

GeoDiff-Geometric diffusion model for molecular conformation generation

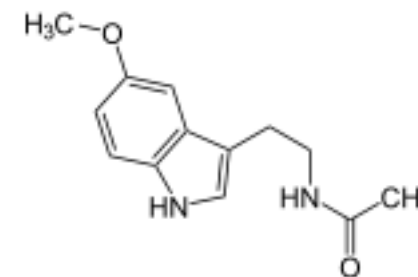
Minkai Xu *et. al*, ICLR 2022

Sourjya Sarkar

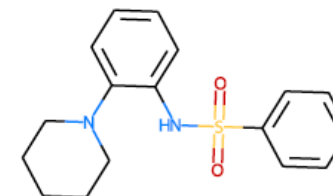
Molecular Graph Representation

- 1D Representation (Smiles)

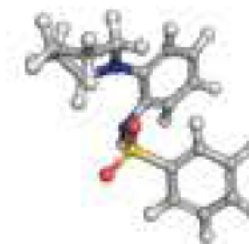
```
CC(=O)NCCC1=CNc2c1cc(OC)cc2  
CC(=O)NCCc1c[nH]c2ccc(OC)cc12
```



- 2D Representation as graphs (nodes are atoms, edges are bonds)

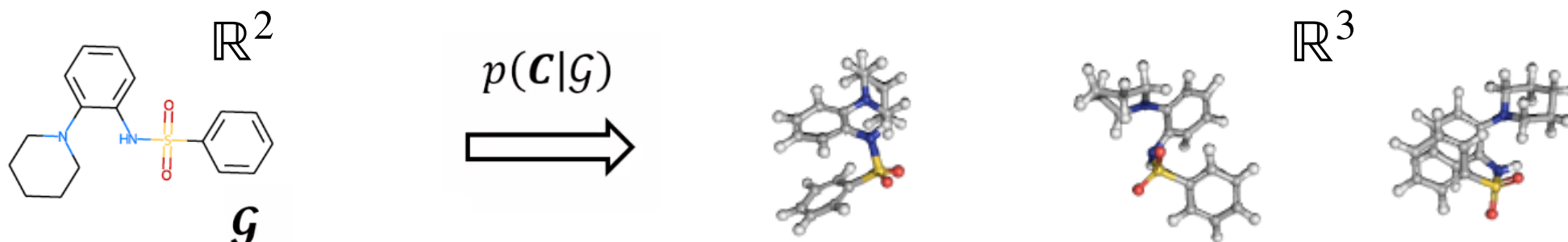


- 3D Representation for graphs (nodes are atoms, edges are bonds, torsion angles)



Research Issue

Generate 3D conformation i.e., representation of atoms (nodes) and bonds (edges) in a molecule in 3D



Challenges

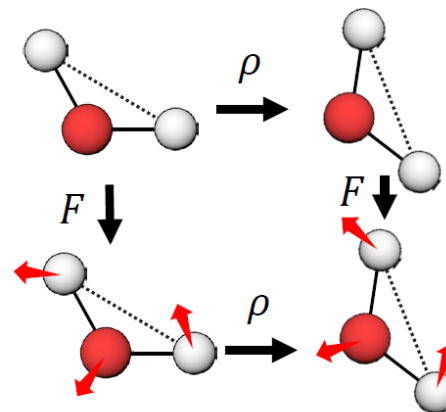
- Roto-Translation Equivariance of conformations
 - (a) The vector of atomic forces should rotate w.r.t. conformation coordinates
 - (b) Inter-atomic distances should not change while rotating the molecule.
- Generating 3D structures are expensive using traditional approaches (X-ray crystallography, Heuristic based)

Roto-Translation Equivariance

- SE(3) i.e., group closed under rotations and translations in \mathbb{R}^3
- Equivariance
 - $f(T_g(x)) = S_g(f(x))$
- SE(3) Equivariance
 - Rotation matrix \mathbf{R}
 - Translation vector \mathbf{t}
 - $\mathbf{R}z_x + \mathbf{t}, z_h = f(\mathbf{R}\mathbf{x} + \mathbf{t}, \mathbf{h})$
 $z_x, z_h = f(\mathbf{x}, \mathbf{h})$

Node features

- Spatial position of atoms $\mathbf{x} \in \mathbb{R}^3$
- Atom Features \mathbf{h} (type, charge etc.)



