giving relative energies when making inferences. Overall, Tora3D aims at achieving a balance among three aspects of performance in the conformational generation including accuracy, validity, and diversity.

Method

Notation

Firstly, the symbols and notations used here were summarized in Table 1. $G = (V, E)$ represents a molecular graph, in which $V = \{v_1, v_2, \ldots, v_{|V|}\}$ is the set of feature

Table 1 List of symbols and notations used in the paper

Symbol	Description
G	The molecular graph
$V = \{v_1, v_2, \ldots, v_{\lvert V \rvert}\}$	The set of feature vectors of atoms (nodes)
$E = \left\{e_{ij} \middle\| (i,j) \in V \times V\right\}$	The set of feature vectors of bonds (edges)
h_v^t	The representation of the atom *v* in the *t* layer
$\boldsymbol{\alpha}_l$	A true normalized torsion angle value
$\widehat{\boldsymbol{\alpha}}_l$	A predicted normalized torsion angle value
$A = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3, \boldsymbol{\alpha}_{4,} \ldots, \boldsymbol{\alpha}_l, \ldots\}$	The sequence of true normalized torsion angle values
$\widehat{A} = \{\widehat{\boldsymbol{\alpha}_1}, \widehat{\boldsymbol{\alpha}_2}, \widehat{\boldsymbol{\alpha}_3}, \widehat{\boldsymbol{\alpha}_4}, \ldots, \widehat{\boldsymbol{\alpha}}_l, \ldots\}$	The sequence of predicted normalized torsion angle values
$\boldsymbol{\tau}_l^0$ or $\boldsymbol{\tau}_l$	The torsion angle representation
$T^0 = \left\{\boldsymbol{\tau}_1^0, \boldsymbol{\tau}_2^0, \boldsymbol{\tau}_3^0, \boldsymbol{\tau}_4^0, \ldots, \boldsymbol{\tau}_l^0, \ldots\right\}$ or $T = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \boldsymbol{\tau}_4, \ldots, \boldsymbol{\tau}_l, \ldots\}$	The sequence of torsion angle representations
k	The true number of conformations of a molecule

vectors of atoms (nodes) and $E = \{e_{ij} | (i,j) \in V \times V\}$ is the set of feature vectors of bonds (edges). The atomic and bond feature vectors were drawn from the input features employed by AttentiveFP, a molecular structural representation scheme based on the graph attention mechanism [26]. The h_v^0 and h_v^T represent the initial and updated atomic representations, respectively. The α_l represents a true normalized torsion angle value (the normalized operation will be discussed later) and $\widehat{\alpha}_l$ is a predicted one. The $A = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4 \ldots, \alpha_l, \ldots\}$ and $\widehat{A} = \{\widehat{\alpha_1}, \widehat{\alpha_2}, \widehat{\alpha_3}, \widehat{\alpha_4}, \ldots, \widehat{\alpha_l}, \ldots\}$ represents the sequence of true and predicted normalized torsion angle values, respectively. Both τ_l^0 and τ_l represent the torsion angle representations, where τ_l^0 represents the torsion angle obtained by the Torsion representation module and used as the initial input of the Transformer module, while τ_l denotes the updated torsion angle representation (these two modules will be described below). The $T^0 = \{\tau_1^0, \tau_2^0, \tau_3^0, \tau_4^0, \ldots, \tau_l^0, \ldots\}$ and $T = \{\tau_1, \tau_2, \tau_3, \tau_4, \ldots, \tau_l, \ldots\}$ denotes all torsion angle representations of a molecule, namely the sequence of torsion angle representations. The k denotes the true number of conformations of a molecule.

Overview

In this section, we want to outline our approach. Tora3D is a neural network model that can predict a series of sequences of torsion angles of all rotatable single bonds in a molecule from a 2D molecular graph (Fig. 1). Inputting a molecular graph (containing information about nodes and edges, as well as topology), Tora3D was trained to predict all torsion angle values of the molecule. Tora3D is divided into two parts: Torsion representation module ($\mathrm{F_r}$) and Transformer module ($\mathrm{F_t}$). The former obtains the sequence of torsion angle representations (T^0) of the molecule from the 2D molecular graph (G) (Eq. 1), and the latter obtains the sequence of normalized torsion angle values $\widehat{A}$ from (T^0) (Eq. 2). Once the torsion angle values have been predicted by Tora3D, they can be used to rebuild conformations of the small molecule from the initial conformation (Fig. 1).

$$T^0 = \mathrm{F_r}(G) \tag{1}$$

$$\widehat{\mathrm{A}} = \mathrm{F_t}\left(T^0\right) \tag{2}$$

To avoid over parametrization, the normalized torsion angle defined by Ganea et al. was used here, which is

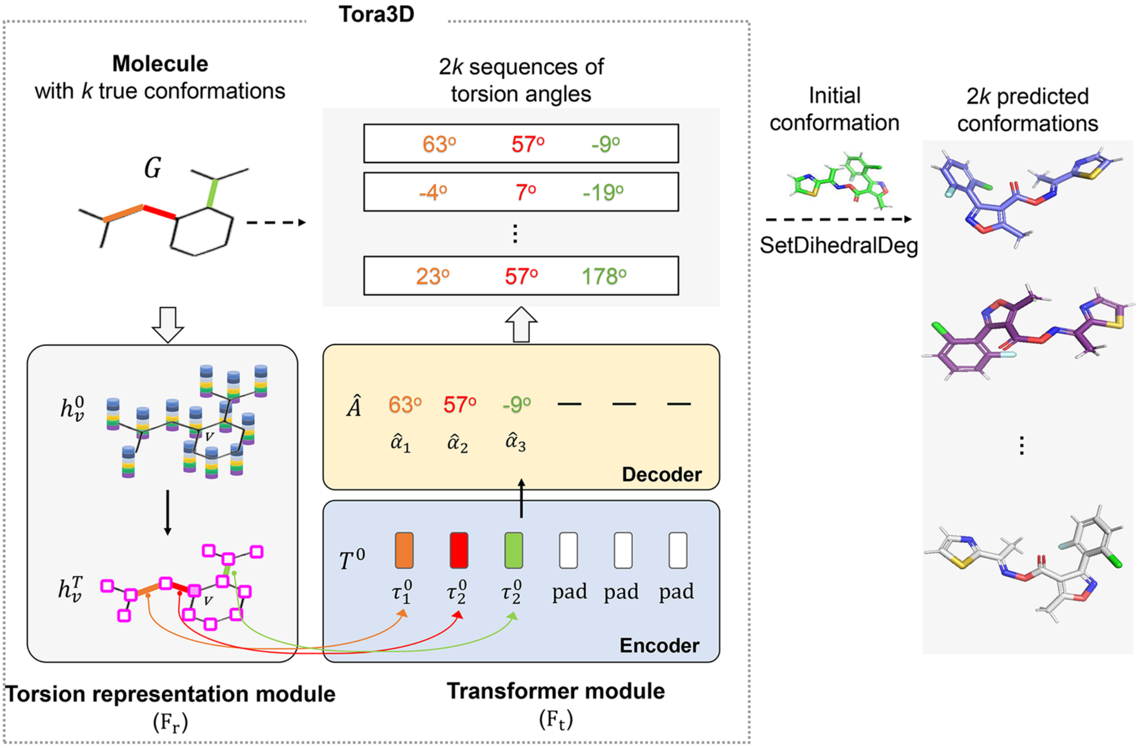


Fig. 1 The framework of Tora3D and the usage of it to generate small molecule conformations. For the case molecule with 3 rotatable bonds (orange, red and green), Tora3D generates 2 k sequences of 3 torsion angles, which can be used to rebuild 2 k predicted conformations

uniquely determined independent of the choice of terminal atoms [19]. Specifically, the normalized torsion angle α_l is calculated as Eq. 3 to Eq. 5.

