GEOMOL

Molecular Team Lecture Series

Dong-hee Shin

10.11.22

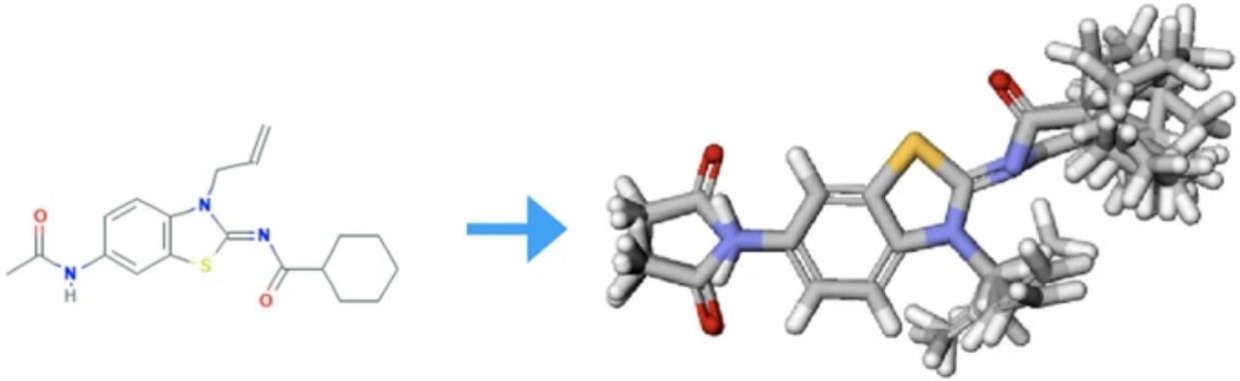
Researchers

Octavian Ganea
(Farewell, Rest in Peace)
ETH Zurich & MIT

ETH Zürich = x 32

3D Conformer Generation

- Molecules have potentially thousands of stable conformations
- **Goal: predict low energy conformers of an input molecular graph**
- More rigid and flexible parts of the 3D structure

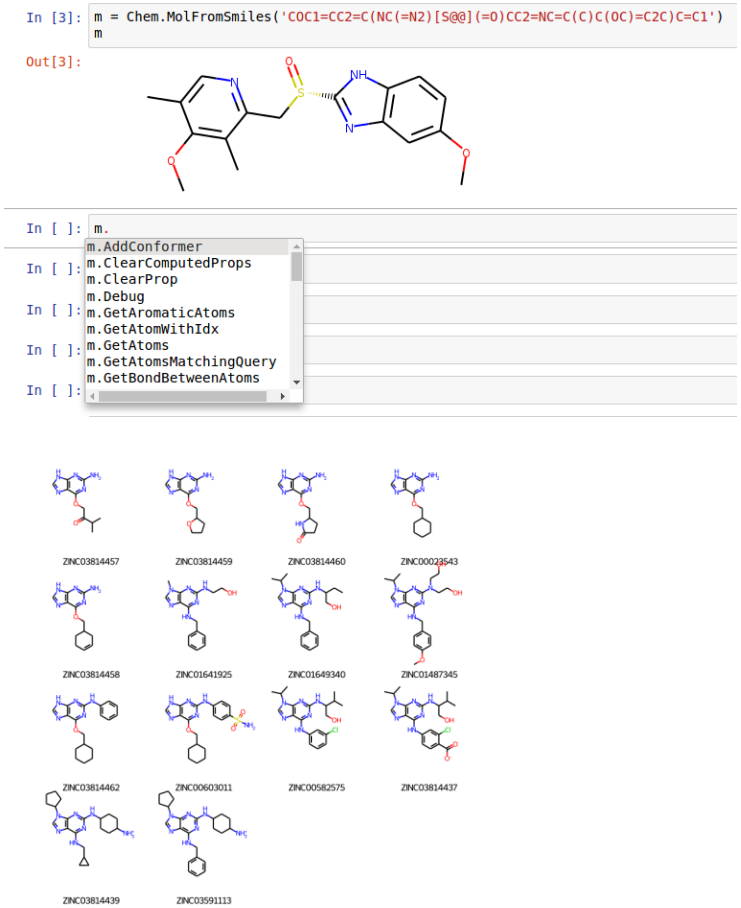


Motivation

- Faster, computationally efficient and more accurate conformer generation using MPNN
- Usable in various 3D downstream tasks
  - Protein –ligand binding
  - Molecular docking poses
  - Generating conformers inside 3D enzyme pockets
- Intermediate representation for various property predictors

