Briefings in Bioinformatics, 2025, **26(1)**, bbae676
https://doi.org/10.1093/bib/bbae676
Problem Solving Protocol

GADIFF: a transferable graph attention diffusion model for generating molecular conformations

Donghan Wang, Xu Dong, Xueyou Zhang, LiHong Hu*

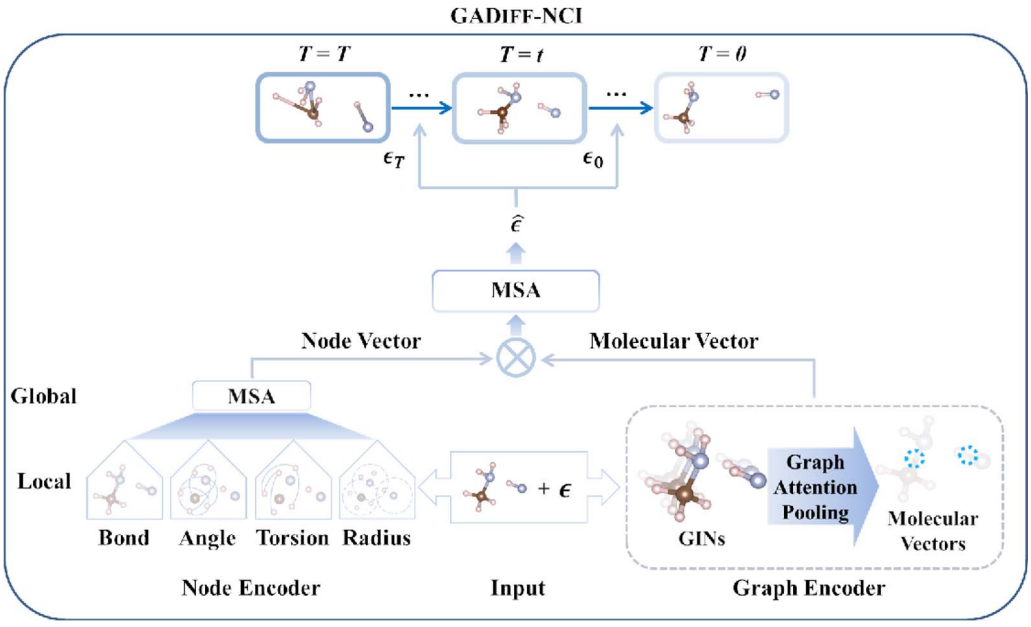
School of Information Science and Technology, Northeast Normal University, 130117 Changchun, China

*Corresponding author. School of Information Science and Technology, Northeast Normal University, 130117 Changchun, China. E-mail: lhhu@nenu.edu.cn

Abstract

The diffusion generative model has achieved remarkable performance across various research fields. In this study, we propose a transferable graph attention diffusion model, GADIFF, for a molecular conformation generation task. With adopting multiple equivariant networks in the Markov chain, GADIFF adds GIN (Graph Isomorphism Network) to acquire local information of subgraphs with different edge types (atomic bonds, bond angle interactions, torsion angle interactions, long-range interactions) and applies MSA (Multi-head Self-attention) as noise attention mechanism to capture global molecular information, which improves the representative of features. In addition, we utilize MSA to calculate dynamic noise weights to boost molecular conformation noise prediction. Upon the improvements, GADIFF achieves competitive performance compared with recently reported state-of-the-art models in terms of generation diversity(COV-R, COV-P), accuracy (MAT-R, MAT-P), and property prediction for GEOM-QM9 and GEOM-Drugs datasets. In particular, on the GEOM-Drugs dataset, the average COV-R is improved by 3.75% compared with the best baseline model at a threshold (1.25 Å). Furthermore, a transfer model named GADIFF-NCI based on GADIFF is developed to generate conformations for noncovalent interaction (NCI) molecular systems. It takes GADIFF with GEOM-QM9 dataset as a pre-trained model, and incorporates a graph encoder for learning molecular vectors at the NCI molecular level. The resulting NCI molecular conformations are reasonable, as assessed by the evaluation of conformation and property predictions. This suggests that the proposed transferable model may hold noteworthy value for the study of multi-molecular conformations. The code and data of GADIFF is freely downloaded from https://github.com/WangDHg/GADIFF.

Graphical Abstract



Received: June 26, 2024. **Revised:** November 4, 2024. **Accepted:** December 15, 2024

Keywords: diffusion generation model; transfer learning; graph neural network; noncovalent interaction; molecular conformation generation; molecular property prediction

Introduction

Molecular Conformation Generation (MCG) plays a crucial role in drug discovery and is an important part of lead drug discovery field [1]. It is designed to generate the low-energy spatial structure of molecules [2]. Moreover, MCG is closely related to many research issues, such as molecular property prediction [3], molecular docking [4], and pharmacophore modeling [5]. The molecular conformation varies with the environment and is highly flexible [2]. The purpose of MCG is to generate 3D coordinates of all the atoms in a molecule and apply the resulting conformation to research fields or specific tasks [6]. However, both chemical experiments (e.g. X-ray crystallography [7]) and theoretical computational methods (e.g. molecular dynamics [8]) are costly or unable to generate plenty of high quality conformations. Therefore, obtaining rational flexible molecular conformations in an efficient way is a significant task yet challenging.

The rapid development of deep learning has driven improvements in MCG models [6, 9–17]. MCG problem can be solved not only by connecting molecular fragments using a set of rules, but also by features learned from large molecular databases [18]. A large number of studies have shown the great potential of deep learning on MCG. Mansimov *et al*. [9] proposed the conditional deep generate graph neural network (CVGAE) that could directly learn to generate energetically favorable molecular conformations. Later, Simm & Hernandez Lobato [10] proposed GRAPHDG to obtain a generation model with roto-translation invariance. Shi *et al*. [13] suggested CONFGF that is able to generate molecular conformations directly, which greatly reduces the cumulative error during model generation. Ganea *et al*. [14] devised a torsional geometric generation approach and came up with new criterion metrics to evaluate the quality of molecular conformations. Xu *et al*. [16] theoretically demonstrated that Markov chains evolving with equivariant Markov kernels can induce an invariant distribution by design under the diffusion framework, and they developed an end-to-end model GEODIFF.

The diffusion model is one of the most powerful generation models. It has achieved remarkable results in many fields such as computer vision, natural language processing and bioinformatics [19]. And many diffusion models also have been applied to solve various cheminformatics problems, such as molecular generation [16] and protein design [20]. These models generally outperform previous generation models such as variational autoencoder (VAE) and generative adversarial network (GAN), showing that diffusion model has great potential in cheminformatics.

Inspired by the work in the reference [16], we propose a graph attention diffusion generation model, GADIFF, by adding noise attention and multi-scale representation for MCG tasks. GADIFF is designed under roto-translational invariant diffusion framework, where molecules can obtain the conformations with invariant likelihood from the Markov chains evolving with equivariant Markov kernels. Besides that, we use E(n) Equivariant Graph Neural Networks (EGNNs) to maintain the equivariant molecular property of GNN model in Markov kernels. The contributions of this research are described as follows:

- Data representation is critical for the performance of diffusion generation model. In order to improve node representation, GADIFF utilizes Graph Isomorphism Network (GIN) to extract local node features by aggregating heterogeneous graphs with different types of edges, atomic bonds, bond angle interactions, torsion angle interactions and long-range interactions. On the other hand, GADIFF computes node features in different surroundings (from the view of various subgraphs related to one node), and then employs Multi-head Self-attention (MSA) mechanism to fuse the node features globally. Therefore, the equivariant node encoder of ϵ_θ (the noise model) consisting of the above two modules can learn both local and global information of all the nodes, which thus improves the node representation.
- Noise calculation is central to the operation of diffusion models. In order to obtain accurate conformation noise, GADIFF implements MSA mechanism to dynamically compute the attention scores for conformation noises from different subgraphs. In GADIFF, the subgraph conformation noise at different spatial scales is weighted by the MSA mechanism to obtain the molecular conformation noise, which automatically learns the weights of different subgraphs on the molecular conformation noise. The module in ϵ_θ showed in Fig. 1(b) is taken as noise encoder to compute the conformation noise.
- Noncovalent interaction (NCI) conformations are remarkably significant for drug discovery; therefore, we further proposed the transferring model GADIFF-NCI for NCI molecule generation. Modeling of long-range interaction features enables GADIFF to fulfill a MCG task of multi-molecular systems. GADIFF-NCI takes GADIFF as the pre-trained model, from which the node encoder parameters on GEOM-QM9 are migrated to GADIFF-NCI, followed by fine-tuning on NCI dataset. For better adaptation to NCI dataset, we use a graph encoder to obtain graph-level features via GIN network and graph attention pooling module, which ensures information completeness of node features. From experiments, the well generated NCI conformations demonstrate that the strategy of GADIFF-NCI can be applied to NCI conformation generation, which provides a new strategy for the similar molecular generation tasks.