Prior ML-based Approaches

- **Likelihood** of conformations **is not rotation and translation invariant** in CVGAE.
- GraphDG, CGCF, ConfVAE, and ConfGF overcome it by **using distances for learning or inference**, which are invariant under rotation and translation. In practice these approaches suffer from the following two key limitations:
 - Noise in generated distances can cause **accumulated error** for 3D coordinate reconstruction, leading to less accurate or even erroneous structures.
 - Learning over noisy distances leading to the **training-test inconsistent** problem.

GraphDG, "A generative model for molecular distance geometry." *arXiv preprint arXiv:1909.11459* (2019)

ConfVAE "An end-to-end framework for molecular conformation generation via bilevel programming." ICML, 2021.

ConfGF "Learning gradient fields for molecular conformation generation." ICML, 2021.

CVGAE "Molecular geometry prediction using a deep generative graph neural network." *Scientific reports* 9.1 (2019): 1-13.

Key Contributions

- (C1) preserve the toto-translation equivariance of conformations
- (C2) directly conduct learning and sampling in the coordinate space (3D)

Problem Formulation

- Diffusion model directly operate on coordinate space

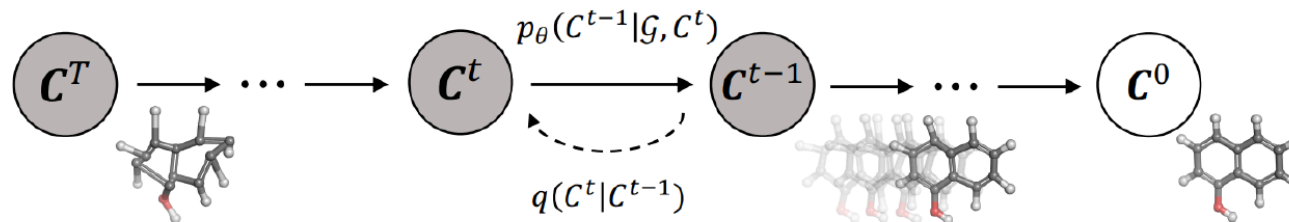


Figure 5. Overview of GeoDiff framework.

- For the **diffusion process**, noise from fixed posterior distributions $q(C^t | C^{t-1})$ is gradually added until the conformation is destroyed. Symmetrically, for the **generative process**, an initial state C^T is sampled from standard Gaussian distribution, and progressively refined via the Markov $p_\theta(C^{t-1} | G, C^t)$ (**satisfy C2**).
- Then, then key problem is how to efficiently optimize the induced likelihood $p_\theta(C^0 | G)$ and meanwhile impose the SE(3)-equivariance (**satisfy C1**).

Generative Model - Diffusion Process

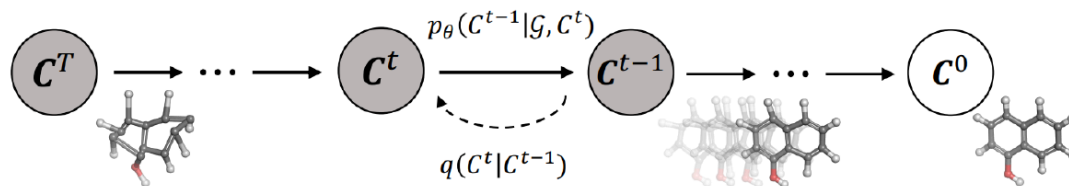
- Fixed diffusion (forward) process:

$$q(\mathcal{C}^{1:T}|\mathcal{C}^0) = \prod_{t=1}^T q(\mathcal{C}^t|\mathcal{C}^{t-1}), \quad q(\mathcal{C}^t|\mathcal{C}^{t-1}) = \mathcal{N}(\mathcal{C}^t; \sqrt{1 - \beta_t}\mathcal{C}^{t-1}, \beta_t I)$$

- Learnable generative (reverse) process:

$$p_{\theta}(\mathcal{C}^{0:T-1}|\mathcal{G}, \mathcal{C}^T) = \prod_{t=1}^T p_{\theta}(\mathcal{C}^{t-1}|\mathcal{G}, \mathcal{C}^t), \quad p_{\theta}(\mathcal{C}^{t-1}|\mathcal{G}, \mathcal{C}^t) = \mathcal{N}(\mathcal{C}^{t-1}; \mu_{\theta}(\mathcal{G}, \mathcal{C}^t, t), \sigma_t^2 I)$$

- Our goal is to learn a generative process that can revert the diffusion process and by gradually eliminating the noise to recover the ground-truth.