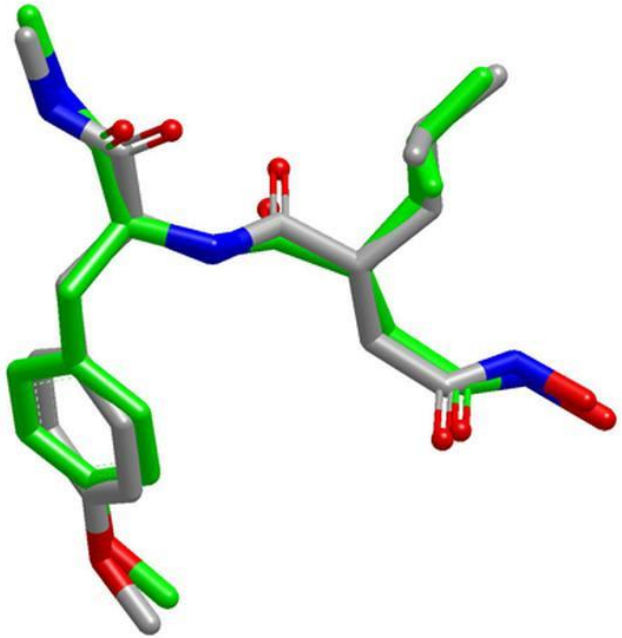
Prior Works

- **Stochastic Methods**
  - Distance geometry initialization + subsequent coordinate optimization
  - Popular open-source method: ETKDG/RDKIT
- **Drawbacks**
  - Difficult to sample diverse and representative conformers
  - Computationally expensive



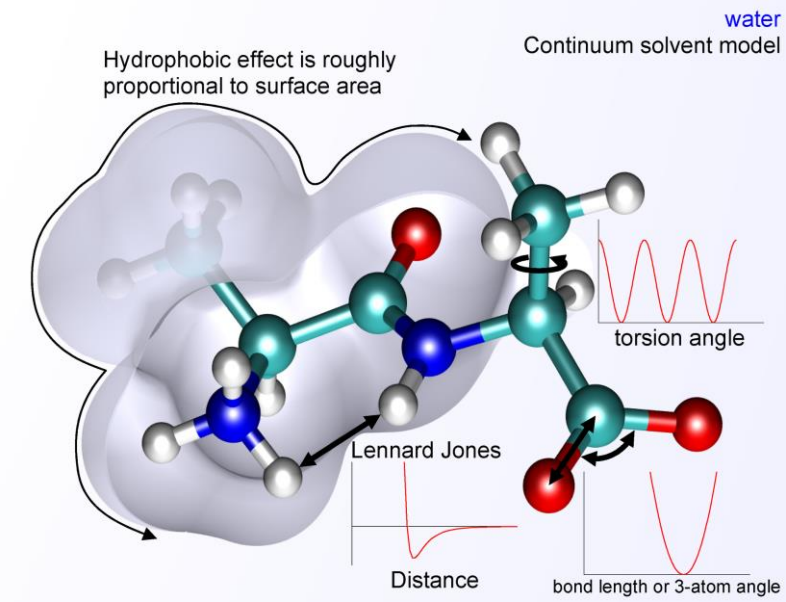
Prior Works

- **Systematic Methods**
  - Exhaustive search over torsion angles
  - Using databases of torsion templates
  - Commercial software: OMEGA
- **Drawbacks**
  - Computational prohibitive for structures with large number of rotatable bonds
  - Poor generalization to unseen structures



Prior Works

- **Fine-Tuning with Force Fields (FF)**
  - Crude approximation of the true energy
  - Experimental quantum mechanics parameters
- **Drawbacks**
  - Strong assumptions (Simplistic formulas)
  - Limitations in accurately capturing subtle, weak interactions in biomolecules