Diffusion models

Generation methods, which aim to model the latent probability distribution of a dataset, are being rapidly developed and widely used. There are various deep generation models emerged, such as VAEs, GANs, flow-based models, autoregressive models, and diffusion models. Currently, diffusion model is one of the most powerful deep generation models with the advantages of being able to learn of complex data distributions and generate diversified data. It can achieve remarkable results in classical fields (e.g. computer vision, natural language processing) and also has been successful in interdisciplinary domains such as computational chemistry, bioinformatics, and medical image reconstruction [19, 20].

In the diffusion framework, the molecular conformation C_t is modeled as a thermodynamic system that changes over time. The C_0 is the initial conformation from the data distribution $q(C_0|G)$. In

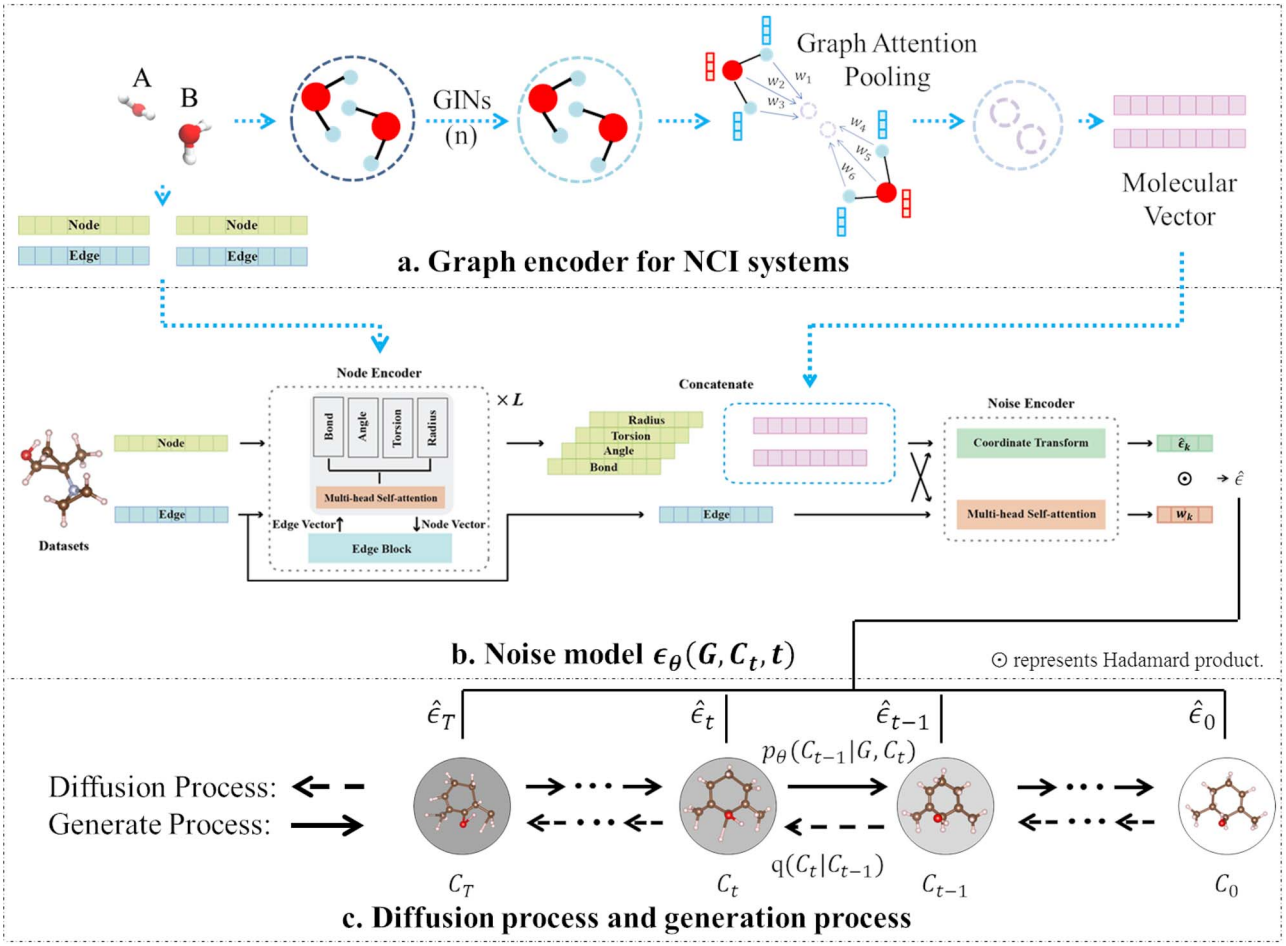


Fig. 1. The illustrations for GADIFF and GADIFF-NCI framework. First, a is the graph encoder of GADIFF-NCI, and the output is the molecule-level feature vector of NCI data. Next, b is $\epsilon_\theta(G, C_t, t)$ of models (GADIFF, GADIFF-NCI), and the node encoder of $\epsilon_\theta(G, C_t, t)$ in GADIFF-NCI is used for fine-tuning based on pre-trained model. Finally, c is the diagram of the forward diffusion process and the reverse generation process of models (GADIFF, GADIFF-NCI). In short, the GADIFF consists of b, c and the GADIFF-NCI consists of a, b, c.

the diffusion process, the white noise C_T of the graph G is obtained by adding noise into the C_0 gradually in T time steps. And the initial conformation C_0 is obtained through gradual noise reduction of C_T in the generation process that is the reverse process of the diffusion process. Figure 1c displays the diffusion process and the generation process of our diffusion model. The $q(C_t|C_{t-1})$ is the fixed posterior distribution in the diffusion process, $p_\theta(C_{t-1}|G, C_t)$ is the Markov kernel in the generation process and ϵ_t represents the noise in each time step $t(t \in [1, T])$. The main objective of diffusion models is to apply the noise model ϵ_θ for learning a data distribution $p_\theta(C_0|G)$ that approximates the real data distribution $q(C_0|G)$ and samples new conformations from data distribution $p_\theta(C_0|G)$.

The reverse generation process can be formulated as a conditional Markov chain with learnable transition kernel, as shown in Equation 1.

$$p_\theta(C_{0:T-1}|G, C_T) = \prod_{t=1}^{T} p_\theta(C_{t-1}|G, C_t),$$
$$p_\theta(C_{t-1}|G, C_t) = N(C_{t-1}; \mu_\theta(G, C_t, t), \sigma_t^2). \quad (1)$$

The μ_θ is the mean parameterized by $\mu_\theta(G, C_t, t) = (C_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(G, C_t, t))/\sqrt{\alpha_t}$, where $\epsilon_\theta(G, C_t, t)$ is an equivariant noise model to learn the noise $\hat{\epsilon}$ with parameters θ. And σ_t^2 is a user-defined variance. The forward diffusion process is defined as a fixed posterior distribution $q(C_{1:T}|C_0)$, as shown in Equation 2.

$$q(C_{1:T}|C_0) = \prod_{t=1}^{T} q(C_t|C_{t-1}),$$
$$q(C_t|C_{t-1}) = N(C_t; \sqrt{1-\beta_t}C_{t-1}, \beta_t I). \quad (2)$$

The $q(C_t|C_{t-1})$ is the transition kernel of Markov chain of the diffusion process at step t, β_t is the fixed variance schedule at step t, and I is the identity matrix. The $q(C_t|C_0)$ is derived as Equation 3 from $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, and Equation 2.

$$q(C_t|C_0) = N(C_t; \sqrt{\bar{\alpha}_t}C_0, (1-\bar{\alpha}_t)I). \quad (3)$$

With t sufficiently large, the forward process will convert conformation C_0 to a whitened isotropic Gaussian distribution conformation C_T, so $p(C_T)$ is set to a standard Gaussian distribution.

Instead of directly optimizing the exact logarithm likelihood, diffusion methods design the training objective according to maximizing the evidence lower bound (ELBO). Ho *et al.* proposed a simplified objective that is a weighted variational bound [21]. Xu *et al.* designed an invariant ELBO based on the simplified objective,

as shown in Equation 4 [16].

$$L_{ELBO}(\theta) = \sum_{t=1}^{T} E_{\{C_0,G\}\sim q(C_0,G),\epsilon\sim N(0,1)}[||\epsilon - \epsilon_\theta(G, C_t, t)||^2]. \quad (4)$$

The corrected noise vector θ which is equivariant to C_t can be calculated by $\frac{\partial_{C^t} d^t \cdot (d^t - \sqrt{\bar{\alpha}_t} d^0)}{\sqrt{1-\bar{\alpha}_t}}$. And the C_t in Equation 4 can be calculated by $C_t = \sqrt{\bar{\alpha}_t} C_0 + \sqrt{1-\bar{\alpha}_t}\theta$ according to Equation 3.

Some other related works are introduced in supplementary information (SI).

