A Plug-and-Play Quaternion Message-Passing Module for Molecular Conformation Representation

Angxiao Yue, Dixin Luo, Hongteng Xu

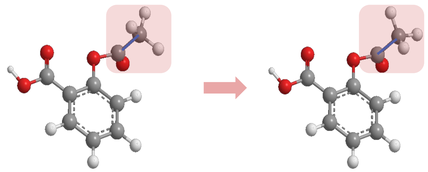
Renmin University of China & Beijing Institute of Technology

AAAI Association for the Advancement of Artificial Intelligence

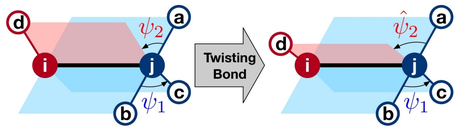
Motivation

What is molecular modeling

- 3D graph neural networks (GNNs)
  1. $G = (\mathcal{V}, \mathcal{E}, H^{(0)}, Z^{(0)}, X)$
  2. $\mathcal{V}$: atoms, $\mathcal{E}$: edges, $H^{(0)}$: atom features, $Z^{(0)}$: bond features, X: coordinates.
  3. Equivariant GNNs: coordinates, orientations, relative position…
  4. Invariant GNNs: bond lengths, angles, dihedrals…
- Conformation



Challenges for conformation representation



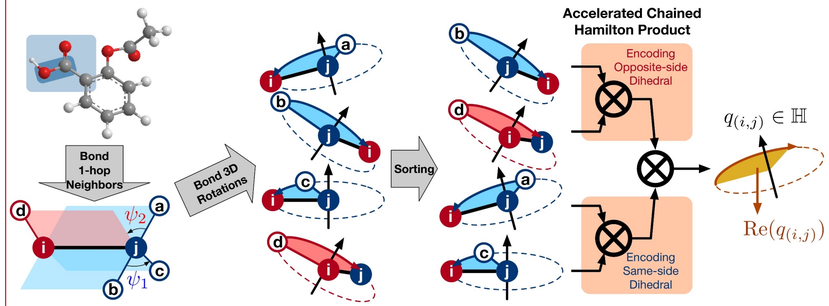
- High computational complexity for complete representation.
  1. $\boldsymbol{\psi_1}$: same-side dihedral, $\boldsymbol{\psi_2}$: opposite-side dihedral.
  2. A molecule with N atoms and D degrees on average: bond lengths $O(ND)$, angles $O(ND^2)$, , dihedrals $O(ND^2)$ for $\boldsymbol{\psi_1}$ and $\boldsymbol{O(ND^3)}$ for $\boldsymbol{\psi_2}$!!!

Our Contributions

- We propose a **plug-and-play** quaternion message-passing (**QMP**) module to improve invariant GNNs in molecular conformation representation and analysis tasks.
  1. **QMP** achieves a **quaternion-based** mechanism to encode $\boldsymbol{\psi_1}$ and $\boldsymbol{\psi_2}$.
  2. **QMP** is a **plug-and-play** design with small computational cost and one-line code.
  3. Computational complexity of **QMP** is $\boldsymbol{O(NDlogD)}$.

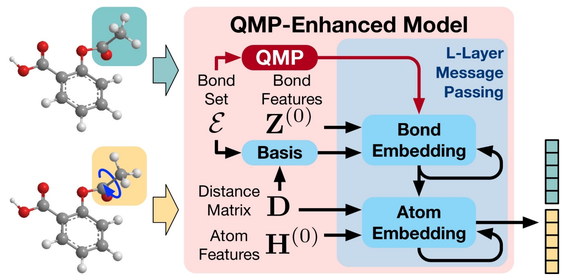
Our Approach

Overview



- High computational complexity for complete representation.
  1. Encode the 3D rotations as a sequence of quaternions.
  2. Aggregate the rotations by the chained Hamilton product of the quaternions.
  3. Plug the real part into existing invariant GNNs.

Plug-and-Play



- Given a graph $G(\mathcal{V}, \mathcal{E}, H^{(0)}, Z^{(0)}, X)$, in the $\boldsymbol{l}$-th message passing layer:

$$z_{ij}^{(l+1)} = f_e\left(e_{ij}\right),$$
$$h_i^{(l+1)} = f_v(h_i^{(l)}, \sum_{j \in \mathcal{N}_i} z_{ij}^{(l+1)})$$

- For DimeNet, DimeNet++, SphereNet

$$e_{ij} = z_{ij}^{(l)} \oplus \sum_{k \in \mathcal{N}_{i \setminus j}} f_a(z_{ik}^{(l)}, b_D, \Theta, \Phi_s || Re\,(q_{(i,j)}))$$