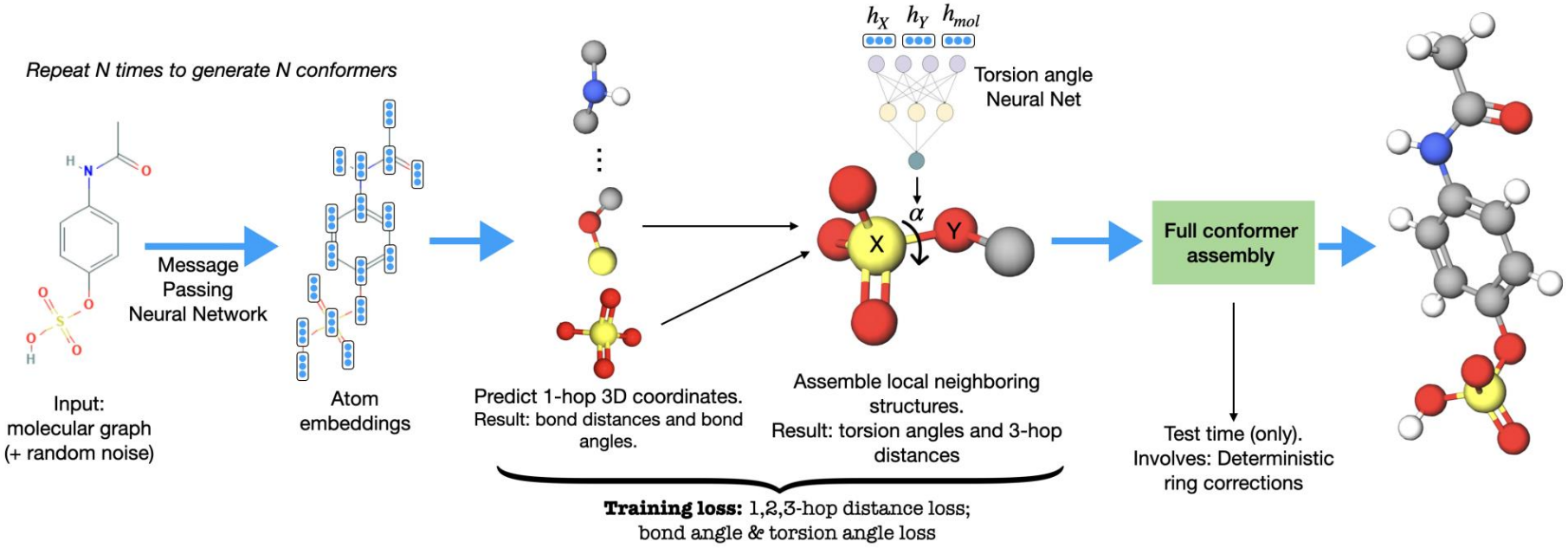
Prior Works

- **ML for Conformer Generation**
  - Multi-stage Models
  - Generate distance matrix, then predict coordinates, then fine-tune the conformer
- **Drawbacks**
  - Need a Force Field or extra energy model
  - No trainable end-to-end (meaning error accumulation)
  - No explicit handling of classic molecular geometry: bond angles, torsion angles
  - Requires an iterative procedure to sample conformers (via Langevin dynamics)