Problem notations

In this research, $G = \langle V, E\rangle$ represents a molecular graph and its related information. The V is the set of atoms in a molecule, $v_i (v_i \in V, 1 \leqslant i \leqslant |V|)$ represents the atomic attributes of the i^{th} atom, such as the elements type, and $|V|$ represents the number of atoms contained in the molecular graph, the embedding for v_i uses h_i; while E is the set of interatomic edges in a molecule, $e_{ij} (e_{ij} \in E, 1 \leqslant i, j \leqslant |E|, i \neq j)$ denotes the embedding of interatomic edge attributes between i^{th} atom and j^{th} atom in the molecular graph.

To obtain more abundant node information, we extend the graph G (a traditional molecular bond graph) to four types of graphs in terms of edge types (atomic bonds, bond angle interactions, torsion angle interactions, and long-range interactions). Therefore, the graph with one edge type is named as a subgraph of G in this paper. The subgraphs are defined on the G with the same atom set V and different edge set E_k{k = (bond, angle, torsion, radius)}. The E_k is the set of e_{ij}^k{k = (bond, angle, torsion, radius)} and the detailed instructions of subgraphs are displayed in Fig. S1 of SI.

Each atom in the molecular graph G can get the coordinates during the sampling process. The conformation set $C(C = [c_1, c_2, \cdots, c_n])$ is the target of the sampling process, where $c_i (c_i \in R^{|V|\times 3}, 1 \leqslant i \leqslant |n|)$ denotes the i^{th} molecular conformation in set C and n means the number of conformations generated from each molecular graph. We describe the importance of equivariance for molecular modeling and its application to GADIFF in SI.

Graph attention diffusion model

Molecules switch between their different conformations and end up in different local minima (stable conformation) because of environmental conditions. Thus, there are many possible rational conformations for each molecular graph, and even these conformations exhibit completely different properties. In this study, GADIFF is proposed to generate multiple possible conformations through molecular graphs and is also generalized to a multi-molecular generation task.

Figure 1 presents the designed framework of the proposed models GADIFF (b, c), GADIFF-NCI(a, b, c), and the diffusion theory (c). The framework (GADIFF, GADIFF-NCI) is optimized by minimizing the difference between the reference conformation noise ϵ and the predicted molecular conformation noise $\hat{\epsilon}$. Then, $\epsilon_\theta(G, C_t, t)$ predicts conformation noise $\hat{\epsilon}_t$ for C_t. When extracting features, GADIFF employs MSA mechanism in the node encoder of $\epsilon_\theta(G, C_t, t)$ to integrate the information from subgraphs. The subgraph conformation noise $\{\hat{\epsilon}_k\}_{k=(bond,angle,torsion,radius)}$, is acquired by jointly training on edge and node features in the molecular graph by the EGNN framework described in SI. And the molecular conformation noise $\hat{\epsilon}$ is obtained by weighting the subgraph conformation noise $\hat{\epsilon}_k$ using MSA mechanism.

Attention graph convolution layer

In GADIFF, we improve the Equivariant Graph Convolutional Layers (EGCL) in EGNN based on MSA mechanism for the tasks with datasets like molecules with multiple edge types. The node encoder of $\epsilon_\theta(G, C_t, t)$ in GADIFF is attention graph convolution layers (AGCLs). In AGCL, the node vectors of various subgraphs can obtain local and global information through GIN network and the MSA mechanism, respectively. The subgraph node vectors $(\{h_{i,k}^l\}_{k=(bond,angle,torsion,radius)})$ and edge embedding e_{ij} are initialized using the Pytorch package's Embedding layer. The l^{th} AGCL of the node encoder is shown in the following Equations 5–8. The diagram of model structure has been shown in Fig. 1(b).

$$h_i^l = \chi(h_{i,bond}^l, h_{i,angle}^l, h_{i,torsion}^l, h_{i,radius}^l) \quad (5)$$

$$m_{ij}^l = \phi_e(h_i^l, h_j^l, e_{ij}) \quad (6)$$

Equations 5 and 6 denote the edge block in AGCL. In Equation 5, χ is a convolution block to obtain node vector h_i^l from node vectors $\{h_{i,k}^l\}_{k=(bond,angle,torsion,radius)}$ of a subgraph. In Equation 6, m_{ij}^l is the edge embedding in each layer and ϕ_e is an MLP network for learning edge vector. Equations 5 and 6 constitute the edge block of AGCL.

$$\tilde{h}_{i,k}^l = \psi_n(h_{i,k}^l, m_{ij,k}^l), k = (bond, angle, torsion, radius) \quad (7)$$

$$h_{i,bond}^{l+1}, h_{i,angle}^{l+1}, h_{i,torsion}^{l+1}, h_{i,radius}^{l+1} = \varphi_n(\tilde{h}_{i,bond}^l, \tilde{h}_{angle}^l, \tilde{h}_{torsion}^l, \tilde{h}_{radius}^l) \quad (8)$$

In Equation 7, ψ_n denotes the GIN network utilized for calculating local node features from subgraphs. In Equation 7, $\{\tilde{h}_{i,k}^l\}_{k=(bond,angle,torsion,radius)}$ are obtained as the input vectors for Equation 8 to get the output node vectors $\{h_{i,k}^{l+1}\}_{k=(bond,angle,torsion,radius)}$. The φ_n in Equation 8 denotes the MSA mechanism designed for fusing the local node vectors. Equations 7 and 8 carry out the computation in the node block of AGCL.

In GADIFF, MSA strategy is used to integrate local information into global molecular information, as shown in Equation 8. Hence, the node vectors $\{h_{i,k}^{l+1}\}_{k=(bond,angle,torsion,radius)}$ from subgraphs are fused in AGCLs for aggregating local information, which grant GADIFF get better representation for molecular data, thus improve the performance accordingly.

Training objective and sampling

Since the purpose of $\epsilon_\theta(G, C_t, t)$ is to predict the conformation noise rather than the conformation itself, solving node coordinates in EGNN is replaced by calculating the conformation noise. The $\epsilon_\theta(G, C_t, t)$ is composed of a node encoder and a noise encoder. And $\hat{\epsilon}$ could be predicted by the weighted sum over different subgraph conformation noises $\hat{\epsilon}_k$ as shown in Equation 10.

$$w_{bond}, w_{angle}, w_{torsion}, w_{radius} = \varphi_\epsilon(h_{bond}, h_{angle}, h_{torsion}, h_{radius}, e_{ij}) \quad (9)$$

$$\hat{\epsilon} = \sum_{k=(bond,angle,torsion,radius)} w_k \sum_{j\in N(i)} \frac{1}{d_{ij}}(x_i - x_j)\phi_x(e_{ij}^k) \quad (10)$$

In Equation 9, φ_ϵ is an MSA mechanism for computing the weights of conformation noise from subgraphs. Equation 10 gets the conformation noise $\hat{\epsilon}$ adaptively by summing over

Algorithm 1 GADIFF Training.

repeat
  $C_0 \sim q(C_0|G), t \sim Uniform(\{1, \cdots, T\}), \epsilon \sim N(0, I)$
  $C_t \sim q(C_t|C_0)$ # Eq. 3
  Initial atom node vector $h_i^0 = v_i$ and edge vector e_{ij} from C_t.
  for $l = 0, \cdots, L-1, L$ **do**
    $h_i^{l+1} = AGCL_\theta(G, e_{ij}, h_i^l)$; # node_encoder, Eq. 5 ∼ Eq. 8
  end for
  $\hat{\epsilon} = noise\_encoder_\theta(G, C_t, e_{ij}, h)$ # noise_encoder, Eq. 9 ∼ Eq. 10
  $\nabla||\theta - \hat{\theta}||^2$ for the gradient descent step # training object, Eq. 4
until converged

the weighted subgraph conformation noises. In Equation 10, $\{e_{ij}^k\}_{k=(bond,angle,torsion,radius)}$ denotes the edge vectors of each subgraph, d_{ij} is the distance norm (interatomic distances) between atom i and j, x_i and x_j represent the coordinates of atom i and j in a conformation C_t, and ϕ_x is an MLP network. As mentioned above, Equation 4 defines L_{ELBO}. With $\epsilon_\theta(G, C_t, t)$ obtained in Equation 10, we can learn training objective L_{ELBO}.

In order to enhance prediction for the noise ϵ in C_t, we improve $\epsilon_\theta(G, C_t, t)$ by utilizing MSA mechanism again. The MSA mechanism is used to output the weights w_k for conformation noises $\hat{\epsilon}_k = \sum_{j \in N(i)} \frac{1}{d_{ij}}(x_i - x_j)\phi_x(e_{ij}^k), k = (bond, angle, torsion, radius)$ of subgraphs, as shown in Equations 9 and 10. In contrast to previous models that learn the noise $\hat{\epsilon}$ directly, the weighting strategy adopts adaptive learning the effects of different interactions on the molecular conformation noise $\hat{\epsilon}$. It could assign the subgraph weights of the same node well. Algorithm 1 gives the GADIFF training procedure based on the designed objective.