$$s_{\mathrm{a}_m\mathrm{b}_n}\underline{\underline{\mathrm{def}}}\left[\frac{\cos\left(\Delta_{\mathrm{a}_m\mathrm{b}_n}\right)}{\sin\left(\Delta_{\mathrm{a}_m\mathrm{b}_n}\right)}\right] \tag{3}$$

$$s\,\underline{\underline{\mathrm{def}}}\sum{}_{m,n}c\ \cdot s_{a_m b_n} \tag{4}$$

$$\alpha_l\,\underline{\underline{\mathrm{def}}} - arctan\left(\frac{s}{\|s\|}\right) \tag{5}$$

where ${}_{\mathrm{a}_m\mathrm{b}_n}$ refers to the angle of twist with terminal atom a_m and b_n as shown in Fig. 2b. And *c* is a constant, to avoid $s_{\mathrm{a}_m\mathrm{b}_n}$ canceling each other out due to summation. It has been demonstrated that when a rotatable bond rotates by an angle γ, the normalized torsion angle α correspondingly rotates γ [19].

Torsion representation module

The first part of Tora3D is the Torsion representation module ($\mathrm{F_r}$) (Fig. 2a), which obtains the sequence of torsion angle representations (T^0) of the molecule from the 2D molecular graph (***G***) (Eq. 1). First, the position information *pos* is concatenated to the initial feature vectors *v* of each atom to obtain the initial representation h_v^0 of each atom (Eq. 6 and Eq. 7). Then, the initial representations $\{h_{v_1}^0, h_{v_2}^0, \ldots, h_{v_{|V|}}^0\}$ of the atoms are put into self-attention based Φ_{update} to updates the atomic representations denoted as h_v^T (Eq. 8). Finally, the updated atomic representations are combined with corresponding edges' representations to obtain the initial representation τ_l^0 for each torsion angle by $\Phi_{readout}$ (Eq. 9 to Eq. 12). Following are the detailed algorithms.

The position vector (pos) in Eq. 7 consists of three parts to ensure that it can contain the position information of nodes throughout the molecular graph. The first part $pos\_wl = WeisfeilerLehman(G)$ is a vector calculated using the Weisfeiler-Lehman (WL) algorithm [27], which is used to detect graph isomorphism. By WL, nodes with the same topology have the same *pos_wl*. The second part $pos\_d = embedding(degree)$ is the degree's embedding of the atom. The third part $pos\_a = FC(Adjacencymatrix)$ is the position vector representation of each atom obtained from the adjacency matrix that can provide connection information of the graph.

Concatenating these three parts (Eq. 6), the final position vector *pos* replaces the position scalar in the original transformer added to each token, and it is concatenated with the initial feature vectors *v* of an atom to obtain the initial representation h_v^0 of each atom (Eq. 7).

$$pos = pos\_wl\|pos\_d\,\|pos\_a \tag{6}$$

$$h_v^0 = v\|pos \tag{7}$$

Φ_{update}^t(Eq. 8) is the atomic update module that updates the initial representation of each atom, i.e. $\{h_{v_1}^0, h_{v_2}^0, \ldots, h_{v_{|V|}}^0\}$. The algorithm is borrowed from the self-attention module in the transformer to overcome the difficulties of traditional GNN's long-distance messaging. After T (hyperparameter, Additional file 1: Table S1) times, the updated representations $\{h_{v_1}^T, h_{v_2}^T, \ldots, h_{v_{|V|}}^T\}$ were obtained, in which each atom interacts with all other atoms of the graph through attention.

$$\begin{aligned} h_{v_1}^{t+1}, h_{v_2}^{t+1}, \ldots, h_{v_{|V|}}^{t+1} &= \Phi_{update}^t\left(h_{v_1}^t, h_{v_2}^t, \ldots, h_{v_{|V|}}^t\right) \\ &= self - attention\left(h_{v_1}^t, h_{v_2}^t, \ldots, h_{v_{|V|}}^t\right) \end{aligned} \tag{8}$$

Given the updated atomic representations $H = \{h_{v_1}^T, h_{v_2}^T, \ldots, h_{v_{|V|}}^T\}$ and the edge information E, the initial representations of torsion angles $T^0 = \left\{\tau_1^0, \tau_2^0, \tau_3^0, \tau_4^0, \ldots, \tau_l^0, \ldots\right\}$ are obtained by Φ_{readout} (Eq. 9).

$$T^0 = \Phi_{readout}(H, E) \tag{9}$$

Specifically, the representation of each torsion angle τ_l^0 is obtained by integrating information about the neighboring edges of each rotatable bond and the corresponding atoms (Eq. 10).

$$\tau_1^0 = h_{atoms}\|e_{edges} \tag{10}$$

Here, e_{edges} for each edge is the concatenated information of itself e_{ij} with its neighboring edges e_{ai} and e_{bj} (Eq. 11).

$$e_{edges} = e_{ai}\|e_{ij}\,\|e_{bj} \tag{11}$$

For example, as shown in Fig. 2b, $e_{ai} = \left(e_{a_1 i} + e_{a_2 i} + \ldots\right)/_{|e_{ai}|}$ is the integrated representation of edges between atom i and atom a_1, a_2 and a_3, and $e_{bj} = \left(e_{b_1 j} + e_{b_2 j} + \ldots\right)/_{|e_{bj}|}$ denotes the integrated representation of edges between atom j and b_1 and b_2.

Similarly, the two atoms of a rotatable bond concatenate information about themselves h_i and h_j, with all terminal atoms $h_a = \left(h_{a_1} + h_{a_2} + \ldots\right)/_{|h_a|}$ and $h_b = \left(h_{b_1} + h_{b_2} + \ldots\right)/_{|h_a|}$ to obtain the representation of atoms h_{atoms} (Eq. 12).

$$h_{\mathrm{atoms}}{=}h_{\mathrm{a}}\|h_{\mathrm{i}}\,\|h_j\,\|h_{\mathrm{b}} \tag{12}$$

