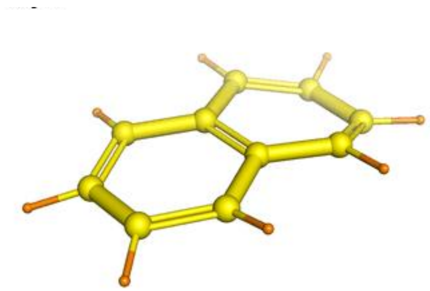
Energy-Motivated Equivariant Pretraining for 3D Molecular Graphs

Rui Jiao, Jiaqi Han, Wenbing Huang, Yu Rong, Yang Liu

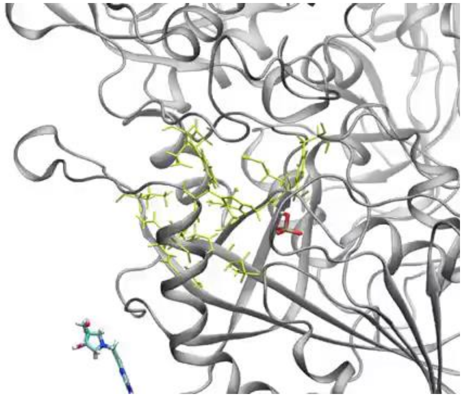
AAAI 2023

Motivation

- Learning informative molecular representation is fundamental for various downstream applications.



MD Simulation

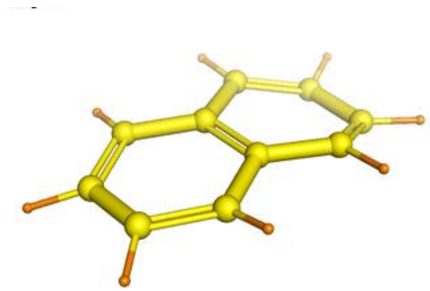


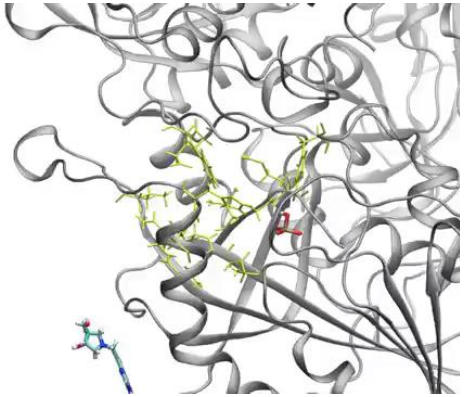
Ligand Binding

- Equivariant Graph networks (EGNs) are capable to model geometric graphs.
  - SchNet, EGNN, PaiNN, TorchMD-Net, e.t.c.
- Community has provided large-scale molecule datasets with rich 3D conformations.
  - GEOM,Molecule3D, PCQM4Mv2, e.t.c.
- Many self-supervised works have shown superiority in 2D graph learning

Motivation

- Learning informative molecular representation is fundamental for various downstream applications.





3D tasks

MD Simulation

Ligand Binding

- Equivariant Graph networks (EGNs) are capable to model geometric graphs.
  - SchNet, EGNN, PaiNN, TorchMD-Net, e.t.c.

3D models

- Community has provided large-scale molecule datasets with rich 3D conformations.
  - GEOM,Molecule3D, PCQM4Mv2, e.t.c.

3D datasets

- Many self-supervised works have shown superiority in 2D graph learning.

2D pretraining

- How about 3D pretraining?

Related Works

- Equivariant Graph Neural Networks
  - Take EGNN as an example
  - E(3) Symmetry

    Invariance $\varphi(g \cdot \boldsymbol{X}, H) = \varphi(\boldsymbol{X}, H)$

    Equivariance $\varphi(g \cdot \boldsymbol{X}, H) = g \cdot \varphi(\boldsymbol{X}, H)$
  - EGNN Message Passing & Aggregration

$$
\begin{aligned}
m_{ij} &= \varphi_m\left(h_i, h_j, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2, e_{ij}\right), \\
\boldsymbol{x}_i' &= \boldsymbol{x}_i + \sum_{j \neq i} (\boldsymbol{x}_i - \boldsymbol{x}_j)\, \varphi_x(m_{ij}), \\
h_i' &= \varphi_h(h_i, \sum_{j \in \mathcal{N}(i)} m_{ij}),
\end{aligned}
$$