Training Objective- ELBO

- Maximize ELBO

$$\begin{aligned}\mathbb{E} [\log p_{\theta}(\mathcal{C}^0|\mathcal{G})] &= \mathbb{E} \left[\log \mathbb{E}_{q(\mathcal{C}^{1:T}|\mathcal{C}^0)} \frac{p_{\theta}(\mathcal{C}^{0:T}|\mathcal{G})}{q(\mathcal{C}^{1:T}|\mathcal{C}^0)} \right] \\ &\geq -\mathbb{E}_q \left[\sum_{t=1}^T D_{\text{KL}}(q(\mathcal{C}^{t-1}|\mathcal{C}^t, \mathcal{C}^0) \| p_{\theta}(\mathcal{C}^{t-1}|\mathcal{C}^t, \mathcal{G})) \right] := -\mathcal{L}_{\text{ELBO}}\end{aligned}$$

- The ELBO can be re-written as follows (Ho *et.al.* 2020)

$$\mathcal{L}_{\text{ELBO}} = \sum_{t=1}^T \gamma_t \mathbb{E}_{\{\mathcal{C}^0, \mathcal{G}\} \sim q(\mathcal{C}^0, \mathcal{G}), \epsilon \sim \mathcal{N}(0, I)} \left[\|\epsilon - \epsilon_{\theta}(\mathcal{G}, \mathcal{C}^t, t)\|_2^2 \right]$$

where $\mathcal{C}^t = \sqrt{\bar{\alpha}_t} \mathcal{C}^0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$. The weights $\gamma_t = \frac{\beta_t}{2\alpha_t(1 - \bar{\alpha}_{t-1})}$ for $t > 1$, and $\gamma_1 = \frac{1}{2\alpha_1}$.

Preserving Group Equivariance Property

- The **marginal likelihood** $p_\theta(C^0|\mathcal{G}) = \int p(\mathcal{C}^T) p_\theta(\mathcal{C}^{0:T-1}|\mathcal{G}, \mathcal{C}^T) d\mathcal{C}^{1:T}$
- Then a non-trivial problem is how to preserve **SE(3) invariance** of $p_\theta(C^0|\mathcal{G})$?

Proposition 1

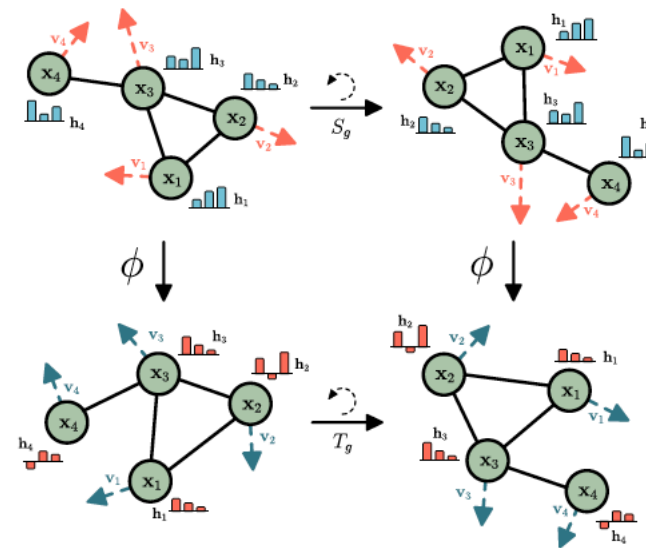
- Given **Invariant** prior density $p(\mathcal{C}^T)$ with zero center of mass (COM)
- Given **Equivariant** Markov Transition Kernel $p(\mathcal{C}^{t-1}|\mathcal{G}, \mathcal{C}^t)$
- Proves that the marginal likelihood is **group invariant**

Question

- How to design a **Equivariant** Markov Transition Kernel $p(\mathcal{C}^{t-1}|\mathcal{G}, \mathcal{C}^t)$?
- Equivalently. How to design a SE(3) **Equivariant** ϵ_θ ?

Graph Field Network

- Graph Field Network ϵ_θ , $\theta = (\theta_m, \theta_h, \theta_x)$
- SE(3) Invariant message passing step
 - $\mathbf{m}_{ij} = \Phi_m \left(\mathbf{h}_i^l, \mathbf{h}_j^l, \|x_i^l - x_j^l\|^2, e_{ij}; \theta_m \right)$
- Feature Update
 - $\mathbf{h}_i^{l+1} = \Phi_h \left(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij}; \theta_h \right)$
- SE(3) Equivariant position update
 - $x_i^{l+1} = \sum_{j \in \mathcal{N}(i)} \frac{1}{d_{ij}} (\mathbf{c}_i - \mathbf{c}_j) \Phi_x (\mathbf{m}_{ij}; \theta_x)$
- Φ_m, Φ_h, Φ_x are feed forward neural networks
- $\mathbf{h}^0 \in \mathbb{R}^{n \times b}$
- $x_i^0 \in \mathbb{R}^{n \times 3} = \mathbf{c}$



Proposition 2 (GEODIFF).