The sampling process adopted in GADIFF is a Markov chain [16]. It transfers the white noise conformation C_T to the equilibrium state conformation C_0. The C_T is sampled from the normal distribution $p(C_T)$ and progressively denoised to get C_{t-1} from $p_\theta(C_{t-1}|G, C_t), t = T, T-1, \cdots, 1$ according to Equation 1. The proposed models use the same sampling algorithm as GEODIFF to get a rational conformation C_0. The Algorithm S1 in SI describes the sampling process in detail.

Graph attention diffusion model for NCI molecules

NCI conformations are important for the study of drug target binding and the generation of NCI conformations is of research value [22]. However, NCI conformation datasets [23–27] are not as rich as the GEOM datasets [28], even though the amount of NCI datasets is too large to manually identify edge types, therefore we adopt Deep Transfer Learning (DTL) to extend the model on NCI conformation generation. In our model, we transfer the node encoder parameters in the pre-trained GADIFF based on the similar size molecule in the dataset (GEOM-QM9), and fine-tune GADIFF on the collated NCI datasets, thus form a transferring model, GADIFF-NCI, for NCI MCG tasks.

$$\tilde{h}_{i,graph} = \psi_g(v_i, e_{ij}) \tag{11}$$

$$h_{graph} = [\sum_{i \in G_j} (\tilde{h}_{i,graph} \cdot \phi_g(\tilde{h}_{i,graph})/n_j)]_{j \in (0,1)} \tag{12}$$

Because systems in NCI datasets are molecular dimers, GADIFF-NCI employs a graph encoder to obtain the graph-level (molecular) vector for obtaining intact molecular pair information. The graph encoder consists of a GNN network and a graph attention pooling module, and their formula are expressed by Equations 11 and 12, respectively, where ψ_g represents GIN networks, ϕ_g is an MLP network to compute the weights for node vectors $\tilde{h}_{i,graph}, i \in G_j$ in a graph $G_j, j \in (0, 1)$, and n_j is the atom number of molecule graph G_j. The detailed training process for GADIFF-NCI is in Algorithm 2.

Experiments

We build and evaluate GADIFF on two databases, small molecules (GEOM-QM9) and drug-like molecules (GEOM-Drugs), and then transfer the trained GADIFF to NCI datasets. Two common MCG evaluation criteria, conformation generation metrics and property prediction, are adopted for GADIFF. And the evaluations for GADIFF-NCI are in terms of interatomic distance distributions difference and property prediction accuracy. In addition, to validate the improving strategies, ablation experiments are conducted on GADIFF and GADIFF-NCI, respectively. We describe the detailed experimental setting of GADIFF and GADIFF-NCI in SI. And the time complexity analysis of GADIFF is list in Table S1 of SI.

Experiment setup

Datasets

The GEOM dataset (the detail is described in SI) is used in modeling and the data segmentation is consistent with the dataset by Shi *et al.* [13]. The training set consists of 40k randomly selected molecules from the two datasets (GEOM-QM9, GEOM-Drugs), with five low-energy conformations per molecule sorted by energies (200k conformations). From the remaining data, 200 molecules are randomly selected as a test set, and the number of conformations per molecule is limited to between 50 and 500. The number of generated conformations is the same as previous researches (twice the number of reference conformations).

In order to generate NCI molecular conformation, the NCI dataset is used to fine-tune GADIFF-NCI. For clearly demonstrating the model validity, only hydrogen-bond dominated molecules among them are chosen for analyses. The NCI dataset used for the GADIFF-NCI consists of two parts: hydrogen-bonded molecules in the dataset (S22×5 [23], S66×8 [24], X40×10 [25]) collected in the previous work [29] and hydrogen-bonded molecules in the NCIAtlas database (HB300SPX×10 [26], HB375×10 [27]). The first dataset consists of 319 conformations including 40 molecular pairs and their derivatives along the dissociation curve, which are dominated by hydrogen bonding [22, 24, 25, 29]. The second dataset from NCIAtlas is created by Jan Řezáč with the aim of building a wide range of NCI dataset using advanced quantum chemical methods. The samples dominated by hydrogen bonding in NCIAtlas are HB300SPX×10 and HB375×10. They are adopted as the second dataset with 675 molecule pairs and 6750

Algorithm 2 GADIFF-NCI Training.

repeat
  $C_0 \sim q(C_0|G), t \sim Uniform(\{1, \cdots, T\}), \epsilon \sim N(0, I)$
  $C_t \sim q(C_t|C_0)$
  Initial atom node vector $h_i^0 = v_i$ and edge vector e_{ij} from C_t.
  $h_{node} = node\_encoder_\theta(G, C_t, h_i^0, e_{ij})$ # using the pre-trained GADIFF
  $h_{graph} = graph\_encoder_\theta(G, h_i^0, e_{ij})$ # Eq. 11 ∼ Eq. 12
  $h = concatenation(h_{node}, h_{graph})$
  $\hat{\epsilon} = noise\_encoder_\theta(G, C_t, e_{ij}, h)$
  $\nabla||\theta - \hat{\theta}||^2$ for the gradient descent step
until converged

conformations [26, 27]. Thus, in total the NCI dataset used for GADIFF-NCI contains 7069 conformations of 715 molecule pairs. The NCI dataset is randomly divided into a training set (6123 conformations of 619 molecule pairs), a validation set (473 conformations of 48 molecule pairs) and a test set (473 conformations of 48 molecule pairs).

Baselines

We compare GADIFF with the SOTA baseline methods: CVGAE [9], GRAPHDG [10], CGCF [11], CONFVAE [12], GEOMOL [14], CONFGF [13], DGSM [15], GEODIFF [16], DMCG [6], RDKIT+Clustering [17]. GADIFF has also been compared to the currently popular open-source software for conformation generation, RDKIT [30], a classical Euclidean distance geometry [31] approach. The detailed introduction of baselines is in SI.

Results and discussions

Conformation generation

Evaluation metrics. In order to evaluate the quality and diversity of the conformations generated by GADIFF, we use two general evaluation metrics as other baselines, Coverage (COV) and Matching (MAT). COV is the coverage rate of one conformation set covered by the other conformation set at a certain RMSD threshold. The generated conformations with better diversity have a larger COV value. And MAT measures the difference between the generated conformations and the reference conformations. The smaller the MAT value, the higher the accuracy of the generated conformations. The evaluation metrics (COV-R, MAT-R, COV-P, MAT-P) derived by COV and MAT based on the root-mean-square deviation (RMSD) are often used in the field of generation modeling [14–16]. The RMSD is defined as the normalized Frobenius norm of the two atomic coordinates matrices aligned by the Kabsch algorithm to measure the similarity between two conformations [32]. Let Sg and Sr represent the generated and reference conformation set respectively, and δ is the defined RMSD threshold. COV-R, MAT-R, COV-P, MAT-P are defined as follows:

$$COV\text{-}R(S_g, S_r) = \frac{1}{|S_r|}|\{C \in S_r | RMSD(C, \hat{C}) \leqslant \delta, \hat{C} \in S_g\}| \quad (13)$$

$$MAT\text{-}R(S_g, S_r) = \frac{1}{|S_r|}\sum_{C \in S_r} \min_{\hat{C} \in S_g} RMSD(C, \hat{C}) \quad (14)$$

$$COV\text{-}P(S_g, S_r) = \frac{1}{|S_g|}|\{C \in S_g | RMSD(C, \hat{C}) \leqslant \delta, \hat{C} \in S_r\}| \quad (15)$$

$$MAT\text{-}P(S_g, S_r) = \frac{1}{|S_g|}\sum_{C \in S_g} \min_{\hat{C} \in S_r} RMSD(C, \hat{C}) \quad (16)$$

According to the experiment setup, the number of S_g conformations per molecule is twice that of S_r. Higher COV scores as well as lower MAT scores indicate that the model has the ability to generate more novel realistic conformations. From the experience of previous researches [13–16], the threshold δs of the GEOM-QM9 and GEOM-Drugs datasets are set 0.5Å and 1.25Å, respectively.

Model comparison. The evaluation metric (mean and median of COV-R, MAT-R, COV-P, MAT-P) values of the models built on GEOM-QM9 and GEOM-Drugs datasets are shown in Tables 1 and 2, respectively. In Tables 1 and 2, the results of DGSM [15] are from its original paper, and the results of RDKIT, DMCG, and RDKIT+Clustering were reported by the paper of RDKIT+Clustering [17]. The results of the rest competitive baselines are from the results recorded in the paper of GEODIFF [16]. It can be seen that GADIFF outperforms those SOTA baselines on most metrics, especially on the more challenging drug-like molecule dataset, GEOM-Drugs. As shown in Table 1, COV-P and MAT-P of GADIFF in GEOM-QM9 give the best results of all baselines in both mean and median values and its results of COV-R and MAT-R are also competitive. In Table 2, it shows GADIFF achieves better performance in GEOM-Drugs than all other baseline methods. With an increasing number of iterations until convergence, GADIFF achieves an average performance of 95.37% and a median of 100.00% on the COV-R metric, which shows GADIFF outperforms previous baselines across all measured criteria. A comparison with computational chemistry energy-based method, Confab, is shown in Table S2 of SI [33]. The result demonstrates that GADIFF outperforms Confab on most evaluation metrics especially on the GEOM-Drugs. Aforementioned results indicate that GADIFF is capable of generating high-quality and diverse conformations, and modeling diversified molecular distributions well.