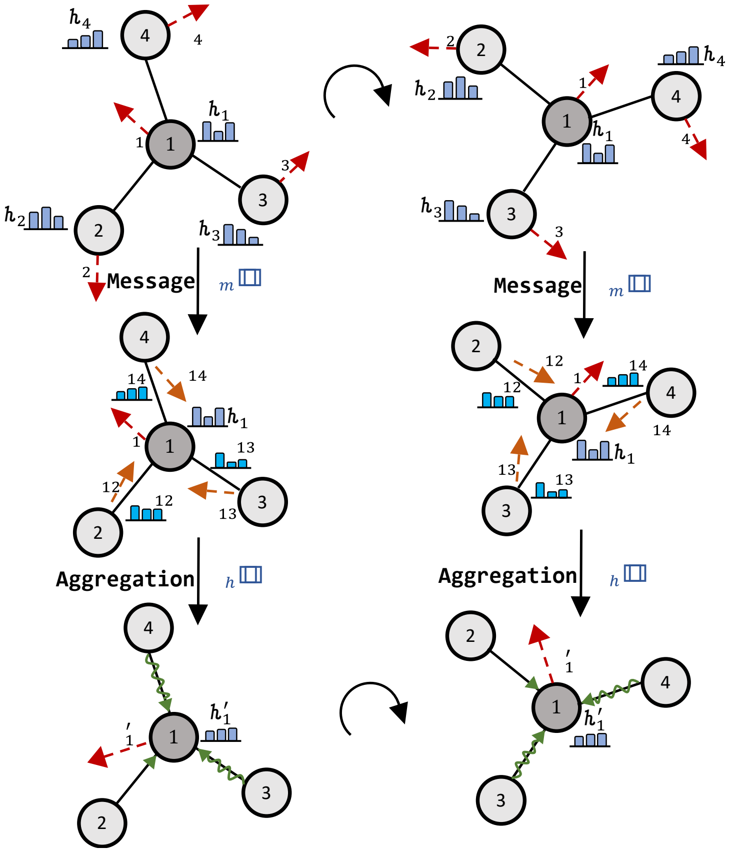


Image taken from Han et. al., 2022

Related Works

- Self-supervised Molecular Pretraining
- 2D Pretraining
  - Contrastive : Maximize MI between different views
    - InfoGraph, GraphCL, JOAO, e.t.c.
  - Generative : Reconstruct the graph components from different levels
    - AttrMask, EdgePred, GPT-GNN, e.t.c.
  - Predictive : Predict domain-specific labels created from the input graphs
    - GROVER, e.t.c.
- 2D + 3D Pretraining
  - GraphMVP
  - 3D Infomax