Proves that GFL is SE(3) Equivariant w.r.t conformations

GeoDiff Sampling Algorithm

Algorithm 1 Sampling Algorithm of GEODIFF.

Input: the molecular graph \mathcal{G} , the learned reverse model ϵ_θ .

Output: the molecular conformation \mathcal{C} .

- 1: Sample $\mathcal{C}^T \sim p(\mathcal{C}^T) = \mathcal{N}(0, I)$
 - 2: **for** $s = T, T - 1, \dots, 1$ **do**
 - 3: Shift \mathcal{C}^s to zero CoM
 - 4: Compute $\mu_\theta(\mathcal{C}^s, \mathcal{G}, s)$ from $\epsilon_\theta(\mathcal{C}^s, \mathcal{G}, s)$ using equation 4
 - 5: Sample $\mathcal{C}^{s-1} \sim \mathcal{N}(\mathcal{C}^{s-1}; \mu_\theta(\mathcal{C}^s, \mathcal{G}, s), \sigma_t^2 I)$
 - 6: **end for**
 - 7: **return** \mathcal{C}^0 as \mathcal{C}
-

$$\mu_\theta(\mathcal{C}^t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathcal{C}^t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathcal{G}, \mathcal{C}^t, t) \right), \quad (4)$$

Experiments

- **Datasets**

- **GEOM-QM9** Small molecules
- **GEOM-Drugs** Medium-Sized Molecules
- **Training Set** 40,000 molecules with 5 conformations for each i.e., 200000 conformations in total
- **Validation Set** Same size as training set
- **Test Set** 200 distinct molecules with 22408 conformations (QM9) and 14,324 (Drugs)

- **Baselines**

- CVGAE(Mansimov *et al* 2019), GeoMol (Xu *et al* 2021) GraphDG,(Simm 2020) ConfGF(Shi *et.al* 2021), GeoMol (Ganea *et. al* 2021)
- RDKit (classical Euclidean Distance Geometry-based approach)

Metrics

- Evaluate the **quality** and **diversity** of generated conformations following the conventional **recall** measurement.
- **Coverage (COV-R)**: the fraction of conformations in the reference set that are matched by at least one conformation in the generated set:

$$\text{COV}(\mathbb{S}_g(\mathcal{G}), \mathbb{S}_r(\mathcal{G})) = \frac{1}{|\mathbb{S}_r|} \left| \left\{ \mathbf{R} \in \mathbb{S}_r \mid \text{RMSD}(\mathbf{R}, \mathbf{R}') < \delta, \mathbf{R}' \in \mathbb{S}_g \right\} \right|$$

- **Matching (MAT-R)**: measure the average distance of the reference conformations with their nearest neighbors in the generated set:

$$\text{MAT}(\mathbb{S}_g(\mathcal{G}), \mathbb{S}_r(\mathcal{G})) = \frac{1}{|\mathbb{S}_r|} \sum_{\mathbf{R}' \in \mathbb{S}_r} \min_{\mathbf{R} \in \mathbb{S}_g} \text{RMSD}(\mathbf{R}, \mathbf{R}').$$

- The other two metrics **COV-P** and **MAT-P** inspired by **precision** can be defined similarly, but with the generated and reference sets exchanged.

Results

Table 1: Results on the **GEOM-Drugs** dataset, without FF optimization.

Models	COV-R (%) \uparrow		MAT-R (\AA) \downarrow		COV-P (%) \uparrow		MAT-P (\AA) \downarrow	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
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
GEODIFF-A	88.36	96.09	0.8704	0.8628	60.14	61.25	1.1864	1.1391
GEODIFF-C	89.13	97.88	0.8629	0.8529	61.47	64.55	1.1712	1.1232

* The COV-R and MAT-R results of CVGAE, GRAPHDG, CGCF, and CONFGF are borrowed from Shi et al. (2021). The results of GEOMOL are borrowed from a most recent study Zhu et al. (2022). Other results are obtained by our own experiments. The results of all models for the GEOM-QM9 dataset (summarized in Tab. 5) are collected in the same way.

