VIETNAM NATIONAL UNIVERSITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Group : Drug Discovery - AI4Science
Student : Dang Cao Cuong
Student ID : 1952598

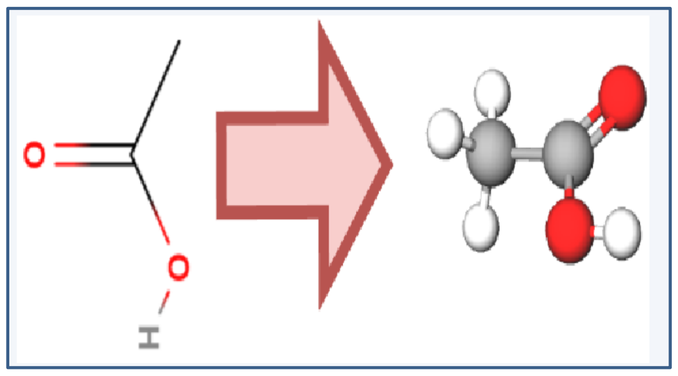
CAPSTONE PROJECT REPORT

Diffusion Model for Conformer Generation

Instructor: Assoc. Prof Quan Thanh Tho, PhD

Demo **Personal Info** **Grading**

Introduction



A conformer is a 3D structure of a molecule and conformer generation is the process of predicting or generating the three-dimensional structure of molecules given their graph, including graph structure, node type, and edge type. Graph structure is the connectivity between nodes of a molecule. Node types and edge types are labeled based on the types of atoms and bonds in a molecule. For example, the image on the left shows that a conformer of the acetic acid molecule is formed given its graph.

Problem Statement/Challenges

- Conformer generation: an important task in several scientific fields, such as bioinformatics, pharmacology, etc.
- Biocomputational method: accurate but time-consuming when dealing with large molecular graphs
- ML methods: difficult to find a suitable GNN architecture for spatiotemporal data commonly used to model 3D structures of molecules in space

→ A stable, fast, diverse, and accurate method is critical

Research Method

- Design ML method to learn the probability distribution of conformers given their graphs
- Propose a score-based diffusion model (SDM) which utilizes the score-based function to approximate small trajectories mapping from a tractable distribution, such as a Gaussian distribution, to the distribution of conformers
- Advantage: small mapping trajectories → easier for GNN to learn the trajectories
- Baseline: two SOTA methods, GeoDiff (ML method) & RDKit (biocomputational method)

Metrics

- RMSD: Root mean square deviation
- $\{C_l^*\}_{l\in[1,L]}$: set of groundtruth conformers
- $\{C_k\}_{k\in[1,K]}$: set of generated conformers
- δ: threshold

$$\text{COV}-\text{R} := \frac{1}{L}\left|\{l \in [1..L] : \exists k \in [1..K], \text{RMSD}(C_k, C_l^*) < \delta\right|$$

Training & Sampling

Algorithm Training procedure

Require: molecules with graphs $[\mathcal{G}_0, ..., \mathcal{G}_N]$ each with true conformers $[\mathcal{C}_{\mathcal{G},1}, ..., \mathcal{C}_{\mathcal{G},K_\mathcal{G}}]$, learning rate α

for $epoch \leftarrow 1$ **to** $epoch_{max}$ **do**
  for $\mathcal{G}$ **in** $[\mathcal{G}_0, ..., \mathcal{G}_N]$ **do**
    Sample $t \in [0, 1]$ and $\mathcal{C} \in [\mathcal{C}_{\mathcal{G},1}, ..., \mathcal{C}_{\mathcal{G},K_\mathcal{G}}]$
    Sample $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$
    $\sigma_t \leftarrow e^t$
    Update $\theta \leftarrow \theta - \alpha\nabla_\theta||\mathbf{z} + \mathbf{s}_\theta(\mathcal{C}^0 + \sigma_t\epsilon, \mathcal{G}, t)||^2$
  end for
end for
return trained score model $\mathbf{s}_\theta$

Algorithm Sampling procedure

Require: molecular graph $\mathcal{G}$, noise levels $\{\sigma_i\}_{i=1}^T$, the small step size ϵ, the small step size ϵ, the learned score model $\mathbf{s}_\theta$, and the number of steps per noise level L.

Sample a conformer $\mathcal{C}^T \sim p(\mathcal{C}^T) = \mathcal{N}(0, \mathbf{I})$
for $t \leftarrow T$ **to** 1 **do**
  Shift $\mathcal{C}^t$ to zero CoM
  $\alpha_t \leftarrow \epsilon \cdot \sigma_t^2/\sigma_T^2$
  for $i \leftarrow 1$ **to** L **do**
    Draw $\mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I})$
    $\mathcal{C}^t \leftarrow \mathcal{C}^t + \alpha_t\mathbf{s}_\theta(\mathcal{G}, \mathcal{C}^t, t) + \sqrt{2\alpha_t}\mathbf{z}_i$
  end for
  $\mathcal{C}^{t-1} \leftarrow \mathcal{C}^t$
end for
return conformer $\mathcal{C}^0$