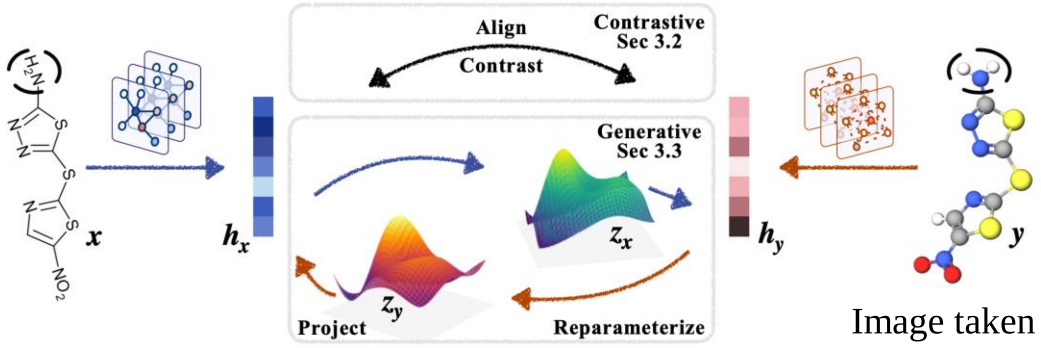
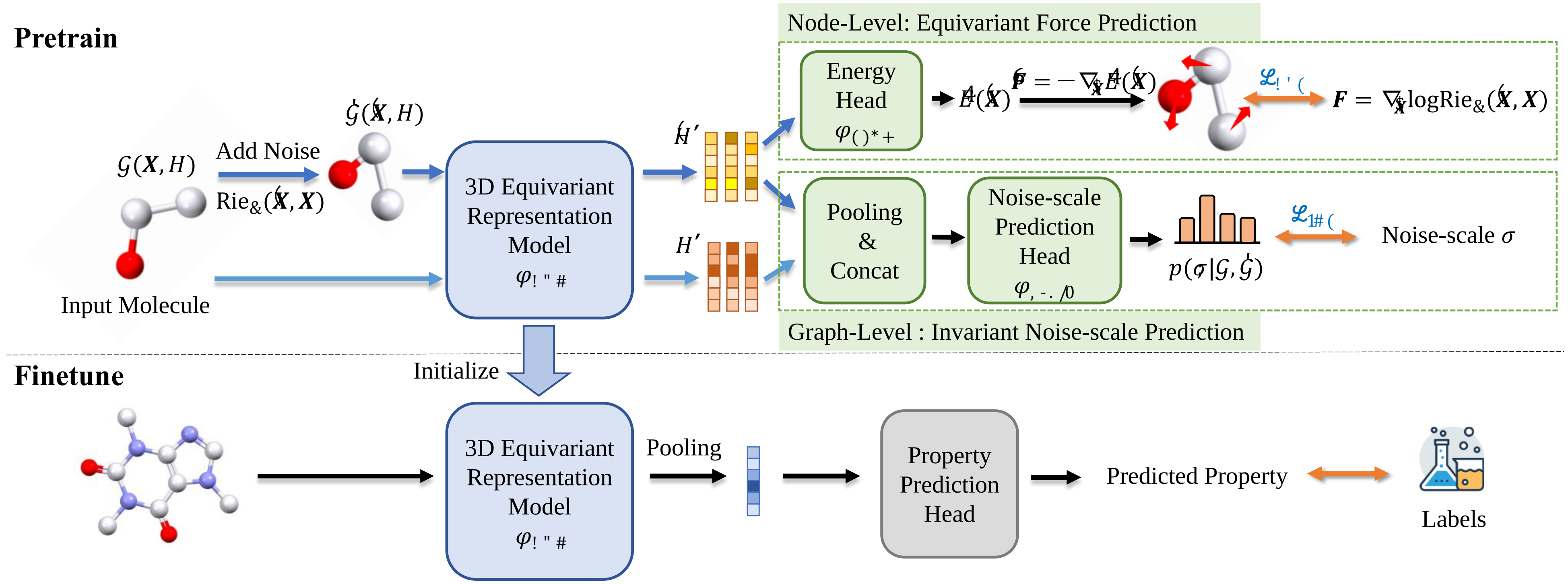


Image taken from Liu et. al., 2022

- Does 2D pretraining methods align well with 3D models?
- Can we design 3D-aware tasks for 3D graphs?

Methods

- Overview
  - Node-level : Equivariant Force Prediction (EFP)
  - Graph-level : Invariant Noise-scale Prediction (INP)



Pretrain
Input Molecule
$\mathcal{G}(\boldsymbol{X}, H)$
Add Noise
$\mathrm{Rie}_{\&}(\tilde{\boldsymbol{X}}, \boldsymbol{X})$
$\tilde{\mathcal{G}}(\tilde{\boldsymbol{X}}, H)$
3D Equivariant Representation Model
$\varphi_{!\,"\,\#}$
$\tilde{H}'$
H'
Node-Level: Equivariant Force Prediction
Energy Head
$\varphi_{()*+}$
$\tilde{E}(\tilde{\boldsymbol{X}})$
$\tilde{\boldsymbol{F}} = -\nabla_{\tilde{\boldsymbol{X}}} \tilde{E}(\tilde{\boldsymbol{X}})$
$\mathcal{L}_{!\,'\,(}$
$\boldsymbol{F} = \nabla_{\tilde{\boldsymbol{X}}} \log \mathrm{Rie}_{\&}(\tilde{\boldsymbol{X}}, \boldsymbol{X})$
Pooling & Concat
Noise-scale Prediction Head
$\varphi_{,\,-.\,/0}$
$p(\sigma | \mathcal{G}, \tilde{\mathcal{G}})$
$\mathcal{L}_{1\#\,(}$
Noise-scale σ
Graph-Level : Invariant Noise-scale Prediction
Initialize
Finetune
3D Equivariant Representation Model
$\varphi_{!\,"\,\#}$
Pooling
Property Prediction Head
Predicted Property
Labels

Methods

- Energy-based Molecular Modeling
- 3D graph : atom representations H + inter-atomic connections $\mathcal{E}$ + 3D positions $\boldsymbol{X}$
- Obtain the node-level representation from the EGN model

