

Deep Generative Models for Molecular Conformation Generation

Jian Tang

Mila-Quebec AI Institute

HEC Montreal

CIFAR AI Chair,

Homepage: www.jian-tang.com



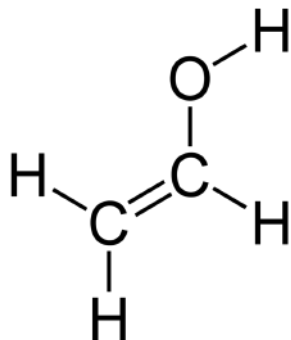
Acknowledgements: Minkai Xu, Shitong Luo, Jian Peng, and Yoshua Bengio

Molecule Representations

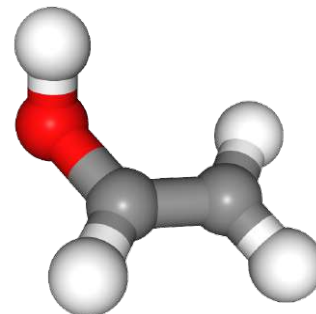
- Understanding properties of molecules is important in a variety of applications
 - Drug discovery, material discovery
- Molecule representations
 - 1D SMILES
 - 2D Molecular graphs
- A more natural and intrinsic representations: **3D conformations**
 - Determines its biological and physical activities
 - E.g., charge distribution, steric constraints, and interaction with other molecules

C1CO

1D SMILES



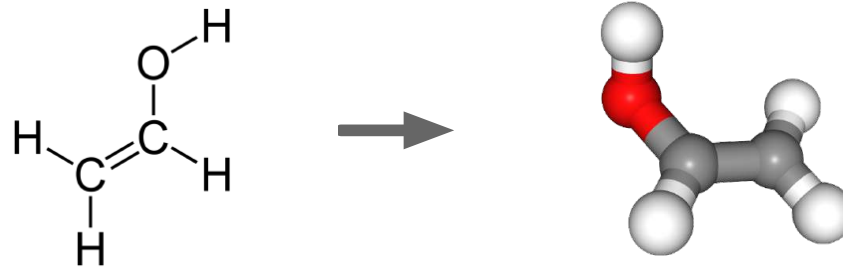
2D Graph



3D Conformation

Conformation Prediction

- For most molecules, their 3D structure are not available
- How to predict valid and stable conformations?
 - Each atom is represented as its 3D coordinates

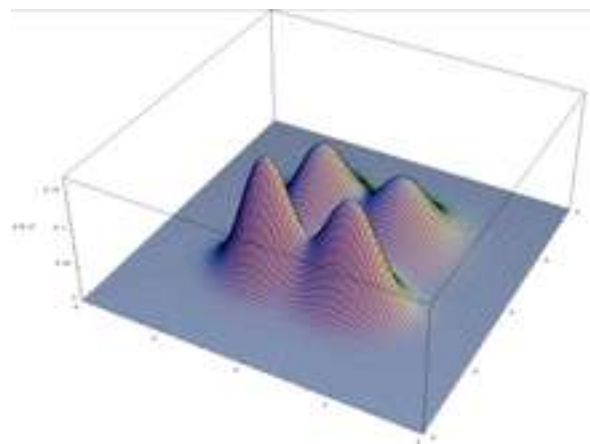


Traditional Approaches

- Experimental methods
 - Crystallography
 - Expensive and time consuming
- Computational methods
 - Molecular dynamics, Markov chain Monte Carlo
 - Very computational expensive, especially for large molecules

Machine Learning Approaches

- Train a model to predict molecular conformations \mathbf{R} given the molecular graph \mathcal{G} , i.e., modeling $p(\mathbf{R}|\mathcal{G})$ (Mansimov et al. 2019, Simm and Hernandez-Lobato 2020)
- Challenges
 - Conformations are rotation and translation equivalent
 - The distribution $p(\mathbf{R}|\mathcal{G})$ is multimodal and very complex