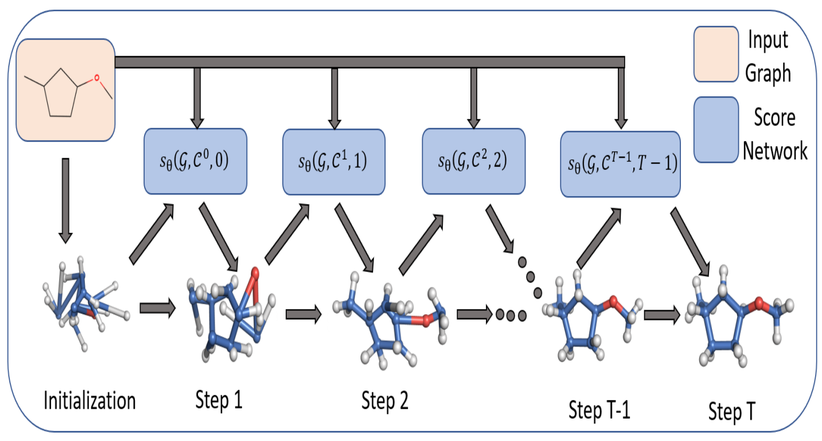
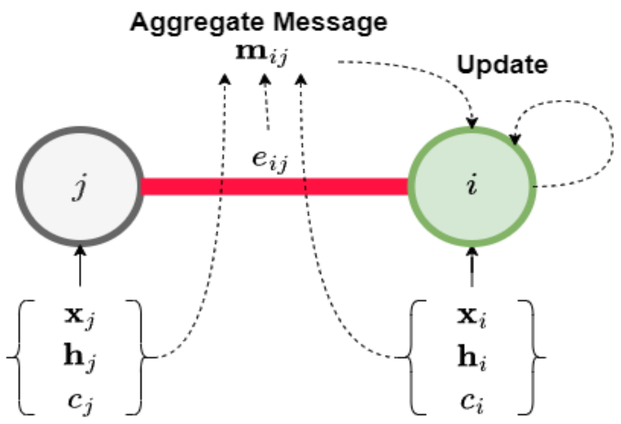


Figure: Generation procedure of the system via Langevin dynamics.

GNN Architecture

- h, x: Node & coordinate embedding
- e, d: edge type and edge length
- c: node coordinate
- Φ: MLP layers



$$\mathbf{m}_{ij} = \Phi_m\left(\mathbf{h}_i^l, \mathbf{h}_j^l, ||\mathbf{x}_i^l - \mathbf{x}_j^l||^2, e_{ij}, t; \theta_m\right)$$

$$\mathbf{h}_i^{l+1} = \Phi_h\left(\mathbf{h}_i^l, \sum_{j\in\mathcal{N}(i)} \mathbf{m}_{ij}; \theta_h\right)$$

$$\mathbf{x}_i^{l+1} = \sum_{j\in\mathcal{N}(i)} \frac{1}{d_{ij}}\left(\mathbf{c}_i - \mathbf{c}_j\right)\Phi_x\left(\mathbf{m}_{ij}; \theta_x\right)$$

Dataset

- GEOM dataset: 37 million conformers
- 133,000 species from QM9 and 317,000 species with experimental data related to biophysics, physiology, and physical chemistry
- Focus on GEOM-QM9, a medium part of GEOM
- Total size: 3.65GB with 133,254 files

Results

Models	Theshold	COV-R(%) Mean ↑	COV-R(%) Median ↑	COV-R(%) Std ↓	COV-P(%) Mean ↑	COV-P(%) Median ↑	COV-P(%) Std ↓
DSM	1.00	0.42	0	4.00	0.07	0	0.65
	1.65	**85.79**	**100**	25.11	**31.15**	**31.21**	15.64
	1.70	**89.49**	**100**	22.05	**39.21**	**40.17**	17.38
	1.80	**94.79**	**100**	15.83	**56.47**	**59.12**	19.45
DDPM(GeoDiff)	1.00	0.50	0	6.03	0.01	0	0.09
	1.65	68.33	76.77	29.19	4.54	4.17	3.13
	1.70	78.86	90.69	25.19	6.09	5.44	3.80
	1.80	93.33	100	15.86	10.30	9.81	5.19
RDKit	1.00	28.4	0	39.44	26.08	0	39.43
	1.65	96.20	100	15.67	92.77	100	20.83
	1.70	97.23	100	13.18	94.32	100	18.88
	1.80	98.31	100	10.65	96.22	100	15.38

Table: The results of conformer generation when applying RDKit, DSM, and DDPM with thresholds of 1.0, 1.65, 1.7 and 1.8

Conclusion

The DSM method gives better results than the DDPM method used in GeoDiff in terms of both metrics, COV-R and COV-P. In terms of mean, the DSM method gives about 25% better results on the COV-R metric and 6.8 times better on the COV-P metric with a threshold of 1.65. Although RDKit biocomputational method has slightly better results than DSM, in practice RDKit method skips 5/200 conformers due to failure to generate a conformer. In brief, the proposed method is the most appropriate one for tasks in the drug discovery industry such as creating a large number of conformers which requires fast, stable, diverse, and accurate conformer generation.

Future Work

- Apply a new generative model
- Use more expressive GNNs
- Inspect the latent space
- Replace GNNs with LLMs
- Generate rotatable bonds