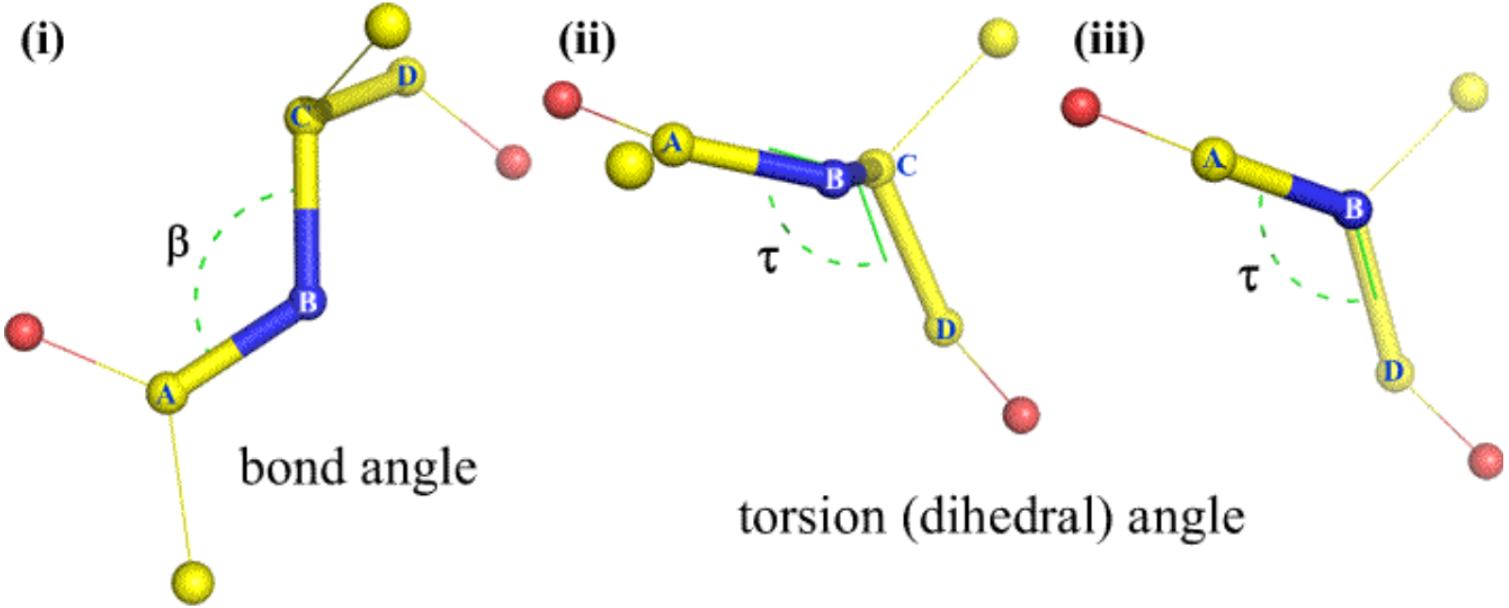
1.

Dive into GEOMOL

GEOMOL Overview



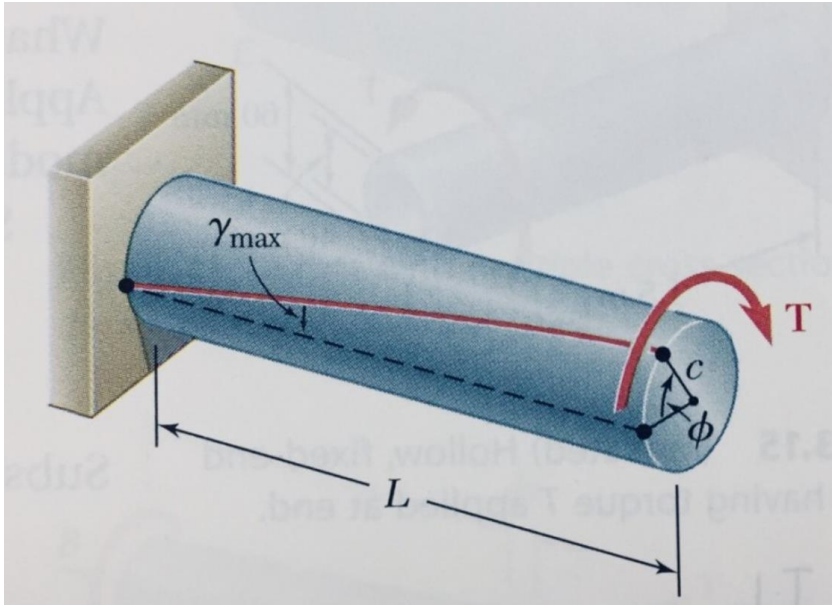
Torsion Angle



Polypeptide main chain dihedral angles: Phi (φ), Psi (ψ), and Omega (ω)

Torsion & Torque

- **Torsion:** a state of being twisted
- **Torque:** a moment that tends to twist a member about its longitudinal axis



제발
지능로봇
들으세요 !!!