Our Solution

- A flexible generative model $p_{\theta}(\mathbf{R}|\mathcal{G})$ based on normalizing flows
 - Treating pairwise distances \mathbf{d} as intermediate variables
 - First generating the distance \mathbf{d} based \mathcal{G} , i. e. $p_{\theta}(\mathbf{d}|\mathcal{G})$
 - Generating conformations based on \mathbf{d} and \mathcal{G} , i.e. $p_{\theta}(\mathbf{R}|\mathbf{d}, \mathcal{G})$

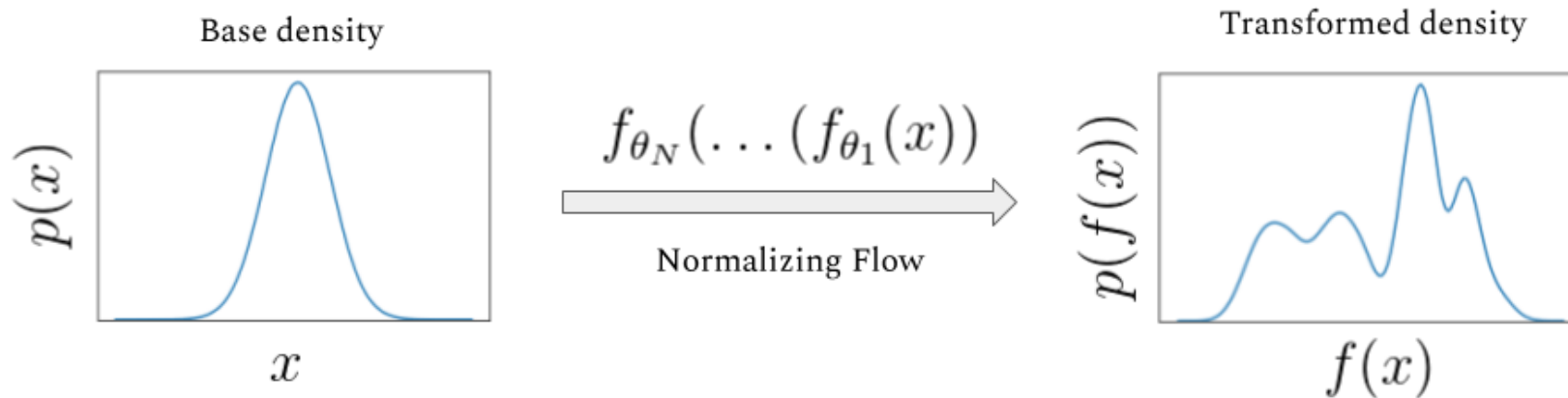
$$p_{\theta}(\mathbf{R}|\mathcal{G}) = \int p(\mathbf{R}|\mathbf{d}, \mathcal{G}) \cdot p_{\theta}(\mathbf{d}|\mathcal{G}) \mathrm{d}\mathbf{d}.$$

- Further correct $p_{\theta}(\mathbf{R}|\mathcal{G})$ with an energy-based tilting term $E_{\phi}(\mathbf{R}, \mathcal{G})$

$$p_{\theta, \phi}(\mathbf{R}|\mathcal{G}) \propto p_{\theta}(\mathbf{R}|\mathcal{G}) \cdot \exp(-E_{\phi}(\mathbf{R}, \mathcal{G}))$$

Normalizing Flows

- Defines an invertible mapping $y = f(x)$ from a base distribution to a complex distribution