The qualitative analysis of generated molecule conformations by GADIFF is shown in Fig. 2. It shows GADIFF generates geometrically diverse conformations that are more similar to reference conformations. And GADIFF outperforms the CONFGF and CONFVAE in all four complex molecules by comparing local structures, which illustrates that GADIFF could perform MCG tasks well. More visualizations of generated drug-like molecule conformations are shown in Fig. S2 of SI. And the detailed discussions of drug-like tautomers generated by GADIFF are illustrated in Fig. S3 of SI.

Property prediction

Setup. Property prediction refers to computing the ensemble properties [28] of generated molecule conformations and comparing them with references, which is a directly assessment of the quality of the generated conformations. Simm and Hernandez-Lobato reported the first work utilizing property prediction to evaluate and analyze the generated molecule conformations [10]. We follow the design scheme of Shi *et al.* [13], where sampling a split containing 30 molecules from the GEOM-QM9 and generating 50

Table 1. Evaluation parameters of models based on the GEOM-QM9 dataset (δ is 0.5 Å)

Models	COV-R(%)↑		MAT-R(Å)↓		COV-P(%)↑		MAT-P(Å)↓	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
RDKIT	83.26	90.78	0.3447	0.2935	–	–	–	–
CVGAE	0.09	0.00	1.6713	1.6088	–	–	–	–
GRAPHDG	73.33	84.21	0.4245	0.3973	43.90	35.33	0.5809	0.5823
CGCF	78.05	82.48	0.4219	0.3900	36.49	33.57	0.6615	0.6427
CONFVAE	77.84	88.20	0.4154	0.3739	38.02	34.67	0.6215	0.6091
GEOMOL	71.26	72.00	0.3731	0.3731	–	–	–	–
CONFGF	88.49	94.31	0.2673	0.2685	46.43	43.41	0.5224	0.5124
DGSM	91.49	95.92	0.2139	0.2137	44.64	43.72	0.5369	0.5023
GEODIFF	90.07	93.39	0.2090	0.1988	52.79	50.29	0.4448	0.4267
DMCG	94.98	98.47	0.2365	0.2312	–	–	–	–
RDKIT+Clustering	**97.65**	**100.00**	**0.1902**	**0.1818**	–	–	–	–
GADIFF	90.50	93.33	0.2142	0.2140	**59.52**	**56.41**	**0.4070**	**0.3579**

Table 2. Evaluation parameters of models based on the GEOM-Drugs dataset (δ is 1.25 Å)

Models	COV-R(%)↑		MAT-R(Å)↓		COV-P(%)↑		MAT-P(Å)↓	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
RDKIT	60.91	65.70	1.2026	1.1252	–	–	–	–
CVGAE	0.00	0.00	3.0702	2.9937	–	–	–	–
GRAPHDG	8.27	0.00	1.9722	1.9845	2.08	0.00	2.4340	2.4100
CGCF	53.96	57.06	1.2487	1.2247	21.68	13.72	1.8571	1.8066
CONFVAE	55.20	59.43	1.2380	1.1417	22.96	14.05	1.8287	1.8159
GEOMOL	67.16	71.71	1.0875	1.0586	–	–	–	–
CONFGF	62.15	70.93	1.1629	1.1596	23.42	15.52	1.7219	1.6863
DGSM	78.73	94.39	1.0154	0.9980	40.08	37.15	1.4994	1.4496
GEODIFF	89.13	97.88	0.8629	0.8529	61.47	64.55	1.1712	1.1232
DMCG	91.27	**100.00**	0.8287	0.7908	–	–	–	–
RDKIT+Clustering	87.93	**100.00**	0.8086	0.7838	–	–	–	–
GADIFF (500 epoch)	94.69	**100.00**	0.6703	0.6579	71.77	76.29	0.9997	0.9517
GADIFF (converged)	**95.37**	**100.00**	**0.6209**	**0.5960**	**74.59**	**80.73**	**0.9282**	**0.8696**

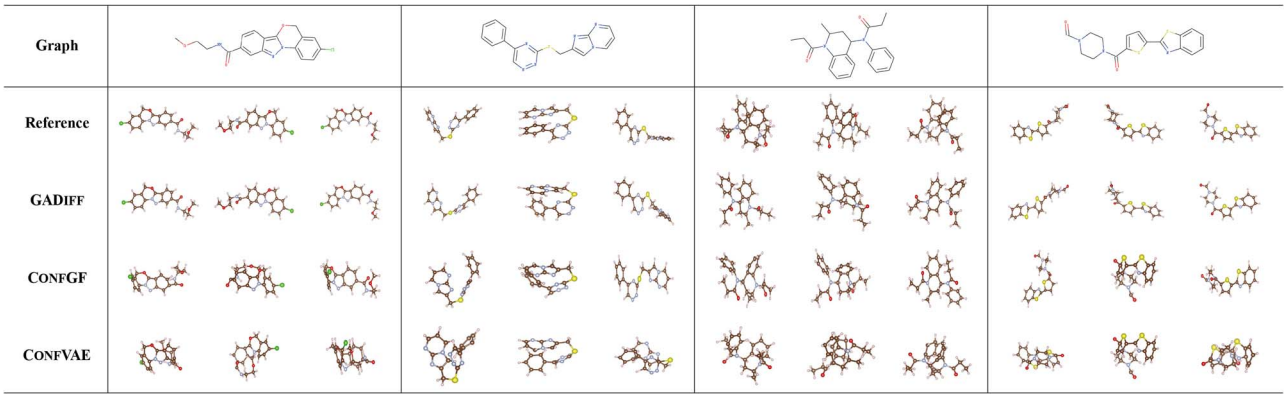


Fig. 2. Generated conformations illustration of GEOM-Drugs test set generated by proposed and baseline models. We present the conformations best-aligned with the ground truth.

conformations for each molecule [13]. The properties, the energy E and HOMO-LUMO gap ϵ of each conformation, are calculated using the chemical toolkit Psi4 [34] and compared with those of the reference conformations on five metrics: average energy $\overline{E}$, lowest energy E_{min}, average gap $\overline{\Delta\epsilon}$, minimum gap $\Delta\epsilon_{min}$ and maximum gap $\Delta\epsilon_{max}$.

Property prediction analyses. The mean absolute errors (MAEs) between the generated conformation properties and the reference conformation properties are shown in Table 3. These properties are highly sensitive to the geometric structure of molecules, so superior performance indicates the high quality of generating molecular conformations. As shown in Table 3, GADIFF achieves the best results among all baselines on $\overline{\Delta\epsilon}$ and $\Delta\epsilon_{min}$. Moreover, GADIFF performs competitively in the other three metrics, with no MAE exceeding 1, and it outperforms the current popular theoretical calculation method RDKIT in the first four metrics.

Table 3. MAEs of predicted ensemble properties (Unit: eV)

Methods	$\overline{E}$	E_{min}	$\overline{\Delta\epsilon}$	$\Delta\epsilon_{min}$	$\Delta\epsilon_{max}$
RDKIT	0.9233	0.6585	0.3698	0.8021	0.2359
G RAPHDG	9.1027	0.8882	1.7973	4.1743	0.4776
CGCF	28.9661	2.8410	2.8356	10.6361	0.5954
CONFVAE	8.2080	0.6100	1.6080	3.9111	0.2429
CONFGF	2.7886	0.1765	0.4688	2.1843	0.1433
DGSM	1.0313	0.0761	0.1963	1.1811	0.1271
GEODIFF	**0.2597**	0.1551	0.3091	0.7033	0.1909
DMCG	0.4324	**0.1364**	0.2057	1.3229	**0.1509**
GADIFF	0.8146	0.4065	**0.1042**	**0.6559**	0.4778

It indicates that GADIFF can ensure the quality of the generated conformations.

NCI conformation generation

Setup. NCIs ubiquitously exist intra- or inter- molecules and they are far weaker than covalent interactions. However, it naturally plays an important role in macromolecular studies, such as drug discovery, supramolecular assembly [22]. Therefore, accurately generating of NCI molecules is of research value [29, 35]. Due to the complexity and diversity of NCIs, and hydrogen bond being one of the strongest interactions of all NCIs, NCI molecules in datasets mainly dominated by hydrogen bonds are selected as the sampling data. To adapt GADIFF to multi-molecular system generation such as NCI molecules, a pre-trained GADIFF on GEOM-QM9 is extending to GADIFF-NCI through DTL technique and we add a graph pooling module to obtain molecular-level information.