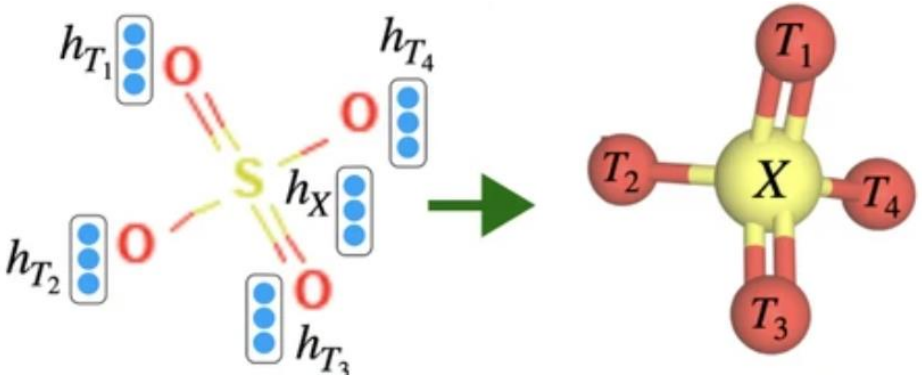
GEOMOL Contribution

- Explicit predictions of bond distances, bond angles, and *torsion angles*
- Trainable end-to-end & non-autoregressive: joint prediction of all atoms 3D coordinates from the molecular graph
- 3D coordinates predicted SE(3)-invariantly (to any global rotation/translation)
- Tetrahedral chiral centers are predicted exactly (No iterative optimization necessary as with traditional distance geometry approaches)
- Diversity of generated conformers: by using a tailored Wasserstein loss

Local Structure (LS) Prediction

- For each non-terminal atom X, predict the relative 3D coordinates of all its 1-hop neighboring assuming X is placed in the origin:

$$f(\mathbf{h}_{T_1}, \ldots, \mathbf{h}_{T_n}; \mathbf{h}_X) = (\mathbf{p}_1, \ldots, \mathbf{p}_n) \in \mathbb{R}^{3\times n}$$



Challenges

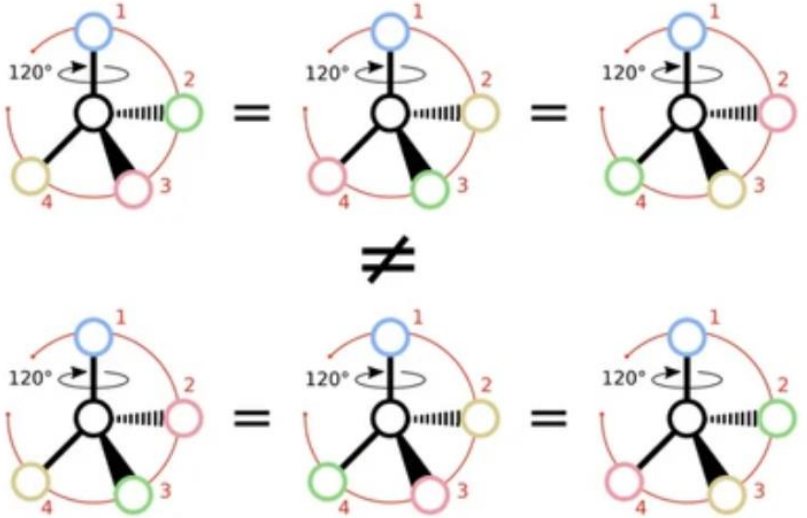
- Equivariant prediction with respect to any permutation of neighbors

$$f(\mathbf{h}_{T_{\pi(1)}}, \ldots, \mathbf{h}_{T_{\pi(n)}}; \mathbf{h}_X) = (\mathbf{p}_{\pi(1)}, \ldots, \mathbf{p}_{\pi(n)}), \forall \pi \in S_n$$

- Bond distance should match symmetrically (XT_1 should have the same length predicted from the LS of X or from the LS of T_1)
  - Solution: a special symmetric transformer encoder that separates distance prediction from direction prediction
- Should explicitly address chirality

Tackling Chirality

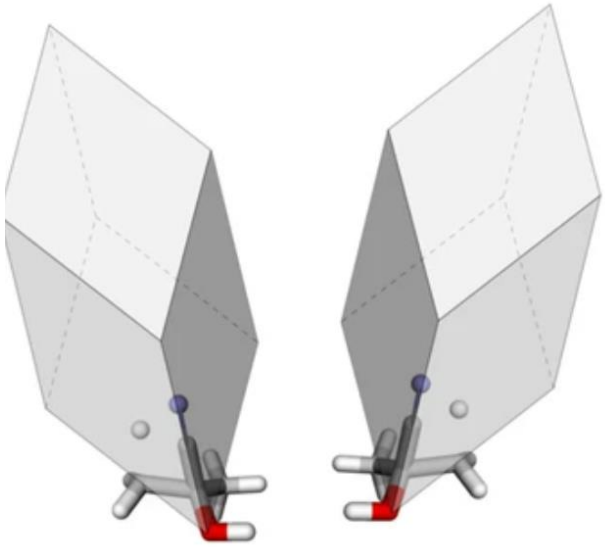
- Chiral Information
  - Bond annotations to describe different molecules with the same molecular graph, but different 3D structures (thus different chemical behaviors
  - Differentiates mirroring structures
  - Bond annotations are not fixed (multiple equivalent annotations)