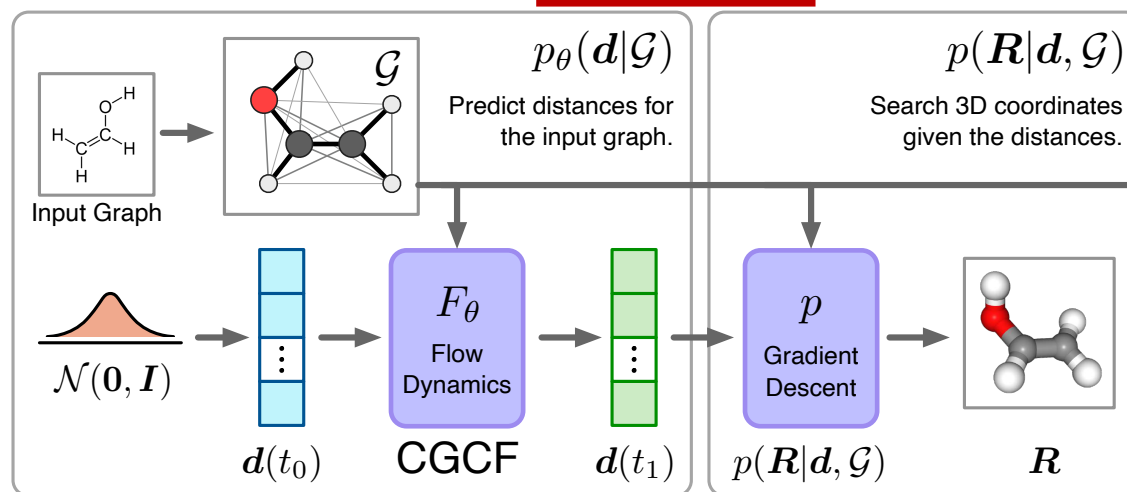
- Change-of-variable theorem

$$\hat{p}(\mathbf{y}) = p(f^{-1}(\mathbf{y})) \left| \det \frac{\partial f^{-1}}{\partial \mathbf{y}} \right| = p(\mathbf{x}) \left| \det \frac{\partial f}{\partial \mathbf{x}} \right|^{-1}$$

Distance Geometry Generation $p_{\theta}(\mathbf{d}|\mathcal{G})$

- Conditional Graph Continuous Flow (CGCF)
 - Defines an invertible mapping between a base distribution and the pairwise atom distance \mathbf{d} conditioning on the molecular graph \mathcal{G}
 - Defines the continuous dynamics of distance \mathbf{d} with Neural Ordinary Differential Equations (ODEs):

$$\mathbf{d} = F_{\theta}(\mathbf{d}(t_0), \mathcal{G}) = \mathbf{d}(t_0) + \int_{t_0}^{t_1} \boxed{f_{\theta}(\mathbf{d}(t), t; \mathcal{G})} dt, \quad \mathbf{d}(t_0) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