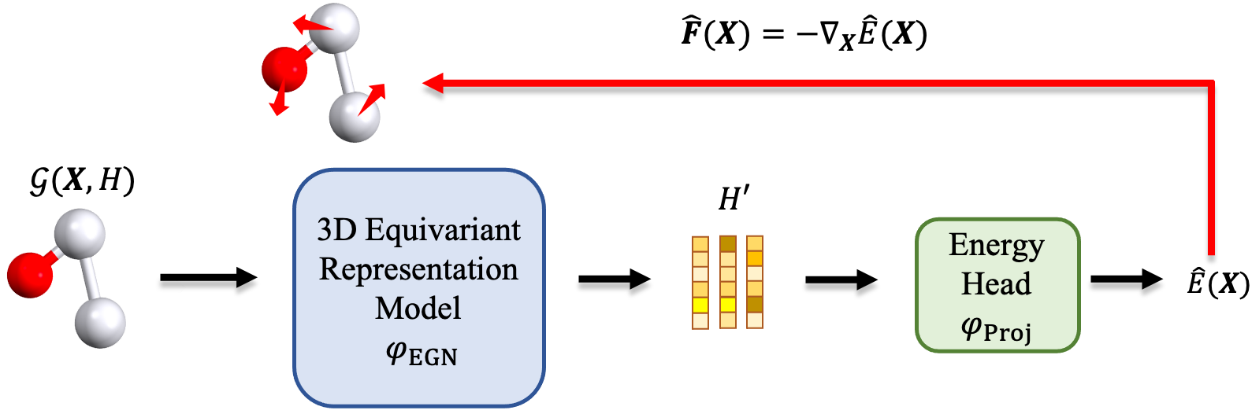
$$H' = \varphi_{\mathrm{EGN}}(\boldsymbol{X}, H, \mathcal{E})$$

- Predict the graph-level energy via a pooling operation

$$\hat{E}(\boldsymbol{X}) = \varphi_{\mathrm{proj}}\left(\sum_{i=1}^{N} h_i'\right)$$

- Forces decrease the potential energy

$$\widehat{\boldsymbol{F}}(\boldsymbol{X}) = -\nabla_{\boldsymbol{X}} \hat{E}(\boldsymbol{X})$$



$\widehat{\boldsymbol{F}}(\boldsymbol{X}) = -\nabla_{\boldsymbol{X}} \hat{E}(\boldsymbol{X})$

$\mathcal{G}(\boldsymbol{X}, H)$

3D Equivariant Representation Model φ_{EGN}

H'

Energy Head φ_{Proj}

$\hat{E}(\boldsymbol{X})$

Methods

- How to fit the predicted forces with reasonable “labels”?
- Assume the training conformers obey a Boltzmann energy distribution

$$p\left(\boldsymbol{X}\right) = \frac{1}{Z}\exp\left(-\frac{E(\boldsymbol{X})}{kT}\right)$$

- Forces are the negative gradients of the energy E over the coordinates $\boldsymbol{X}$

$$\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X}) \propto -\nabla_{\boldsymbol{X}} E(\boldsymbol{X}) := \boldsymbol{F}.$$

Methods

- How to fit the predicted forces with reasonable “labels”?
- Assume the training conformers obey a Boltzmann energy distribution

$$p\left(\boldsymbol{X}\right) = \frac{1}{Z}\exp\left(-\frac{E(\boldsymbol{X})}{kT}\right)$$

- Forces are the negative gradients of the energy E over the coordinates $\boldsymbol{X}$

$$\boxed{\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X})} \propto -\nabla_{\boldsymbol{X}} E(\boldsymbol{X}) := \boldsymbol{F}.$$

Force “labels” !

$$\mathcal{L}_{\text{EFP}} = \mathbb{E}_{\mathcal{G}\sim\mathbb{G}}\left[\|\hat{\boldsymbol{F}}(\boldsymbol{X}) - \nabla_{\boldsymbol{X}} \log p(\boldsymbol{X})\|_F^2\right]$$

Methods

- How to fit the predicted forces with reasonable “labels”?
- Assume the training conformers obey a Boltzmann energy distribution

$$p\left(\boldsymbol{X}\right) = \frac{1}{Z} \exp\left(-\frac{E(\boldsymbol{X})}{kT}\right)$$

- Forces are the negative gradients of the energy *E* over the coordinates ***X***

$$\boxed{\nabla_{\boldsymbol{X}} \log p(\boldsymbol{X})} \propto -\nabla_{\boldsymbol{X}} E(\boldsymbol{X}) := \boldsymbol{F}.$$

Force “labels” !