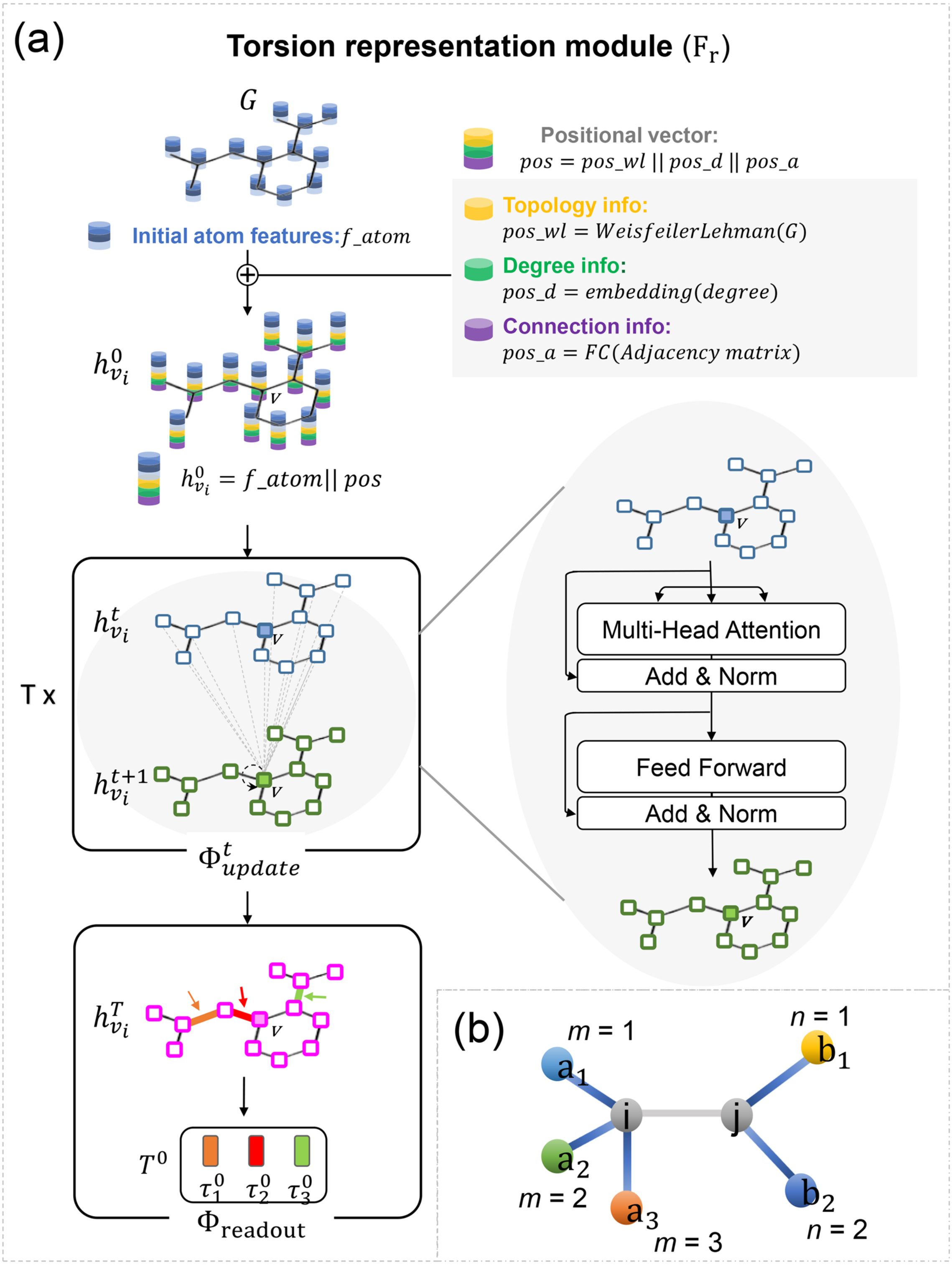


Fig. 2 **a** Torsion representation module $(\mathrm{F_r})$. **b** The calculation of the normalized torsion angle α of bond ij. There are three options for end atom a and two options for end atom b in this case

Thus, by the above process of $\Phi_{readout}$, the initial representation of torsion angles T^0 is obtained from the overall atomic representation of each side of a rotatable bond. Unlike some models that predict atomic extrinsic coordinates of a molecule from atom representation, Tora3D is SE (3)-invariant by focusing on torsional space specific to the molecule (intrinsic coordinates).

Transformer module

The Transformer module (F_t) (Fig. 3) is used to accept the sequence of torsion angle representations T^0 as input and output the sequence of predicted torsion angle values $\widehat{A}$ (Eq. 2). Compared to the original Transformer's framework, the Transformer module has a few changes as detailed below.

Transformer encoder

Transformer encoder maps an input sequence of torsion angle representations T^0 to the sequence of updated torsion angle representations T (Fig. 3). Gaussian noise (hyperparameter, Additional file 1: Table S1) with a mean of zero and standard deviation of 5.0, which is determined by hyperparameter searching, was added to T^0 to allow the model generates multiple conformations by introducing an element of randomness. The position coding of the original Transformer was removed since τ_i^0