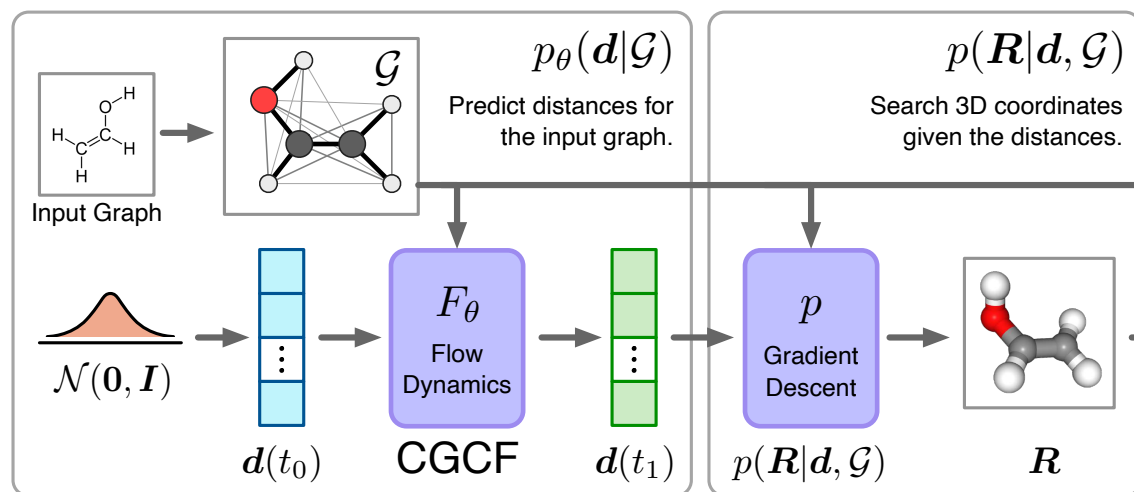


Conformation Prediction $p(\mathbf{R}|\mathbf{d}, \mathcal{G})$

- Defines the distribution of conformation \mathbf{R} given the molecular graph \mathcal{G} and the pairwise atom distance \mathbf{d}

$$p(\mathbf{R}|\mathbf{d}, \mathcal{G}) = \frac{1}{Z} \exp \left\{ - \sum_{e_{uv} \in \mathcal{E}} \alpha_{uv} (\|\mathbf{r}_u - \mathbf{r}_v\|_2 - d_{uv})^2 \right\}$$

- Trying to find the conformations \mathbf{R} that satisfy the distance constraints

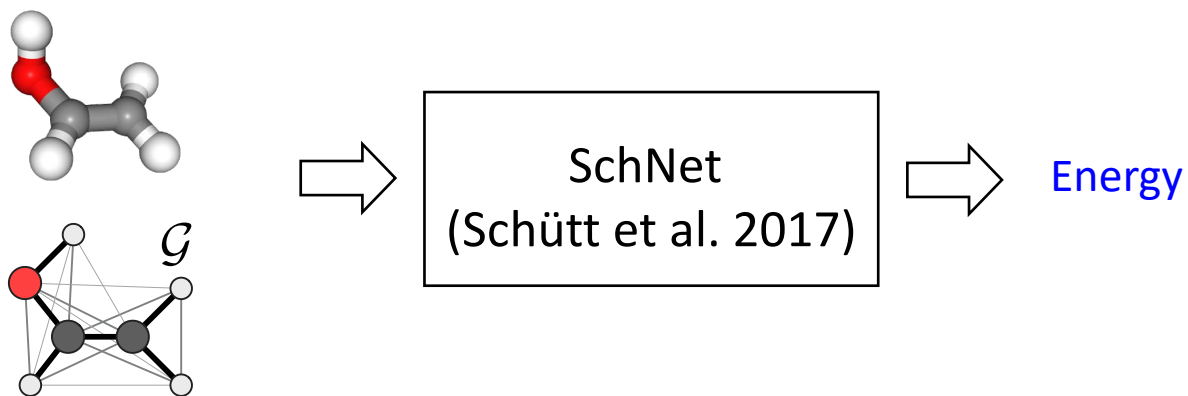


Energy-based Tilting Model

- Further correct $p_{\theta}(\mathbf{R}|\mathcal{G})$ with an energy-based tilting term $E_{\phi}(\mathbf{R}, \mathcal{G})$

$$p_{\theta, \phi}(\mathbf{R}|\mathcal{G}) \propto p_{\theta}(\mathbf{R}|\mathcal{G}) \cdot \exp(-E_{\phi}(\mathbf{R}, \mathcal{G}))$$

- Explicitly learn an energy function $E_{\phi}(\mathbf{R}, \bar{\mathcal{G}})$ with SchNet (Schütt et al. 2017)
 - Neural message passing in 3D space