- For SchNet:

$$e_{ij} = f_s(h_j^{(l)}) \odot f_{cf}(b_D) || Re(q_{(i,j)})$$

Experiment

- MD17 & MD17@CCSD

Dataset	Molecule	sGDML	DimeNet	SchNet	DimeNet++	SphereNet	Q-SchNet	Q-DimeNet++	Q-SphereNet
MD17	Aspirin	0.68	0.499	1.339	0.325	0.400	1.289↑	**0.316**↑	0.356↑
	Benzene	0.20	0.187	0.346	0.168	0.193	0.316↑	**0.151**↑	0.177↑
	Ethanol	0.33	0.230	0.738	0.150	0.181	0.468↑	**0.148**↑	0.169↑
	Malonaldehyde	0.41	0.383	1.559	0.263	0.379	1.540↑	**0.241**↑	0.321↑
	Naphthalene	0.11	0.215	0.723	**0.100**	0.159	0.521↑	0.106	0.138↑
	Salicylic acid	0.28	0.374	1.001	**0.231**	0.261	0.971↑	0.238	0.321
	Toluene	0.14	0.216	0.747	0.117	0.142	0.550↑	**0.107**↑	0.136↑
	Uracil	0.24	0.301	1.351	0.189	0.228	1.010↑	**0.178**↑	0.241
MD17@CCSD	Aspirin	-	-	1.471	0.387	0.458	1.298↑	**0.365**↑	0.475
	Benzene	-	-	0.352	0.049	0.061	0.287↑	**0.048**↑	0.063
	Ethanol	-	-	1.321	0.129	0.158	0.786↑	**0.108**↑	0.148↑
	Malonaldehyde	-	-	1.531	**0.211**	0.289	1.086↑	0.231	0.264↑
	Toluene	-	-	0.855	0.133	0.160	0.637↑	0.128↑	**0.052**↑

Table 1: Results in terms of force MAE ($\frac{\text{kcal}}{\text{molÅ}}$). The best results are in bold. The second best results are with underlines. ↑ indicates the performance is improved. The results of sGDML and DimeNet are quoted from (Liu et al. 2022).

- OC20

Metrics		CGCNN	SchNet	DimeNet	DimeNet++	Q-SchNet	Q-DimeNet	Q-DimeNet++
Energy MAE	ID	0.9773	1.0480	0.9314	0.9261	1.0363↑	**0.9132**↑	0.9177↑
	OOD Ads	0.9818	1.0450	1.0720	0.9400	1.0640	0.9865↑	**0.9295** ↑
	OOD Cat	0.9269	1.0630	0.8945	0.8828	1.0370 ↑	**0.8753**↑	0.8816↑
	OOD Both	0.8828	1.0076	0.9643	0.8572	0.9646 ↑	0.9317↑	**0.8513**↑
	Average	0.9422	1.0409	0.9656	0.9015	1.0255 ↑	0.9267↑	**0.8950**↑
EwT (%)	ID	1.84	1.58	1.80	1.79	1.59↑	1.84↑	**1.91**↑
	OOD Ads	1.72	1.53	1.62	1.67	1.52	**1.85**↑	1.79↑
	OOD Cat	1.93	1.56	1.93	1.81	1.46	1.96↑	**2.13**↑
	OOD Both	1.69	1.48	1.57	**1.80**	1.61↑	1.70↑	1.75↑
	Average	1.80	1.54	1.73	1.77	1.55↑	1.84↑	**1.90**↑

Table 2: Results on OC20 IS2RE task in terms of energy MAE (eV) and the percentage of EwT of the ground truth energy. Performance is reported for models trained on the 10k training dataset.

- Ablation Study

QMP Components			Force MAE	
j-side	i-side	sorting	Aspirin	Benzene
✓	X	✓	0.323	0.152
X	✓	✓	0.360	0.166
✓	✓	X	0.390	0.173
✓	✓	✓	**0.316**	**0.151**

Table 3: Effects of ablating components of QMP.

- Potentials to Equivariant GNNs

Method	SE(3) Trans.	TFN	EGNN	Q-EGNN
MSE	2.440	1.550	0.724	**0.668**

Table 4: Mean Squared Error ($\times 10^{-2}$) for the future position estimation in the N-body system experiment.

Code: https://github.com/AngxiaoYue/QMP

Future work

- We would like to utilize the imaginary part of **QMP** to enhance equivariant GNNs.
- We plan to design more hypercomplex neural networks for molecular modeling based on Clifford Algebra

Reference

[1] Liu M, Luo Y, Wang L, et al. DIG: A turnkey library for diving into graph deep learning research[J]. Journal of Machine Learning Research, 2021, 22(240): 1-9.
[2] Chanussot L, Das A, Goyal S, et al. Open catalyst 2020 (OC20) dataset and community challenges[J]. Acs Catalysis, 2021, 11(10): 6059-6072.