Distributions over distances. Previously, references [10, 15] reported the detailed analysis on the edge distance distribution. They consider the distribution of interatomic distances derived from the generated conformations as indicative of the variability in atomic spacing. Due to the highly flexible structure of multi-molecular systems, RMSD-based metrics such as COV and MAT are not applicable in the evaluation of generated NCI molecular conformations any more. Therefore, distance distribution evaluation is applied to evaluate the gap between the distributions of molecular conformations generated by the GADIFF-NCI and the reference NCI dataset.

Figure 3 illustrates the instances of the generated molecular conformations from NCI datasets (S22, S66, X40, HB300SPX, HB375) and some typical hydrogen bonding types (X-H· · ·O, X-H· · ·N, X-H· · ·F) in the test set. Figure 3(a) shows five generated conformation groups from different datasets and each with one reference conformation and two generated conformations. Figure 3(b) displays three typical types of generated hydrogen-bonding molecular conformations in a manner consistent with Fig. 3(a). Moreover, a three-body water cluster is attempted to be generated by GADIFF-NCI as representing in Fig. 3(c), which displays three generated conformations and one reference conformation and demonstrates the generated conformations of the water cluster are reasonable. It has seen from Fig. 3 that GADIFF-NCI can generate molecular conformations that agree well with references from varieties of datasets. More conformation examples of NCI systems are shown in Fig. S4 of SI.

NCI property prediction. The property prediction is also use to assess the quality of the NCI molecular conformations generated by GADIFF-NCI. The evaluation metrics are consistent with that of GADIFF. Properties are calculated by a single point calculation for each generated molecule conformation using the Gaussian09 [36] program package (DFT M062X method and the 6-31G* basis set under vacuum conditions [29]). The calculated molecular interaction energy E and HOMO-LUMO gap of NCI molecular systems are compared with the properties of the reference NCI dataset on the metrics ($\overline{E}$, E_{min}, $\overline{\Delta\epsilon}$, $\Delta\epsilon_{min}$, $\Delta\epsilon_{max}$) in MAEs.

The distributions over interatomic distances are demonstrated in Fig. 4. The kernel density curve of interatomic distance distribution of the generated conformation (blue) is very close to that (orange) of the reference NCI dataset in the test set, especially the distances of (H,O) and (H,F). Moreover, the histogram of the generation conformations and reference conformations is highly overlapped, which means that generated conformation quality is similar to the reference NCI dataset. It indicates that GADIFF could be efficiently scaled to the NCI dataset and learn the reference conformation distribution.

MAEs between the generated molecular properties and the ground truth are shown in Table 4. Except for the $\Delta\epsilon_{max}$, MAEs of the other metrics are less than 0.4, which suggests that the properties of the generated NCI conformations are very similar to those in the reference NCI dataset conformations. The performance of these ensemble properties also indicates the high quality of the NCI molecular conformations generated by GADIFF-NCI.

Ablation study

In order to verify the validity of improvement on the noise model $\epsilon_\theta(G, C_t, t)$, ablation experiments are conducted on GADIFF and GADIFF-NCI, respectively.

Multi-head self-attention mechanism in GADIFF

GADIFF utilizes the MSA mechanism to fuse the atom/node vectors obtained from subgraphs, which is able to obtain significant and global information of molecular conformations. Meanwhile, GADIFF also employs the MSA mechanism to weight the conformation noise from different subgraphs. This helps us better understand the impact of noise from various subgraphs on the overall conformational noise. We name the former MSA mechanism as **node attention** and the latter MSA mechanism as **noise attention**. Table 5 shows the performance (COV, MAT) of the models on datasets (GEOM-QM9 and GEOM-Drugs) without the MSA mechanism. From the results presented in Table 5, it is clearly that the MSA mechanism impacts on both the fusion of atom vectors and processing subgraph conformation noise. On the two datasets, GADIFF with both node attention and noise attention achieves the best performance. The t-SNE visualization of the node feature vector in Fig. 5 displays the node feature vector variation after adding MSA mechanism in node attention. It is

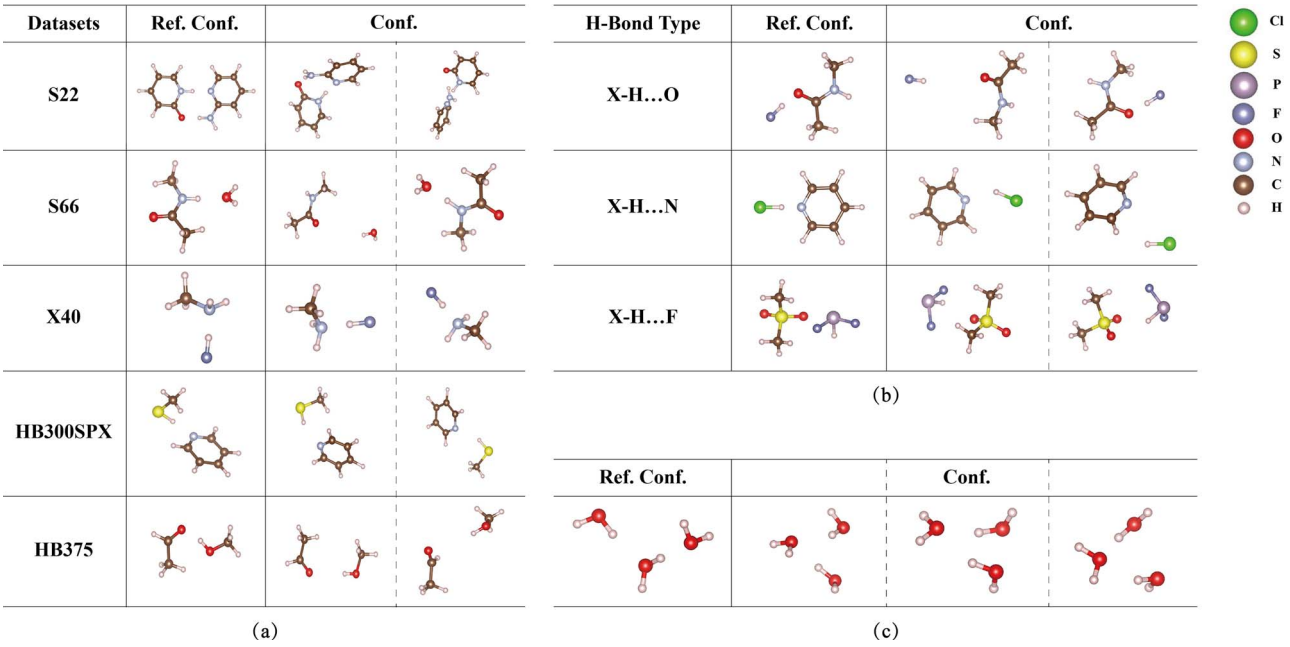


Fig. 3. Generated NCI molecular conformations in the test set by the proposed transfer model GADIFF-NCI.

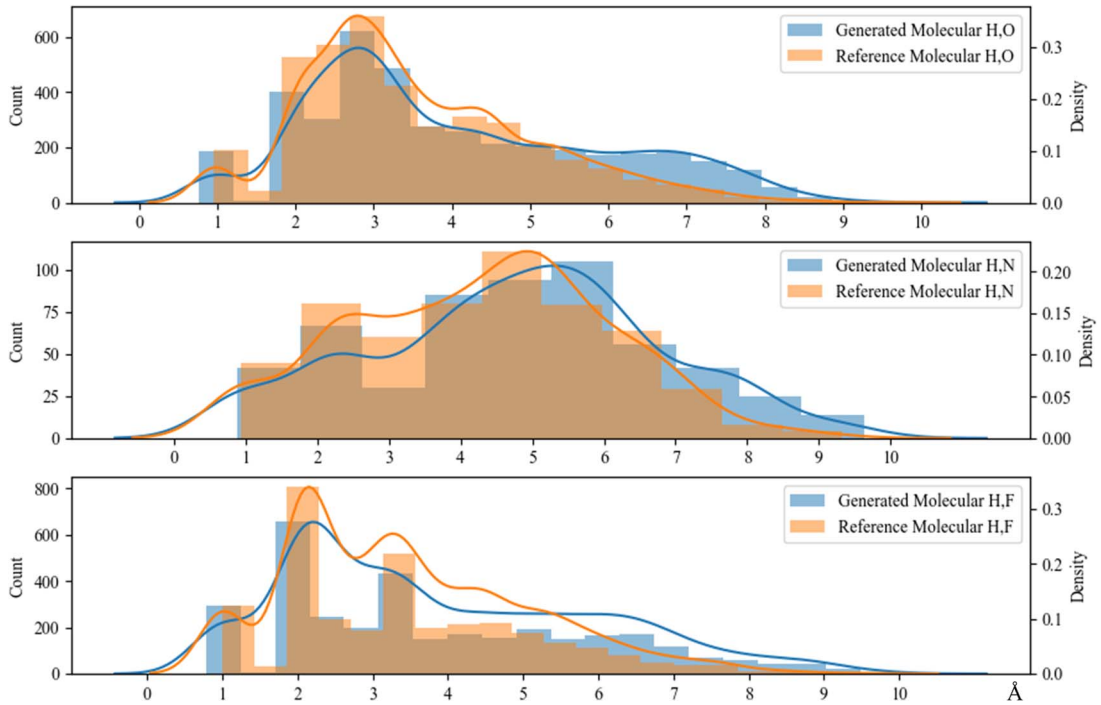


Fig. 4. The interatomic distances (H,O; H,N; H,F) distribution histogram and kernel density curves.