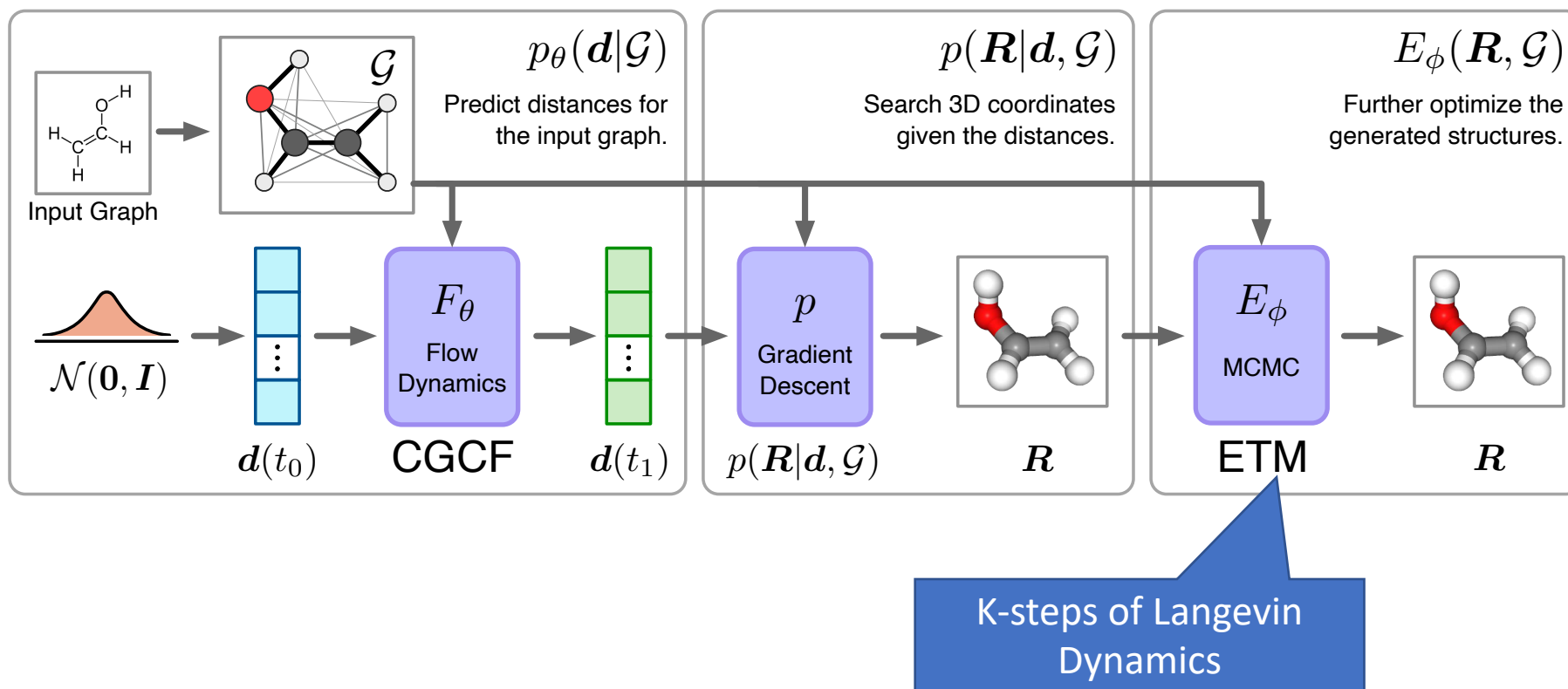


Training Energy Model

- Directly training EBMs with maximum likelihood is difficult
 - Involving a slow sampling process from the model distribution (e.g. with Langevin dynamics)
- Training EBMs with negative sampling
 - Treating observed conformations as positive examples
 - Generating negative conformations through the flow-based model $p_{\theta}(\mathbf{R}|\mathcal{G})$

$$\mathcal{L}_{\text{nce}}(\mathbf{R}, \mathcal{G}; \phi) == -\mathbb{E}_{p_{\text{data}}} \left[\log \frac{1}{1 + \exp(E_{\phi}(\mathbf{R}, \mathcal{G}))} \right] - \mathbb{E}_{p_{\theta}} \left[\log \frac{1}{1 + \exp(-E_{\phi}(\mathbf{R}, \mathcal{G}))} \right]$$