Tackling Chirality Exactly

- Given a chiral center, we compute the oriented volume
- The sign of the oriented volume changes depending on chirality
- If we get the incorrect sign, we simply reflect the structure by flipping against the z-axis
- No iterative optimization is needed

$$OV(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) \stackrel{\text{def}}{=} sign \left(\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} \right)$$



$OV(C_S) = -1$ $OV(C_R) = +1$

Assembling LS via Torsion Angle

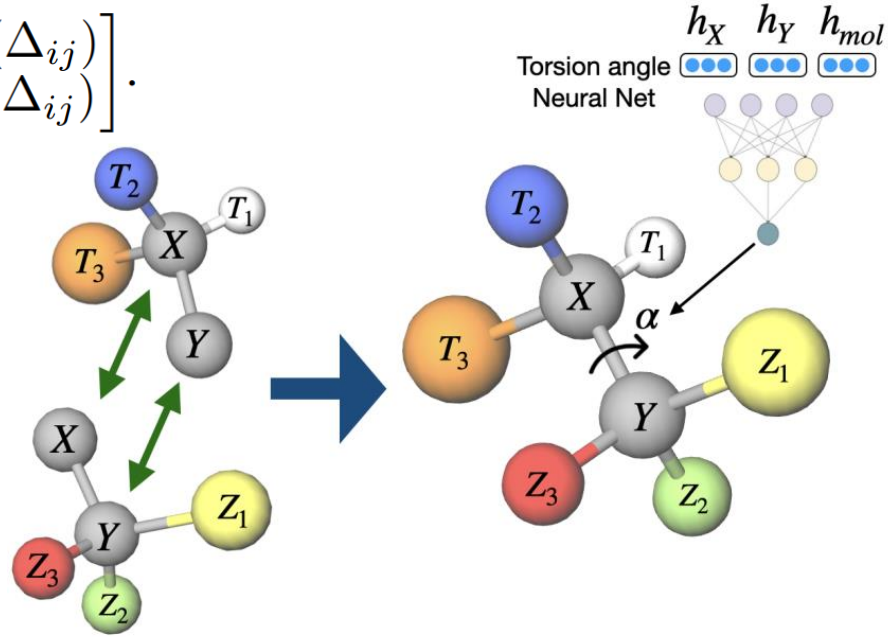
- Assemble every two local neighboring structures by predicting the torsion angle
- **Challenges:**
  - Parameterize a single canonical torsion angle per rotatable bond
    - All dihedral angle (XYT, XYZ) are couped via a single canonical torsion when LS of X and Y are fixed
  - Torsion angle should be predicted in a *rotation-translation and permutation invariant*
- **Solution:** Novel Torsion Angle Neural Network

Assembling LS via Torsion Angle

$$\Delta_{ij} \overset{\text{def}}{=} \angle(XYT_i, XYZ_j) \text{ and } \mathbf{s}_{ij} \overset{\text{def}}{=} \begin{bmatrix} \cos(\Delta_{ij}) \\ \sin(\Delta_{ij}) \end{bmatrix}.$$

torsion angle as $\alpha \overset{\text{def}}{=} atan2(\frac{\mathbf{s}}{\|\mathbf{s}\|})$.

satisfies both invariances



Proposition 1. *Given 3D coordinates of nodes X, Y, T_i, Z_j and fixed weights $c_{ij} \in \mathbb{R}$ such that $\sum_{i,j} c_{ij}\mathbf{s}_{ij} \in \mathbb{R}^2$ is not the null vector, then $\alpha \overset{\text{def}}{=} atan2(\frac{\mathbf{s}}{\|\mathbf{s}\|})$ is unique, i.e., if we change the torsion angle of bond XY, then α will change. Formally, if we rotate the set of bonds $\{XT_i\}_i$ jointly around the line XY with the same angle γ, then α will be exactly shifted with γ.*