Table 4. MAEs of predicted ensemble properties on NCI test set (unit: eV)

Methods	$\overline{E}$	E_{min}	$\overline{\Delta\epsilon}$	$\Delta\epsilon_{min}$	$\Delta\epsilon_{max}$
GADIFF-NCI (w/o graph encoder)	0.3957	0.4507	0.3439	0.2917	0.5039
GADIFF-NCI (w/o DTL)	0.4234	0.3621	0.3146	0.2345	**0.4816**
GADIFF-NCI	**0.3648**	**0.3540**	**0.2382**	**0.2044**	0.5855

obviously that the node features from different types of subgraphs after MSA mechanism tend to be distinguishable by categories, especially long-range interaction. The reason may be that atomic bonds, bond angle interactions and torsion angle interactions are all covalent interactions, so some feature values of them are close; while long-range interaction/NCI is essentially different

Table 5. Ablations of GADIFF on GEOM-QM9 and GEOM-Drugs

Datasets	Models	COV-R(%)↑		MAT-R(Å)↓		COV-P(%)↑		MAT-P(Å)↓	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
GEOM-QM9	GADIFF (w/o node attention)	90.14	**93.75**	0.2245	0.2184	58.09	54.85	0.4159	0.3765
	GADIFF (w/o noise attention)	89.97	92.31	0.2170	0.2197	59.61	56.11	**0.4010**	0.3644
	GADIFF (used in paper)	**90.50**	93.33	**0.2142**	**0.2140**	**59.52**	**56.41**	0.4070	**0.3579**
GEOM-Drugs	GADIFF (w/o node attention)	92.66	98.76	0.7438	0.7380	69.54	76.14	1.0520	0.9847
	GADIFF (w/o noise attention)	94.42	98.92	**0.6663**	**0.6451**	71.67	**78.43**	**0.9889**	**0.9172**
	GADIFF (used in paper)	**94.69**	**100.00**	0.6703	0.6579	**71.77**	76.29	0.9997	0.9517

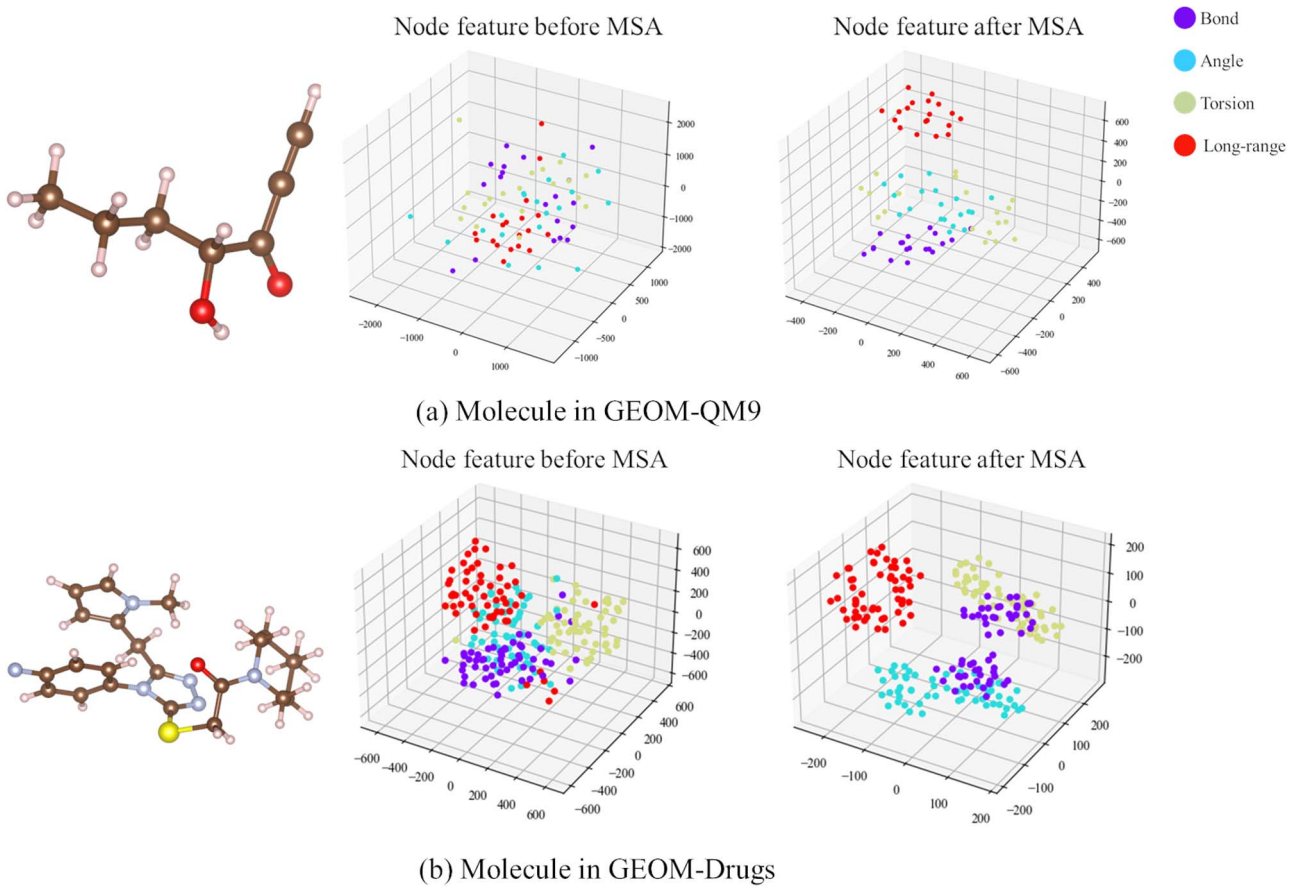


Fig. 5. The t-SNE visualization of atom/node features before and after a MSA mechanism process applied for molecules from (a) GEOM-QM9 and (b) GEOM-Drugs.

from covalent interaction, the boundary between them is evident. This clearly shows MSA mechanism in our proposed models has captured representative information of atom/node features.

Ablation experiments for GADIFF-NCI

We construct the GADIFF-NCI based on GADIFF and carry out the multi-molecular conformation generation task on the NCI datasets. Since the samples in the NCI dataset are in the form of molecular dimer unlike the single molecule form in GEOM-QM9, a graph encoder is added to obtain molecular-level features. Additionally, transfer learning is applied for introducing knowledge from GEOM-QM9 to GADIFF-NCI, which can compensate the data information deficiencies, i.e., lacking of conformation diversity and the detailed edge features in the NCI dataset.

The ablation experiments are performed on the above two items, the former is named **graph encoder** and the latter is named **DTL**. Table 4 demonstrates the ablation results, which suggests effectiveness of the graph encoder and the DTL in the GADIFF-NCI. Obviously, both DTL and graph encoder can enhance the accuracy of property prediction. It can be noticed that both the molecular-level characteristics and the source domain knowledge of multiple conformations impact on the GADIFF-NCI performance.

Additionally, an ablation study on subgraph representation is conducted to verify the importance of subgraphs to GADIFF. The experiment results are shown in Table S3 of SI.

Conclusion and future work

Diffusion generation models have exhibited great potential across various domain, but there are many issues to be addressed. In this paper, we propose a diffusion generation model, GADIFF, that focuses on data representation and noise computation. In the Markov chain framework based on equivariant Markov kernels, GADIFF utilizes MSA mechanism to improve molecular graph representation and conformation noise prediction for MCG tasks. On one hand, GADIFF fuses node vectors from subgraphs of different edge types to obtain both local and global information for a molecule; on the other hand, MSA weights the subgraph conformation noises so that the GADIFF can learn the accurate prediction of conformation noises. Comprehensive experiments on multiple tasks show that GADIFF is competitive with existing SOTA models based on GEOM-QM9 and GEOM-Drugs datasets. We further adopt fine-tune transfer learning strategy on GDAIFF to achieve a DTL model, GADIFF-NCI, for generating multi-molecular systems. GADIFF-NCI takes GADIFF as the pre-trained model that has been added a graph encoder to obtain multi-molecular graph information instead of one molecule. Then, the information learned from the GEOM-QM9 is transferred to the NCI dataset, which enable GADIFF-NCI to generate reasonable conformations for molecular pair and even trimer water clusters. The distance distribution analyses and property prediction performance on the NCI dataset indicate that GADIFF-NCI is capable of generating NCI molecule conformations of high quality. Future work includes further improving the model efficiency and capability with advanced algorithms, and extending the model to more challenging structures, such as protein-ligands.