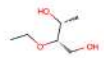










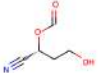











































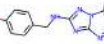










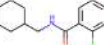










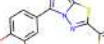










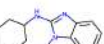










The Final Sampling Process:

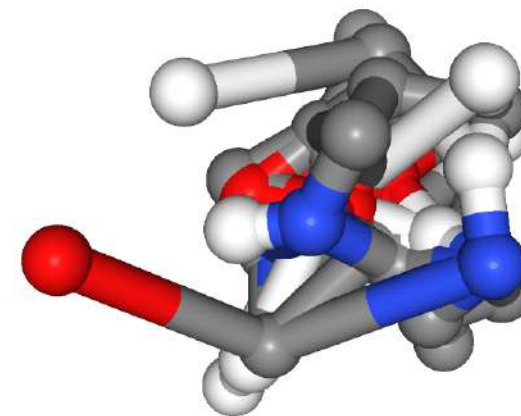
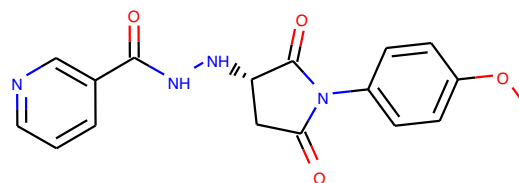
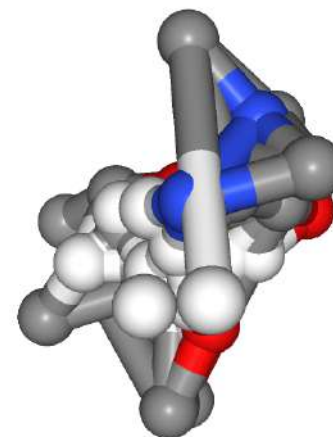
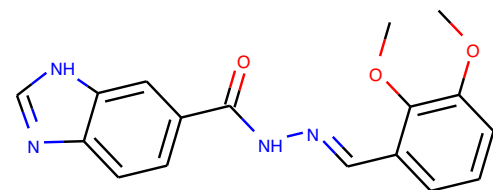


Experiments

- Data Sets
 - **GEOM**: > 33 million molecular conformers by **Rafael**'s group, including both small molecules in QM9 and medium-sized drug-like molecules.
 - **ISO17**: built on QM9, including 197 molecules, each with 5000 conformations
- Baselines
 - **CVGAE(Mansimov et al. 2019)**: learning atom representations with GNNs and then predict the coordinates of atoms
 - **GraphDG(Simm&Hernandez-Lobato, 2020)**: predicting the pairwise distances between atoms with GNNs and then generate conformers based on distances
 - **RDKit**: a classical force field in molecular dynamics

Examples

Graph	Conformations									
										
										
										
										
										
										
										
										
										



Conformation Generation

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

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

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

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

Results

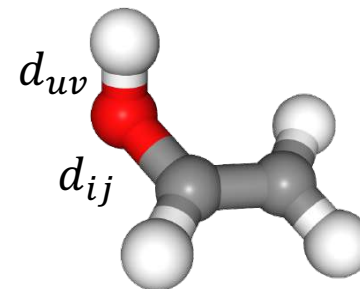
Dataset Metric	GEOM-QM9				GEOM-Drugs			
	COV* (%)		MAT (Å)		COV* (%)		MAT (Å)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
CVGAE	8.52	5.62	0.7810	0.7811	0.00	0.00	2.5225	2.4680
GraphDG	55.09	56.47	0.4649	0.4298	7.76	0.00	1.9840	2.0108
CGCF	69.60	70.64	0.3915	0.3986	49.92	41.07	1.2698	1.3064
CGCF + ETM	72.43	74.38	0.3807	0.3955	53.29	47.06	1.2392	1.2480
RDKit	79.94	87.20	0.3238	0.3195	65.43	70.00	1.0962	1.0877
CVGAE + FF	63.10	60.95	0.3939	0.4297	83.08	95.21	0.9829	0.9177
GraphDG + FF	70.67	70.82	0.4168	0.3609	84.68	93.94	0.9129	0.9090
CGCF + FF	73.52	72.75	0.3131	0.3251	92.28	98.15	0.7740	0.7338
CGCF + ETM + FF	73.54	72.58	0.3088	0.3210	92.41	98.57	0.7737	0.7616