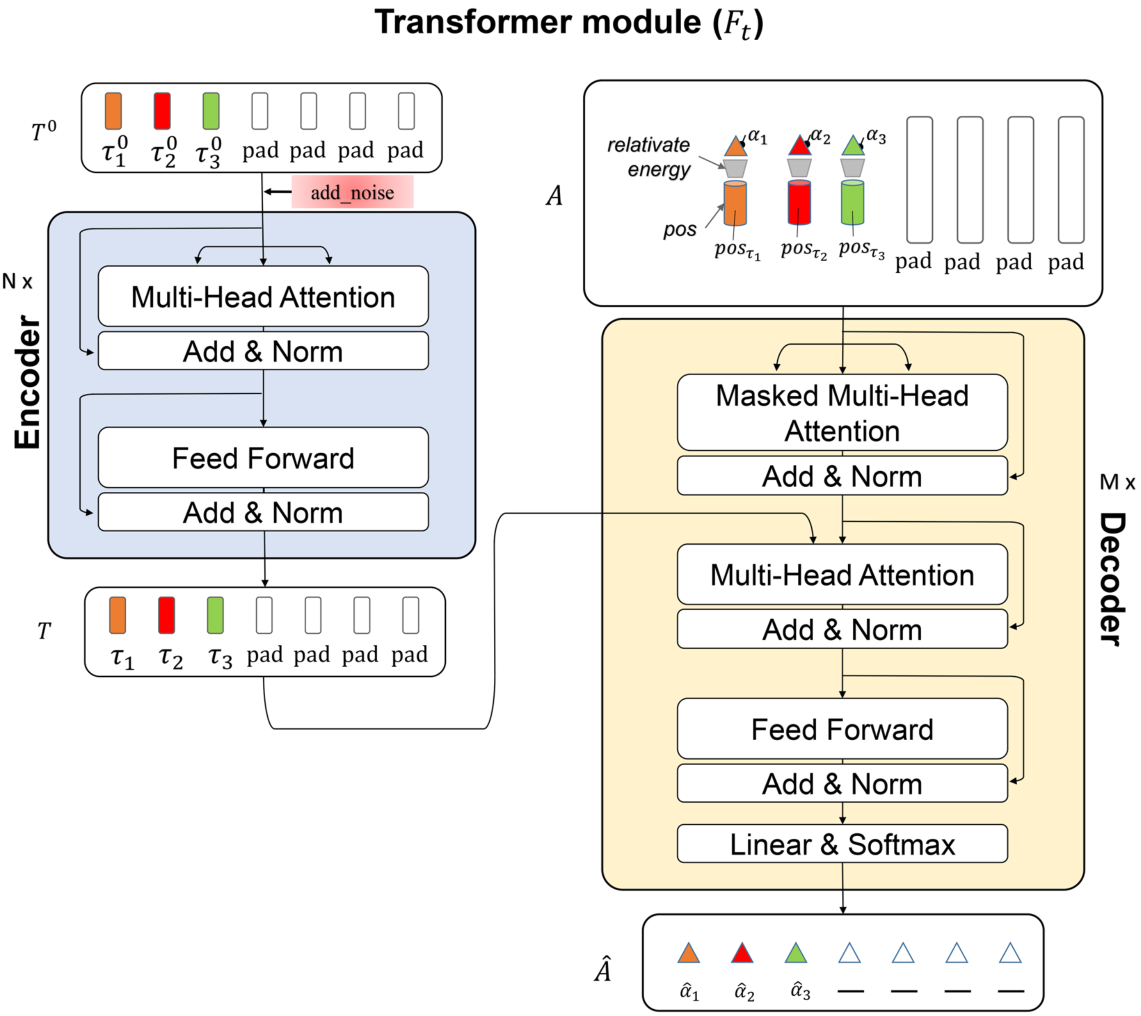

Transformer module (F_t)
T^0
τ_1^0 τ_2^0 τ_3^0 pad pad pad pad
add_noise
N x
Encoder
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
T
τ_1 τ_2 τ_3 pad pad pad pad
A
relativate energy
pos
α_1 α_2 α_3
pos_{τ_1} pos_{τ_2} pos_{τ_3} pad pad pad pad
Masked Multi-Head Attention
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Linear & Softmax
M x
Decoder
$\hat{A}$
$\hat{\alpha}_1$ $\hat{\alpha}_2$ $\hat{\alpha}_3$

Fig. 3 The transformer module has an encoder-decoder structure that uses stacked self-attention and fully connected layers. The sequence of initial torsion angle representations T^0 was input into the encoders (left) and updated for N (hyperparameter, Additional file 1: Table S1) times to obtain a sequence of continuous representations T. Given T, the M (hyperparameter, Additional file 1: Table S1) stack of decoders (right) generates an output sequence, i.e., normalized torsion angle values $\widehat{A}$, one element at a time. At each step the model is autoregressive, consuming the previously generated angle values as additional input when generating the next

already contains the position information, as defined in the Torsion representation module.

Transformer decoder

Given T, the decoder generates the sequence of predicted normalized torsion angle values $\widehat{A}$ (Fig. 3). The model predicts each $\widehat{A}$ with reference to the previously predicted torsion angle values i.e., $\widehat{a}$, and their corresponding position encoding (pos_{τ_i}), as well as the relative energy (kcal/mol) for each molecular conformation. Such an autoregressive approach avoids local structural clashes. Moreover, the relative energy as input also allows the model to generate energy-specific conformations.

Conformation generation

As shown in Fig. 1, Tora3D predicts the torsion angle value of rotatable bonds consisting of heavy atoms. Once the torsion angles of all the rotatable bonds of a small molecule have been predicted, the fragments of the small molecule can be assembled to form the overall conformation. There are already some accurate and effective knowledge-based algorithms for generating conformation ensembles from fragments have been demonstrated, represented by the commercial algorithm Omega [3] and the freely available algorithm Conformator [2]. Thus, we directly use one of the conformations generated by the Conformator as the initial conformation and twist it to obtain the predicted conformation based on the torsion angle values predicted by the Tora3D. The comparison between the initial conformations and Tora3D's generated conformations are shown in supporting information, Additional file 1: Figure S4 and Additional file 1: Table S4. To be specific, we reset the torsion angles of an initial conformation by the SetDihedralDeg function in RDKit [28].

Experiments

Dataset and split

Following previous works [17, 20], the Geometric Ensemble Of Molecules (GEOM)-Drugs dataset was used for building the model. The GEOM-drugs dataset contains 118,434,901 molecular conformations of 304,466 unique molecules, generated by advanced sampling and semi-empirical DFT. Relative energy of each conformation is also included in GEOM-drugs, which is the difference between the absolute energy of a conformation and that of the lowest-energy conformation. A value of 0 kcal/mol signifies the energy of the lowest-energy conformation. The molecules in GEOM-Drugs are annotated by experimental data related to biophysics, physiology, and physical chemistry [1]. The test set of Shi et al. containing 200 molecules was also used here for performance evaluation (test set I) [20]. Analysis showed that the number of conformations of each molecule in test set I is less than 100, while the number of conformations of each molecule in the GEOM-Drugs dataset ranges from 0 to 12,000 (Additional file 1: Figure S1). Thus, Test set I is not reflective of the overall modeling dataset. Therefore, an additional test set II was collected, which contains randomly selected 1000 molecules and their conformation number ranges from 0 to 500, same as the range of conformation number of molecules in the GEOM-Drugs dataset. With a similar distribution to the entire dataset, test set II consists of more conformations with higher diversity than test set I, which was used to further evaluate model performance affected by conformational flexibility.

Evaluation indicators

Coverage (COV) and Matching (MAT) score are used to measure the diversity and accuracy respectively. COV score reports the percentage of reference conformers that are produced by the predicted ensemble. MAT score reports the minimum RMSD between a generated conformer and the references. Following the conventional Recall measurement, COV-R and MAT-R can be defined as [21]:

$$\mathrm{COV\text{ - }R}\left(S_g, S_r\right) = \frac{1}{|S_r|}\left|\left\{R \in S_r | RMSD\left(R, \hat{R}\right) < \delta, R \in S_g\right\}\right| \tag{13}$$

$$\mathrm{MAT\text{ - }R}\left(S_g, S_r\right) = \frac{1}{|S_r|}\sum_{R \in S_r} \min_{\hat{R} \in S_g} RMSD\left(R, \hat{R}\right) \tag{14}$$

Here, S_g is the set of generated conformations and S_r is the set of reference conformations of a molecule. $\widehat{R}$ and R refer to a generated conformation and a reference conformation, respectively. The δ is set as 1.25. The above equations are used for calculating COV-R and MAT-R (Recall). And to calculate COV-P and MAT-P (Precision), S_g and S_r should be swapped. Generally, higher COV rates or lower MAT score suggest that more realistic conformations are generated. And the Recall metrics concentrate more on the diversity, while the Precision metrics depend more on the quality. Consistent with previous work, we predicted and generated twice as many conformations as the number of true conformations for each molecule.