Optimal Transport Loss

- We predict a single conformer *C* and then calculates a loss $\mathcal{L}(\mathcal{C}, \mathcal{C}^*)$
  - matches 1,2,3-hop distances, bond and torsion angles
- In practice, multiple ground truth conformers $\{\mathcal{C}_l^*\}_{l\in[1..L]}$ and predicted $\{\mathcal{C}_k\}_{k\in[1..K]}$
  - However, we do not know a priori the number *L* of true conformers or the matching between generated and true conformers
  - We wish to avoid expensive and problematic adversarial training
  - How to generate diverse conformers (to cover all modes of true distributions)
- Solution: Optimal-Transport based loss function

$$\mathcal{L}^{ensemble} \stackrel{\text{def}}{=} EMD_{\mathcal{L}(\cdot,\cdot)}(\{\mathcal{C}_k\}_k, \{\mathcal{C}_l^*\}_l) = \min_{\mathbf{T}\in\mathcal{Q}_{K,L}} \sum_{k,l} T_{kl}\mathcal{L}(\mathcal{C}_k, \mathcal{C}_l^*)$$

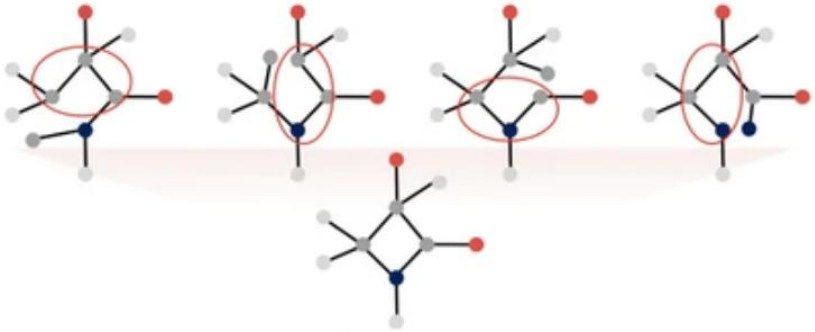
$\mathbf{T}$ is the ***transport plan***

Total Loss

$$\mathcal{L}(\mathcal{C},\mathcal{C}^*) \stackrel{\text{def}}{=} \xi_1 \cdot \frac{1}{\#\{(u,v)\in E\}} \sum_{\{(u,v)\in E\}} (d(u,v)-d^*(u,v))^2$$
$$+\xi_2 \cdot \frac{1}{\#\{u,v : \text{2-hops away}\}} \sum_{\{u,v:\text{2-hops away}\}} (d(u,v)-d^*(u,v))^2$$
$$+\xi_3 \cdot \frac{1}{\#\{u,v : \text{3-hops away}\}} \sum_{\{u,v:\text{3-hops away}\}} (d(u,v)-d^*(u,v))^2$$
$$-\xi_4 \cdot \frac{1}{\#(u,v)\in E,(v,w)\in E} \sum_{(u,v)\in E,(v,w)\in E} \cos\left(\angle uvw - \angle^* uvw\right)$$
$$-\xi_5 \cdot \frac{1}{\#(u,v),(v,w),(w,y)\in E} \sum_{(u,v),(v,w),(w,y)\in E} \cos\left(\angle(uvwy) - \angle^*(uvwy)\right)$$

Assemble Full Conformer at Test Time

- We can assemble any tree-like molecule using predicted local structures and torsion angles



- We correct rings by averaging over all ring spanning trees and using *Kabsch superimposition algorithm*
  - *Kabsch algorithm* is method for calculating the optimal rotation matrix that minimizes the RMSD between two paired sets of points (very useful method for comparing molecular/protein structures)

Results

Table 1: Results on the **GEOM-DRUGS** dataset. All models are without FF fine-tuning. "R" and "P" denote Recall and Precision. Note: OMEGA is an established commercial (C) software.