Refined by classical
**Merck Molecular
Force Field (MMFF)**

* For the reported COV score, the threshold δ is set as 0.5Å for QM9 and 1.25Å for Drugs. More results of COV scores with different threshold δ are given in Appendix H.

Distribution over Pairwise Distances

- Evaluate the distribution of the pairwise distance between atoms for each molecular graph
 - Marginal distribution $p(d_{uv}|\mathcal{G})$
 - Pairwise distribution $p(d_{uv}, d_{ij}|\mathcal{G})$
 - Joint distribution $p(\mathbf{d}|\mathcal{G})$
- Evaluation Metrics: **maximum mean discrepancy (MMD)** between the distributions over the reference set and the generated set



Results

	Single		Pair		All	
	Mean	Median	Mean	Median	Mean	Median
RDKit	3.4513	3.1602	3.8452	3.6287	4.0866	3.7519
CVGAE	4.1789	4.1762	4.9184	5.1856	5.9747	5.9928
GraphDG	0.7645	0.2346	0.8920	0.3287	1.1949	0.5485
CGCF	0.4490	0.1786	0.5509	0.2734	0.8703	0.4447
CGCF + ETM	0.5703	0.2411	0.6901	0.3482	1.0706	0.5411

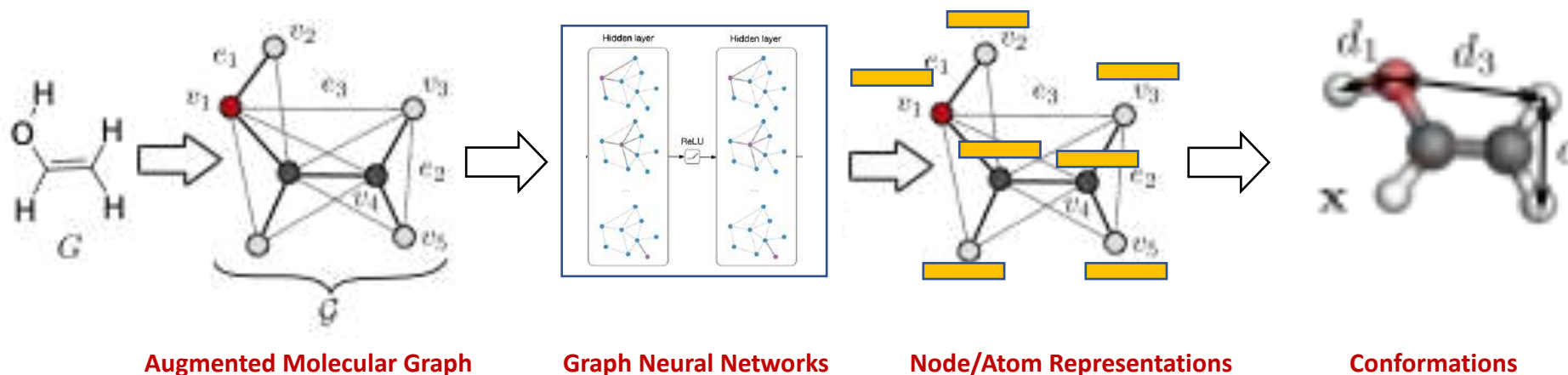
ETM slight hurts the performance as it will sharpen the distribution