Results

Model performance in conformational diversity and accuracy

We have compared Tora3D with several recent popular models of molecular 3D conformation prediction: CVGAE [7], GraphDG [16], CGCF [17], ConfVAE[18],

Table 2 Performance comparison of models on the GEOM-drugs dataset (Test set I)

Models	COV-R(↑)	MAT-R(↓)	COV-P(↑)	MAT-P(↓)	Speed (s/ molecule)
CVGAE	0.00	3.0702	–	–	–
GraphDG	8.27	1.9722	2.08	2.4340	–
CGCF	53.96	1.2487	21.68	1.8571	-
ConfVAE	55.20	1.2380	22.96	1.8287	10–16
GeoMol	67.16	1.0875	–	–	1–4
ConfGF	62.15	1.1629	23.42	1.7219	> 600
GeoDiff	**82.96**	0.9525	48.27	1.3205	540–600
Tora3D	80.37	**0.9272**	**62.22**	**1.1524**	**5–8.4**

* The results of CVGAE[7], GraphDG[16], CGCF[17], ConfGF[20], and GeoDiff[21] are borrowed from Shi et al.[20]. The experiments of ConfVAE and GeoMol[19] were implemented by ourselves. The inference speed for a molecule of each model was tested by ourselves

Table 3 Performance of different position embedding

Models	COV-R(↑)	MAT-R(↓)	COV-P(↑)	MAT-P(↓)
Without position embedding	57.32	1.1742	62.05	1.4200
Learnable-position embedding	73.21	1.0109	**62.85**	1.4459
Tora3D	**81.92**	**0.9297**	62.16	**1.1600**

GeoMol [19], ConfGF [20], and GeoDiff [21]. In addition, we have conducted the comparisons with torsion angle prediction methods including Torsion Library [24] and TorsionNET [25]. The implementation is shown in supporting information, Additional file 1: Figure S3 and Additional file 1: Table S3. As shown in Table 2, Tora3D shows superior performance compared to the above-mentioned models in conformational diversity (higher COV) and accuracy (lower MAT) on the same test set (Test set I) that was reported in Shi et al. [20]. Although its COV-R is slightly lower than GeoDiff, Tora3D makes a trade-off between accuracy and efficiency. Tora3D is relatively fast and can predict all conformations of a molecule within 5 to 8.4 s, while GeoDiff needs about 10 min for a molecule.

Position embedding is devised to capture the position/location of the node within the broader context of the graph structure to tackle the problem that conventional GNN architecture hardly learns long-range patterns in graphs. The importance of the position embedding in Tor3D is verified by an ablation experiment by removing the position embedding from the model or replacing it with a learnable position embedding that represents the position of atoms. It can be seen in Table 3 that Tora3D with the specially designed position embedding provides better performance, especially on conformation accuracy and coverage. The results in Table 3 demonstrate that removing positions embedding for nodes in Tora3D, which is just like a conventional GNN architecture, does harm the quality of conformation generation. And our strategy addresses the issue of capturing long-range node dependencies, leading to better accurate and diverse conformations than learnable-position embedding.

In addition, Tora3D uses a basic assumption same to systematic methods that the only factor changing the conformation of a molecule is rotatable bonds. The number of rotatable bonds (nRotb) plays a decisive role in molecular flexibility, as the space of possible conformations grows exponentially with it. Thus, nRotb would affect the prediction performance (Fig. 4). We have implemented ConfVAE and GeoMol [19] to test the effect of nRotb based on test set II. Here ConfVAE and GeoMol were selected for comparison because they

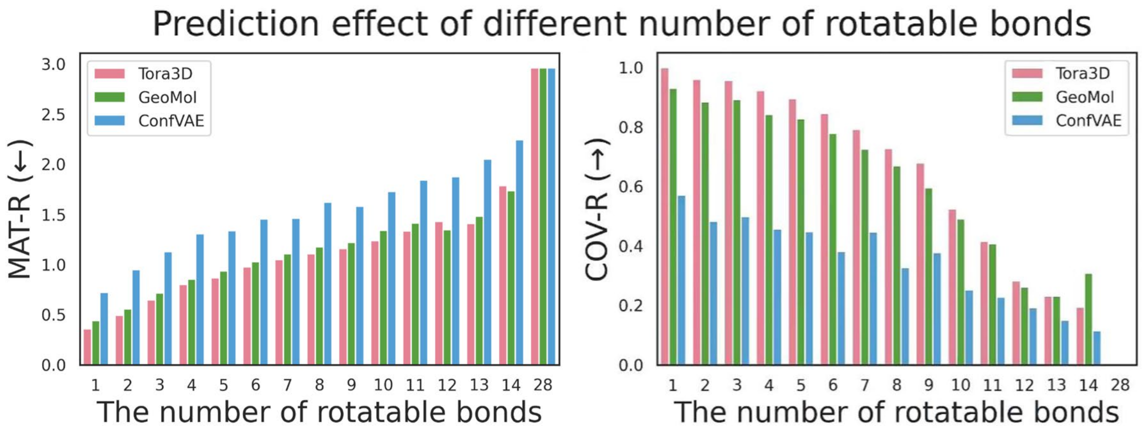


Fig. 4 Prediction performance of the model for different numbers of rotatable bonds. The x-axis indicates the number of rotatable bonds, and the y-axis indicates the prediction performance

showed relatively good performance and acceptable speed (Table 2). As shown in Fig. 4, for each of the models, essentially, the more rotatable bonds a molecule has, the more difficult it is to predict its conformations. More importantly, the performance of Tora3D consistently surpasses ConfVAE and GeoMol when the molecule has less than 10 rotatable bonds (Table 4), which is also an important criterion for the drug-likeness of a molecule [29].

Conformational validity

Given that some inherent defects in typical conformation prediction models would cause conformational invalidity and undermine model performance, we have introduced some strategies into Tora3D to improve its validity.

The first and most important strategy of Tora3D is its autoregressive algorithm. Tora3D predicts the torsion angles of rotatable bonds in a molecule one by one to avoid arising conflicts among local structures, and thus the current torsional angle value is determined not only by the molecular graph but also by the previously predicted torsion angle values. Most deep learning-based conformation generation models do not consider the dependencies among the local structures of a molecule and predict each dihedral angle as an independent variable, which will inevitably lead to invalidity of the overall molecular conformation. For example, topologically distant fragments of a molecule may conflict with each other in space. Figure 5 shows examples of the torsion angle predictions of Tora3D, GeoMol and ConfVAE. With the autoregressive algorithm, Tora3D could consider every rotatable bond sequentially to avoid clashes among local structures, but GeoMol and ConfVAE can not explicitly capture the global interactions as the torsion angles are predicted independently.

At the same time, the autoregressive approach could further ensure the spatial rationality of the whole molecule by attention mechanism. As shown in Fig. 5, the 1th bond of the first molecule shows higher attention with respect to the more distant 5th and 4th bonds but lower attention to the closer 2nd bond. The attention scores are consistent with the observation that the incorrect rotations of the 5th and 4th bond would cause the spatial conflict between the trifluoromethyl and 1-methylimidazole, and thus the 1st bond have a stronger relation to the 5th and 4th bond torsion angle than the closer 2nd bonds. In the second molecule, the 1st bond shows higher attention to the 2nd bond, whose improper rotation would cause serious spatial conflict between the terminal structures in the molecule.

Table 4 Performance comparison of models on the GEOM-drugs dataset (Test set II)

Models	COV-R(↑)	MAT-R(↓)	COV-P(↑)	MAT-P(↓)
ConfVAE	40.06	1.3771	-	-
GeoMol	72.50	1.1000	61.15	1.2009
Tora3D	**81.92**	**0.9297**	**62.16**	**1.1600**
ConfVAE (nRotb ≤ 10)	42.43	1.3296	–	–
GeoMol (nRotb ≤ 10)	76.36	0.9380	57.29	1.1611
Tora3D (nRotb ≤ 10)	**83.03**	**0.8704**	**63.81**	**1.0906**

The other strategy is to incorporate prior knowledge of local 3D structures of each non-terminal atom, to ensure the validity of conformational generation. The main challenge in molecular conformation generation comes from the enormous size of the 3D structure space consisting of bond lengths, bond angles, and torsion angles. However, the molecular graph imposes specific constraints on the set of possible stable local structures, which can be predicted by fast cheminformatics methods. Thus, Tora3D incorporates the prior knowledge about bond lengths and angles to guarantee validity by assembling fixed local structures directly from an initial conformation.