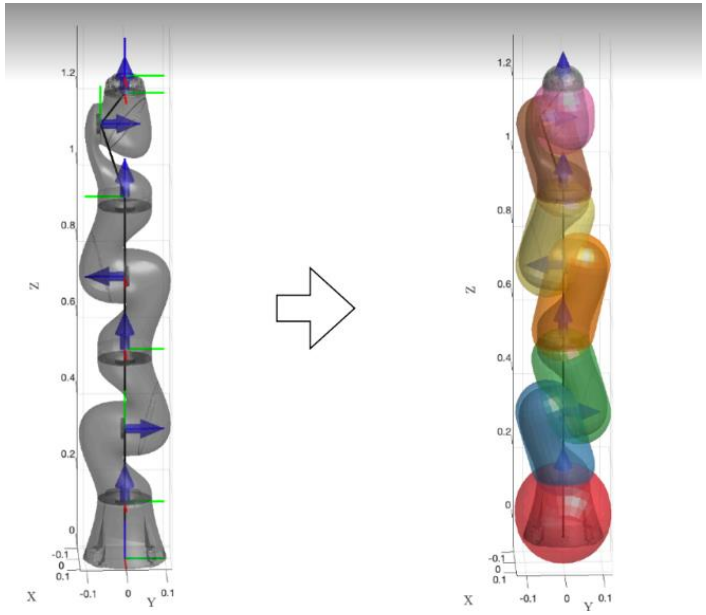
Models	COV - R (%) ↑		AMR - R (Å) ↓		COV - P (%) ↑		AMR - P (Å) ↓	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GraphDG *(ML)*	10.37	0.00	1.950	1.933	3.98	0.00	2.420	2.420
CGCF *(ML)*	54.35	56.74	1.248	1.224	24.48	15.00	1.837	1.829
RDKit/ETKDG	68.78	76.04	1.042	0.982	71.06	88.24	1.036	0.943
OMEGA *(C)*	81.64	97.25	0.851	**0.771**	77.18	**96.15**	0.951	**0.854**
GeoMol ($s = 9.5$)	**86.07**	**98.06**	**0.846**	0.820	71.78	83.77	1.039	0.982
GeoMol ($s = 5$)	82.43	95.10	0.862	0.837	**78.52**	94.40	**0.933**	**0.856**

Table 2: Results on the **GEOM-QM9** dataset. See caption of table 1.

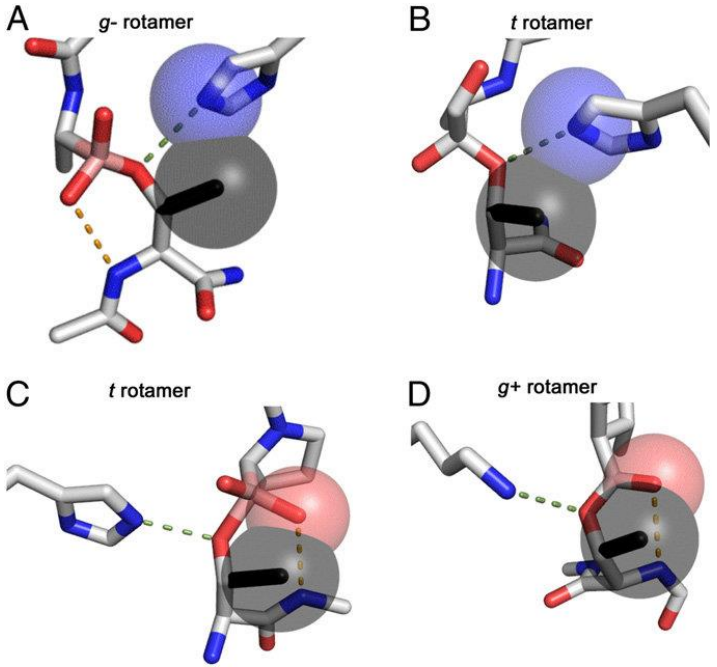
Models	COV - R (%) ↑		AMR - R (Å) ↓		COV - P (%) ↑		AMR - P (Å) ↓	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
GraphDG *(ML)*	74.66	100.00	0.373	0.337	63.03	77.60	0.450	0.404
CGCF *(ML)*	69.47	96.15	0.425	0.374	38.20	33.33	0.711	0.695
RDKit/ETKDG	85.13	**100.00**	0.235	0.199	**86.80**	**100.00**	0.232	0.205
OMEGA *(C)*	85.51	**100.00**	**0.177**	**0.126**	82.86	**100.00**	**0.224**	**0.186**
GeoMol ($s = 5$)	**91.52**	**100.00**	0.225	0.193	**86.71**	**100.00**	0.270	0.241

Limitations

- Weakness in capturing some long-range interactions expecially of sturctures that are scarce in the train set (e.g., macrocycles)
- Steric Clashes
- Large Rings



In robotics, they check self-collision by using joint capsules



ISMB D-Day 101

Thank You