Conclusion

- Molecule representations: moving from 2D graphs to 3D conformations
- Predicting molecular conformations is challenging
 - Multimodal
- A normalizing flow and energy model based framework
 - A flexible flow-based model for conformation generation
 - Energy model is further used for correcting the flow model
- Future work
 - Integrating the [physic model](#)
 - Other tasks such as [protein structure prediction](#)

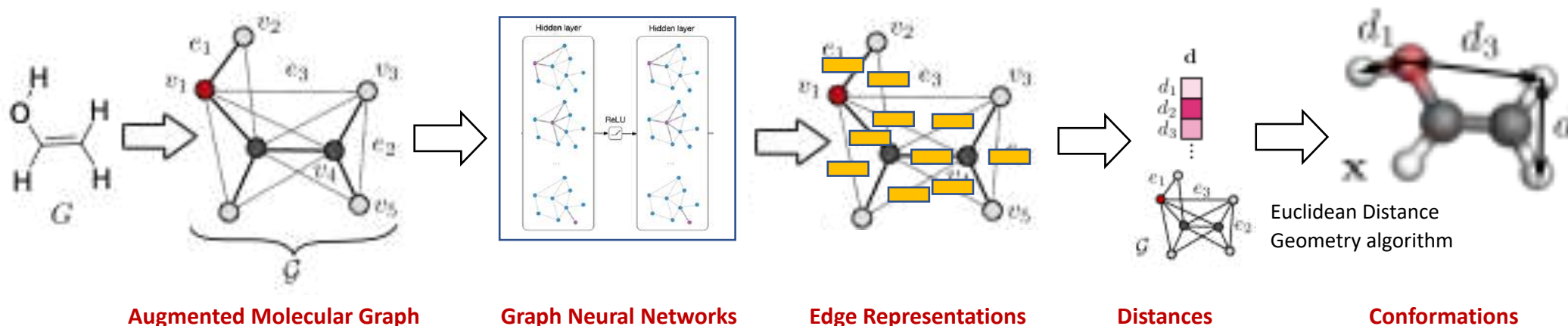
Machine Learning-based Approaches

- Train a model to predict molecular conformations \mathbf{R} given the molecular graph \mathcal{G} , i.e., modeling $p(\mathbf{R}|\mathcal{G})$
- Deep Generative Graph Neural Network (Mansimov et al. 2019)
 - Learning atom representations with graph neural networks
 - Predicting atom coordinates based on atom representations
- Limitations
 - Conformations are rotation and translation equivalent



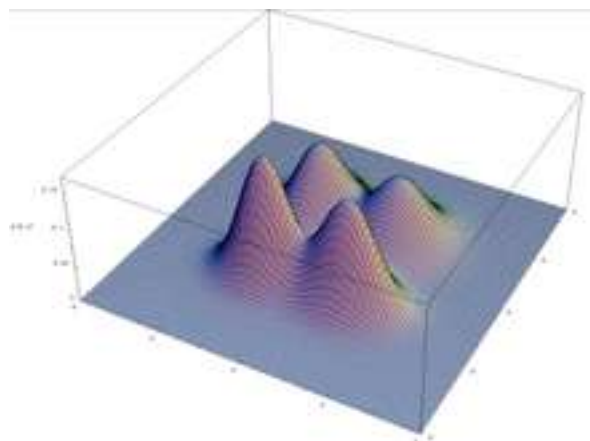
A Generative Model for Molecular Distance Geometry (Simm and Hernandez-Lobato 2020)

- Two stage generation: distance geometry generation and conformation generation
 - The distances between atoms are rotation and translation equivalent
 - Predict the conformations based on molecular graph and distances
- Distance prediction
 - Graph neural networks are used to learn the edge representations
 - Predict the edges based on edge representations



Limitations

- The model capacity is still very limited
 - The distribution $p(\mathbf{R}|\mathcal{G})$ is multimodal
 - Each molecule could have multiple stable conformations



- We need to find more flexible models!!