Using fixed local structures can avoid the prediction error for symmetric graph nodes (i.e., nodes with the same topology in graph) and ring structures that seriously undermines the accuracy of many 3D prediction model. For example, GeoMol and ConfVAE generate invalid conformation of non-planar rings, such as hexahydropyridine shown in Fig. 6. GeoMol explicitly models and predicts bond angles and length, but the accumulated errors cause flattened or severely distorted ring; the distance matrix used in ConfVAE is difficult to enforce geometric graph constraint and inevitably lead to seriously implausible structures. Whereas Tora3D could correctly maintain the chair conformation that conforms to the chemical rules by assembling from the initial conformation. Another case is predicting pairs of atoms that are completely structurally symmetrical in a molecule. The classical message-passing neural networks (MPNNs) will embed symmetric graph nodes to the same point in the embedding space and thus generate identical coordinates for them. Previous works often add noise, augment atom features or design complex loss functions to avoid the overlapping of symmetric graph nodes [30]. In the case of the 1,2,3-trimethoxybenzene group (Fig. 6), though appending initial random noise feature vectors does avoid the overlapping of the symmetrical benzene ring and the methoxyl groups in GeoMol and ConfVAE, as Ganea et al. stated that symmetric graph nodes that are less than 3 hops away are indistinguishable by MPNNs in general, the matching information between the ring

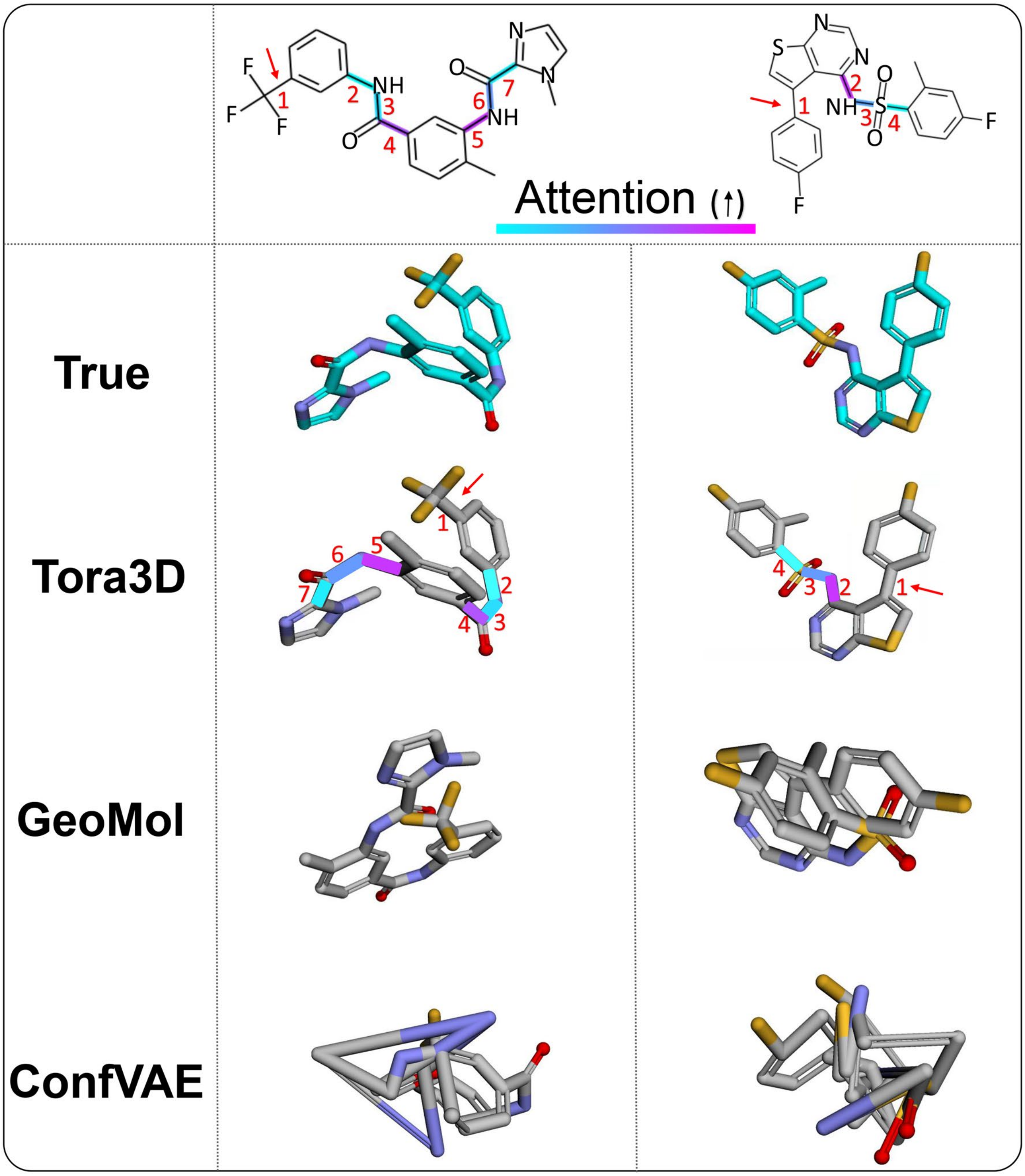


Fig. 5 Comparisons of the conformations predicted by Tora3D, GeoMol, and ConfVAE. The 2D molecule graphs in the first row are marked by the score of attention paid by one of the torsion angles to other bonds (i.e., the bond pointed by the red arrow). Their attention scores toward other torsion angles are indicated by the highlighting. Bonds colored magenta refer to high attention and cyan refer to low attention

plane of benzene with the ground true conformation nodes is lost.

Hence, as the accuracy of local structures significantly impacts many models' performance, Tora3D reconstructing the conformation by a two-stage generation procedure that utilizes predicted torsion angles to assemble fixed local structures can be of practical value. Even if a conformation generated by Tora3D is not in the provided conformation set, it still conforms to the chemical rules and is thus valid and usable.