Key Points

- We improve the diffusion generation model performance by implementing MSA mechanism on computing dynamical attention scores for conformation noises. With that, the proposed model, GADIFF, can automatically learn the weights of different subgraph conformation noises at multi-scale feature spaces to better predict the molecular conformation noise.
- We enhance molecular representation by fusing local and global information of all the nodes. The local node features are extracted by GIN, and then they are fused by MSA mechanism into global representation. Therefore, GADIFF can have comparably intact molecular information.
- Our proposed models GADIFF and GADIFF-NCI can generate not only single molecule, but also noncovalent interacting molecular pair, even a trimer water cluster, owing to transfer learning technique. Taking the pre-trained transferring strategy, GADIFF are pre-trained on GEOM-QM9 and migrated to GADIFF-NCI to model multi-molecular systems. In addition, a graph encoder is employed to obtain graph-level features via GIN network and graph attention pooling module, which ensures complete information coverage for each noncovalent molecular system.

Author contributions statement

D.W. designed models and programs, conducted experiments and result analyses, and wrote the original draft. X.D. and X.Z. conducted experiments and maintained software. L.H.H. conceived and supervised the research, analyzed results, wrote, and reviewed the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared

Funding

This work was supported by National Natural Science Foundation of China (22273010) and the Department of Science and Technology of Jilin Province (20210402075GH).

References

1. Ou-Yang SS, Lu JY, Kong XQ. *et al*. Computational drug discovery. *Acta Pharmacol Sin* 2012;**33**:1131–40. https://doi.org/10.1038/aps.2012.109.
2. Hawkins PCD. Conformation generation: the state of the art. *J Chem Inf Model* 2017;**57**:1747–56. https://doi.org/10.1021/acs.jcim.7b00221.
3. Park YJ, Kim H, Jo J. *et al*. Deep contrastive learning of molecular conformation for efficient property prediction. *Nat Comput Sci* 2023;**3**:1015–22. https://doi.org/10.1038/s43588-023-00560-w.
4. Ascherl L, Sick T, Margraf J. *et al*. Molecular docking sites designed for the generation of highly crystalline covalent organic frameworks. *Nature Chem* 2016;**8**:310–6. https://doi.org/10.1038/nchem.2444.
5. Schaller D, Šribar D, Noonan T. *et al*. Next generation 3D pharmacophore modeling. *Nature Chem* 2020;**10**:e1468. https://doi.org/10.1002/wcms.1468.
6. Zhu J, Xia Y, Liu C. *et al*. Direct molecular conformation generation. *Transact Mach Learn Res* 2022.
7. Smyth MS, Martin JH. X ray crystallography. *Mol Pathol* 2000;**53**:8–14. https://doi.org/10.1136/mp.53.1.8.
8. De Vivo M, Masetti M, Bottegoni G. *et al*. Role of molecular dynamics and related methods in drug discovery. *J Med Chem* 2016;**59**:4035–61. https://doi.org/10.1021/acs.jmedchem.5b01684.

9. Mansimov E, Mahmood O, Kang S. *et al*. Molecular geometry prediction using a deep generative graph neural network. *Sci Rep* 2019;**9**:20381. https://doi.org/10.1038/s41598-019-56773-5.
10. Simm GNC, Hernández-Lobato JM. A generative model for molecular distance geometry. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Daumé HIII, Singh A (eds.), pp. 8949–58. Vienna, Austria: PMLR, 2020.
11. Xu M, Luo S, Bengio Y. *et al*. Learning neural generative dynamics for molecular conformation generation. In: *International Conference on Learning Representations (ICLR)*. OpenReview, 2021.
12. Xu M, Wang W, Luo S. *et al*. An end-to-end framework for molecular conformation generation via bilevel programming. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Meila M, Zhang T (eds.), pp. 11537–47. PMLR, 2021.
13. Shi C, Luo S, Xu M. *et al*. Learning gradient fields for molecular conformation generation. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Meila M, Zhang T (eds.), pp. 9558–68. PMLR, 2021.
14. Ganea O, Pattanaik L, Coley C. *et al*. GeoMol: torsional geometric generation of molecular 3D conformer ensembles. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ranzato M, Beygelzimer A, Dauphin Y. *et al*. (eds.), pp. 13757–69. Curran Associates Inc., 2021.
15. Luo S, Shi C, Xu M. *et al*. Predicting molecular conformation via dynamic graph score matching. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ranzato M, Beygelzimer A, Dauphin Y. *et al*. (eds.), pp. 19784–95. Curran Associates Inc., 2021.
16. Xu M, Yu L, Song Y. *et al*. GEODIFF: a geometric diffusion model for molecular conformation generation. In: *International Conference on Learning Representations (ICLR)*, 2022.
17. Zhou G, Gao Z, Wei Z. *et al*. Do deep learning methods really perform better in molecular conformation generation? In: *ICLR 2023-Machine Learning for Drug Discovery workshop*. Kigali, Rwanda: OpenReview, 2023.
18. Hu W, Liu Y, Chen X. *et al*. Deep learning methods for small molecule drug discovery: a survey. *IEEE Trans Artif Intel* 2024;**5**: 459–79. https://doi.org/10.1109/TAI.2023.3251977.
19. Yang L, Zhang Z, Song Y. *et al*. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput Surv* 2023;**56**:1–39.
20. Gruver N, Stanton S, Frey N. *et al*. Protein design with guided discrete diffusion. In: *Advances in Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA*. Oh A, Naumann T, Globerson A. *et al*. (eds.), pp. 12489–517. Curran Associates Inc., 2023.
21. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Larochelle H, Ranzato M, Hadsell R (eds.), pp. 6840–51. Vancouver, Canada: Curran Associates Inc., 2020.
22. Beno BR, Yeung KS, Bartberger MD. *et al*. A survey of the role of noncovalent sulfur interactions in drug design. *J Med Chem* 2015;**58**:4383–438. https://doi.org/10.1021/jm501853m.
23. Gráfová L, Pitoňák M, Řezáč J. *et al*. Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. *J Chem Theory Comput* 2010;**6**:2365–76. https://doi.org/10.1021/ct1002253.
24. Řezáč J, Riley KE, Hobza P. S66: a well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J Chem Theory Comput* 2011;**7**:2427–38. https://doi.org/10.1021/ct2002946.
25. Řezáč J, Riley KE, Hobza P. Benchmark calculations of noncovalent interactions of halogenated molecules. *J Chem Theory Comput* 2012;**8**:4285–92. https://doi.org/10.1021/ct300647k.
26. Řezáč J. Non-covalent interactions atlas benchmark data sets 2: hydrogen bonding in an extended chemical space. *J Chem Theory Comput* 2020;**16**:6305–16. https://doi.org/10.1021/acs.jctc.0c00715.
27. Řezáč J. Non-covalent interactions atlas benchmark data sets: hydrogen bonding. *J Chem Theory Comput* 2020;**16**:2355–68. https://doi.org/10.1021/acs.jctc.9b01265.
28. Axelrod S, Gomez-Bombarelli R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci Data* 2022;**9**:185. https://doi.org/10.1038/s41597-022-01288-4.
29. Wang D, Li W, Dong X. *et al*. TFRegNCI: interpretable noncovalent interaction correction multimodal based on transformer encoder fusion. *J Chem Inf Model* 2023;**63**:782–93. https://doi.org/10.1021/acs.jcim.2c01283.
30. Riniker S, Landrum GA. Better informed distance geometry: using what we know to improve conformation generation. *J Chem Inf Model* 2015;**55**:2562–74. https://doi.org/10.1021/acs.jcim.5b00654.
31. Liberti L, Lavor C, Maculan N. *et al*. Euclidean distance geometry and applications. *SIAM Rev* 2014;**56**:3–69. https://doi.org/10.1137/120875909.
32. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr Sect A* 1976;**32**:922–3. https://doi.org/10.1107/S0567739476001873.
33. O'Boyle NM, Vandermeersch T, Flynn CJ. *et al*. Confab - systematic generation of diverse low-energy conformers. *J Chem* 2011;**3**: 1–9. https://doi.org/10.1186/1758-2946-3-8.
34. Smith DG, Burns LA, Simmonett AC. *et al*. psi4 1.4: open-source software for high-throughput quantum chemistry. *J Chem Phys* 2020;**152**:184108. https://doi.org/10.1063/5.0006002.
35. Li W, Wang D, Yang Z. *et al*. DeepNCI: DFT noncovalent interaction correction with transferable multimodal three-dimensional convolutional neural networks. *J Chem Inf Model* 2022;**62**:5090–9. https://doi.org/10.1021/acs.jcim.1c01305.
36. Frisch MJ, Trucks GW, Schlegel HB. *et al*. *Gaussian 09, Revision d.01*. Wallingford CT: Gaussian Inc, 2013.

Briefings in Bioinformatics, 2025, **26(1)**, bbae676
https://doi.org/10.1093/bib/bbae676
Problem Solving Protocol