- GeoDiff achieves the state-of-the-art performances on all four metrics.

Results

Table 2: Results on the **GEOM-Drugs** dataset, with FF optimization.

Models	COV-R (%) \uparrow		MAT-R (\AA) \downarrow		COV-P (%) \uparrow		MAT-P (\AA) \downarrow	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
RDKit	60.91	65.70	1.2026	1.1252	72.22	88.72	1.0976	0.9539
GEODIFF + FF	92.27	100.00	0.7618	0.7340	84.51	95.86	0.9834	0.9221

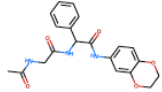
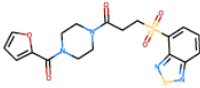
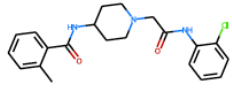
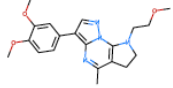
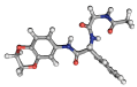
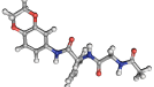
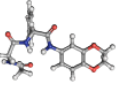
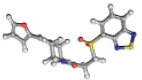
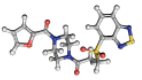
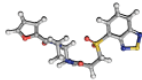
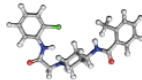
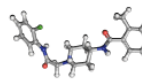
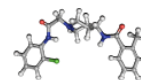
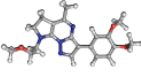
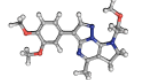
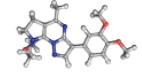
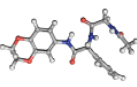
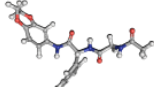
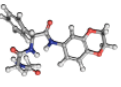
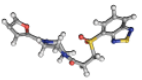
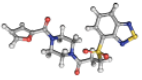
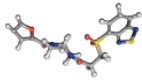
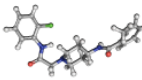
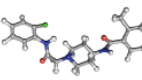
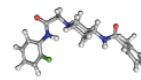
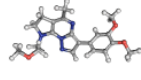
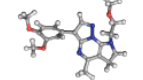
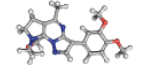
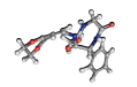
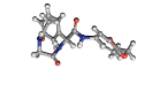
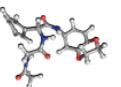
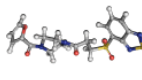
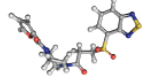
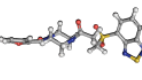
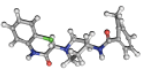

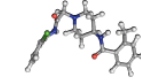
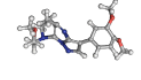
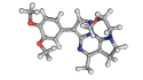
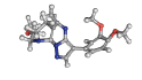




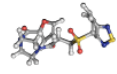



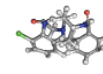



- GeoDiff outperform classic method with Force-Field optimization.

Table 6: Additional results on the **GEOM-Drugs** dataset, without FF optimization.

Models	COV-R (%) \uparrow		MAT-R (\AA) \downarrow		COV-P (%) \uparrow		MAT-P (\AA) \downarrow	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GEODIFF (T=1000)	82.96	96.29	0.9525	0.9334	48.27	46.03	1.3205	1.2724

- GeoDiff with fewer timesteps, which still outperform existing ML models.

Conformations generated

Graph												
Reference												
GeoDiff												
ConfGF												
GraphDG												

Future Work

- More challenging structures such as protein.
Translation and rotation Equivariant architectures are often also symmetric under reflection, violating fundamental structural properties of protein like chirality.
- Accelerated sampling for diffusion models (choice of kernels)
- Understanding the molecule(ligand) and protein binding process (docking)

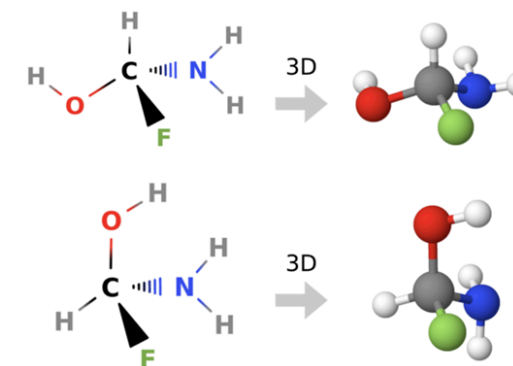


Figure 4: Chirality: even if the two shown graphs are isomorphic, they have distinct 3D structures that can be distinguished by the order of the carbon center's neighbors.

Thank You