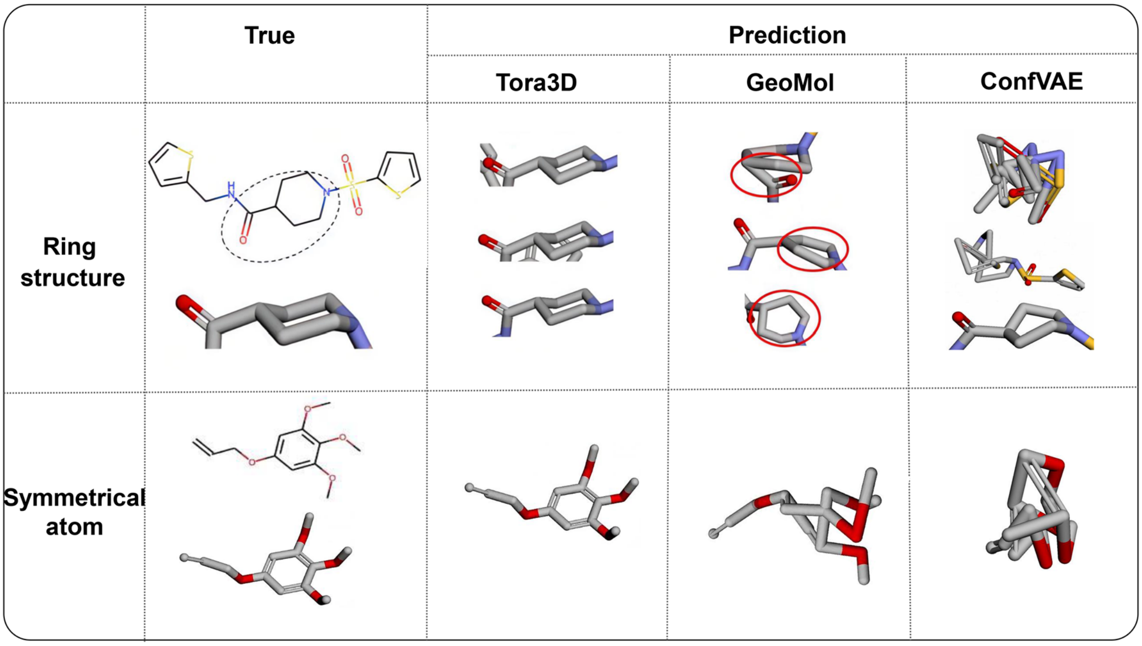


Fig. 6 Comparison of conformations predicted by different models. The first column shows the real conformation, and the other columns show the conformations predicted by Tora3D, GeoMol, and ConfVAE, respectively. The first row shows their differences in the predicted structure (hexahydropyridine), and the second row shows their differences in the predicted symmetrical structure (1,2,3-trimethoxybenzene)

Energy-guided conformational generation

Common methods of conformational sampling in machine learning-based models are random initialization and RDKit initialization. An RDKit initialization can achieve better accuracy by providing a more accurate starting point, while a random initialization can achieve better coverage by the sufficient sampling of the space [31]. To promote diverse conformation ensembles with both good coverage and accuracy, a mixture of random initialization and energy-specific input is used in Tora3D for the conformational generation process. In addition to Gaussian noise that is added to T^0 that allows the model generates multiple conformations, Tora3D can generate a set of conformations with geometrical diversity by varying relative energies as model input. As shown in Fig. 7, the Tora3D predictions of conformations with various relative energies could reproduce the ground true conformations of depicted molecules. The high structural quality, as well as the competitive COV score achieved by Tora3D, suggest that relative energies can be used to guide the generation of a diverse collection of conformations.

Conclusion

Due to the extension of the application scope of molecular 3D structure in the field of drug development, the methodology of molecular conformation generation continues to develop. Here, combining systematic search methods and advanced deep learning models, we propose a deep learning-based model to predict the torsion angles of rotatable bonds in a molecule, thereby predicting molecular conformations. Tora3D is superior to a series of baseline models with comparatively high accuracy but does not sacrifice efficiency. In the aspect of conformational validity, Tora3D employs an autoregressive approach to predict all torsion angles, so that the problem of the collision between local structures can also be solved in an interpretable way. The autoregressive algorithm could consider every rotatable bond sequentially to avoid clashes among local structures, and further improve the spatial rationality of the whole molecule by attention mechanism. At the same time, reconstructing the conformation by a two-stage generation procedure avoids many invalid local structures. In the aspect of conformational diversity, by varying relative energies as model input, Tora3D can generate energy-specific conformation ensemble with good coverage. In addition, as an improvement in model structure to promote accuracy, we proposed a new method of position encoding on graphs that compensates for the difficulties of traditional GNN long-distance messaging. The ablation test of the position vector verified that Tora3D outperformed traditional GNN to solve the problem of long-distance information passing.

Tora3D is a promising tool to generate valid and diverse molecular conformation sets with competitive

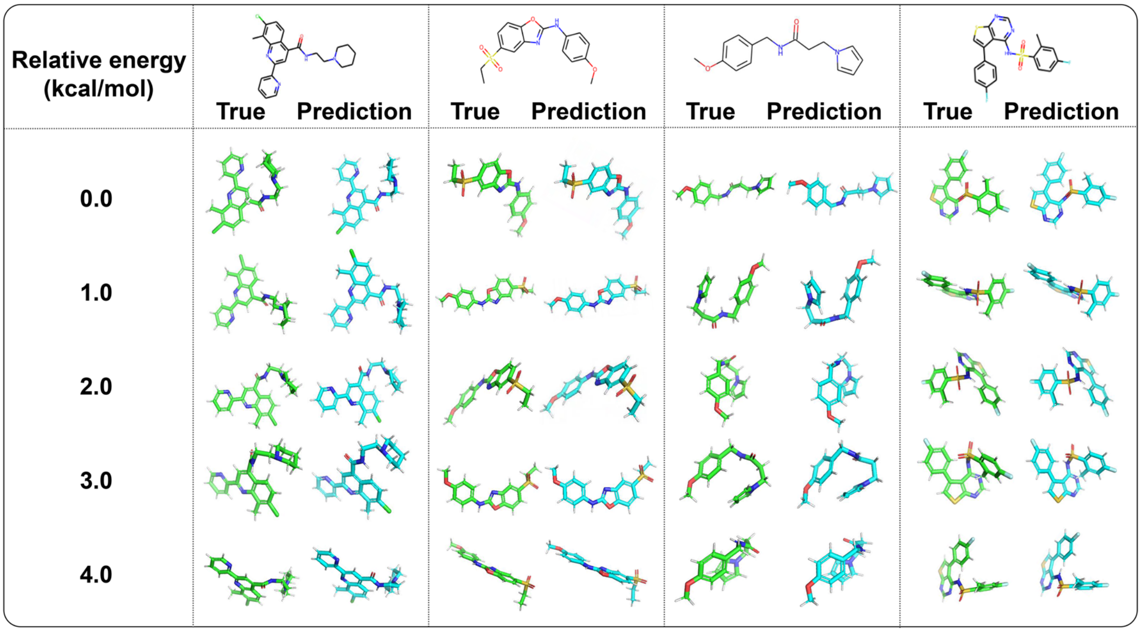


Fig. 7 The predicted results of Tora3D. The relative energy is the absolute energy of a conformation minus the absolute energy of the lowest-energy conformation. And 0 kcal/mol indicates the lowest-energy conformation. The true conformations (green) are on the left and the predicted conformations (blue) are on the right of each column (image source: Pymol [32])

accuracy and efficiency. Its performance is particularly high for drug-like molecules with rotatable bonds less than 10. Especially, energy-guided conformational generation provides many possibilities for model application in the field of drug design, as conformational energy is crucial to understand how a molecule binds to a specific target protein. In future work, we will do more rigorous explorations and we expect that Tora3D will be applied in a variety of downstream tasks including large-scale virtual screening, molecular property prediction, and drug-target interaction prediction, thus speeding up areas of drug discovery.

Abbreviations

3D	Three-dimensional
DFT	Density functional theory
QSAR	Quantitative structure–activity relationships
MC	Monte Carlo simulation
DG	Distance geometry
GA	Genetic Algorithm
WL	Weisfeiler-Lehman
GNN	Graph neural network
MPNN	Message-passing neural network
nRotb	Number of rotatable bonds

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-023-00726-8.

Additional file 1: **Figure S1.** The distribution of the number of conformations.**Figure S2.** The true and predicted conformations for chiral-molecule, spirans and macrocycles. **Figure S3.** The comparison of thetorsion angles predicted by Tora3D with the corresponding statisticaldistribution provided by Torsion Library. **Figure S4.** The conformationcomparison between Tora3D and Conformator. **Table S1.** Hyperparameters.**Table S2.** The prediction performance for chiral molecules, spiransand macrocycles. **Table S3.** Performance comparison with TorsionNET onthe GEOM-drugs dataset (Test set II). **Table S4.** Performance comparisonwith Conformator's initial conformations on the GEOM-drugs dataset (Testset II). **Implementation of Torsion Library. Implementation of TorsionNet.The comparison between the initial conformations and Tora3D's generatedconformations. Loss** Tora3D's Loss.

Author contributions

MYZ and XTL conceived the project. ZMZ implemented the Tora3D model and conducted the computational analysis. GW, RL, LN, KYC, RZZ, QR, XTK, XJL and SKN collected and analyzed the data. XCT, LL, and DYW wrote the paper. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Project supported by National Natural Science Foundation of China (T2225002 and 82273855 to MYZ, 82204278 to XTL, 91953203 to XJL), Lingang Laboratory (LG202102-01-02 to MYZ), the National Key Research and Development Program of China (2022YFC3400504 to MYZ), China Postdoctoral Science Foundation (2022M720153 to XTL), and Shanghai Sailing Program (22YF1460800 to DYW).

Availability of data and materials

The supplementary materials provide the distribution of data sets, some hyperparameters of the model, the settings of loss functions and the implementation and results of some model comparisons. The source code and related datasets are provided for academic use: https://github.com/zimeizhng/Tora3D.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Division of Life Science and Medicine, University of Science and Technology of China, Hefei 230026, Anhui, China. [2]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China. [3]University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China. [4]Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China. [5]School of Pharmacy, China Pharmaceutical University, 639 Longmian Road, Nanjing 211198, China. [6]Lingang Laboratory, Shanghai 200031, China. [7]Precision Pharmacy & Drug Development Center, Department of Pharmacy, Tangdu Hospital, Fourth Military Medical University, Xi'an 710038, China.

Received: 12 January 2023 Accepted: 20 May 2023
Published online: 07 June 2023

References
1. Axelrod S, Gómez-Bombarelli R (2022) GEOM, energy-annotated molecular conformations for property prediction and molecular generation. Sci Data 9:185
2. Friedrich N-O, Flachsenberg F, Meyder A, Sommer K, Kirchmair J, Rarey M (2019) Conformator: A novel method for the generation of conformer ensembles. J Chem Inf Model 59:731–742
3. Hawkins PCD, Nicholls A (2012) Conformer generation with OMEGA: learning from the data set and the analysis of failures. J Chem Inf Model 52:2919–2936
4. Poli G, Seidel T, Langer T (2018) Conformational sampling of small molecules with iCon: performance assessment in comparison with OMEGA. Front Chem 6:229
5. Roy K, Kar S, Das RN (2015) Chapter 10—Other Related Techniques. Academic Press, Boston
6. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, Zhang L, Ke G (2022) Uni-Mol: a universal 3D molecular representation learning framework. ChemRxiv.
7. Mansimov E, Mahmood O, Kang S, Cho K (2019) Molecular geometry prediction using a deep generative graph neural network. Sci Rep 9:20381
8. Parr RG, Weitao Y (1995) Density-functional theory of atoms and molecules. Oxford University Press, Oxford
9. Hu W, Fey M, Ren H, Nakata M, Dong Y, Leskovec J (2021) OGB-LSC: a large-scale challenge for machine learning on graphs. arXiv. https://doi.org/10.48550/arXiv.2103.09430
10. Rappe AK, Casewit CJ, Colwell KS, Goddard WA III, Skiff WM (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 114:10024–10035
11. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parametrization, and performance of MMFF94. J Comput Chem 17:490–519
12. Kanal IY, Keith JA, Hutchison GR (2018) A sobering assessment of small-molecule force field methods for low energy conformer predictions. Int J Quantum Chem 118:e25512
13. Deng Q, Han Y, Lai L, Xu X (1991) Application of monte-carlo simulated annealing on conformational analysis. Chin Chem Lett 2:809–812
14. Spellmeyer DC, Wong AK, Bower MJ, Blaney JM (1997) Conformational analysis using distance geometry methods. J Mol Graphics Modell 15:18–36
15. Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, Cambridge
16. Simm G, Hernandez-Lobato JM (2020) A Generative Model for Molecular Distance Geometry. In: Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, 8949–8958
17. Xu M, Luo S, Bengio Y, Peng J, Tang J (2021) Learning neural generative dynamics for molecular conformation generation. arXiv. https://doi.org/10.48550/arXiv.2102.10240
18. Xu M, Wang W, Luo S, Shi C, Bengio Y, Gomez-Bombarelli R, Tang J (2021) an end-to-end framework for molecular conformation generation via bilevel programming. In: Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, 11537–11547 2021.
19. Ganea O, Pattanaik L, Coley C, Barzilay R, Jensen K, Green W, Jaakkola T (2021) Geomol: Torsional geometric generation of molecular 3d conformer ensembles. Adv Neural Inform Proc Syst NeurIPS 34:13757–13769
20. Shi C, Luo S, Xu M, Tang J (2021) Learning gradient fields for molecular conformation generation. In: Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, 9558–9568
21. Xu M, Yu L, Song Y, Shi C, Ermon S, Tang J (2022) Geodiff: a geometric diffusion model for molecular conformation generation. arXiv. https://doi.org/10.48550/arXiv.2203.02923
22. Riniker S, Landrum GA (2015) Better informed distance geometry: using what we know to improve conformation generation. J Chem Inf Model 55:2562–2574
23. Hawkins PCD (2017) Conformation generation: the state of the art. J Chem Inf Model 57:1747–1756
24. Penner P, Guba W, Schmidt R, Meyder A, Stahl M, Rarey M (2022) The torsion library: Semiautomated improvement of torsion rules with SMARTScompare. J Chem Inf Model 62:1644–1653
25. Rai BK, Sresht V, Yang Q, Unwalla R, Tu M, Mathiowetz AM, Bakken GA (2022) TorsionNet: a deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. J Chem Inf Model 62:785–800
26. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H et al (2020) Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 63:8749–8760
27. Niepert M, Ahmed M, Kutzkov K (2016) Learning convolutional neural networks for graphs. In: Proceedings of the 33rd international conference on machine learning, proceedings of machine learning research. 2014–2023
28. RDKit: Open-source cheminformatics. https://www.rdkit.org.
29. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1:337–341
30. You J, Ying R, Leskovec J (2019) Position-aware graph neural networks. In: Proceedings of the 36th international conference on machine learning, proceedings of machine learning research, 7134–7143
31. Guan J, Qian WW, Ma W-Y, Ma J, Peng J (2021) Energy-inspired molecular conformation optimization. In: international conference on learning representations
32. Schrödinger L, DeLano W (2020) PyMOL. https://www.pymol.org/pymol

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.