- Denoise score matching

$$\mathcal{L}_{\text{EFP-DN}} = \mathbb{E}_{\mathcal{G}\sim\mathbb{G}, \tilde{\boldsymbol{X}}\sim p(\tilde{\boldsymbol{X}}\mid\boldsymbol{X})}\left[\|\hat{\boldsymbol{F}}(\tilde{\boldsymbol{X}}) - \nabla_{\tilde{\boldsymbol{X}}} \log p(\tilde{\boldsymbol{X}} \mid \boldsymbol{X})\|_F^2\right]$$

Methods

- Doubly E(3)-invariance

$$p(g_1 \cdot \tilde{\boldsymbol{X}} \mid g_2 \cdot \boldsymbol{X}) = p(\tilde{\boldsymbol{X}} \mid \boldsymbol{X}), \forall g_1, g_2 \in \mathrm{E}(3)$$

- Our proposed Riemann-Gaussian distribution

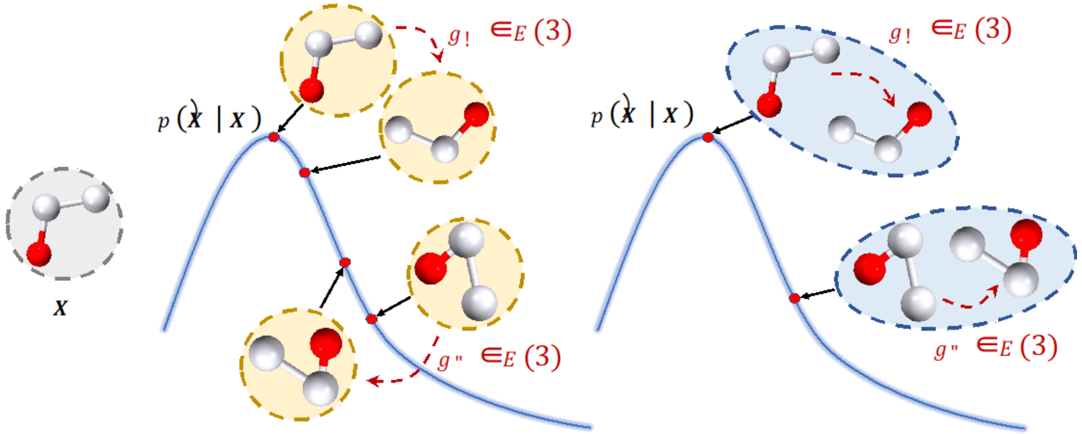
$$p_\sigma(\tilde{\boldsymbol{X}} \mid \boldsymbol{X}) = \mathrm{Rie}_\sigma(\tilde{\boldsymbol{X}} \mid \boldsymbol{X}) := \frac{1}{Z(\sigma)} \exp\left(-\frac{d^2(\tilde{\boldsymbol{X}}, \boldsymbol{X})}{4\sigma^2}\right)$$

- Let $\boldsymbol{Y} = \boldsymbol{X} - \mu(\boldsymbol{X})$ be the zero-mean positions, the distance kernel is determined on the inner-product matrix to satisfy the doubly E(3)-invariance

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \|\boldsymbol{Y}_1^\top \boldsymbol{Y}_1 - \boldsymbol{Y}_2^\top \boldsymbol{Y}_2\|_F$$

- Targeted gradients

$$\nabla_{\tilde{\boldsymbol{X}}} \log p_\sigma(\tilde{\boldsymbol{X}} | \boldsymbol{X}) = -\frac{1}{\sigma^2}[(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top)\tilde{\boldsymbol{Y}} - (\tilde{\boldsymbol{Y}}\boldsymbol{Y}^\top)\boldsymbol{Y}]$$



(a) Gaussian

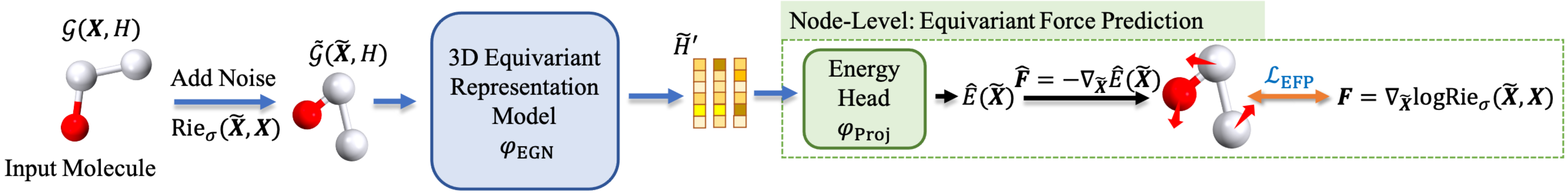
(b) Riemann-Gaussian

Methods

- Final Equivariant Force Prediction loss

$$\mathcal{L}_{\text{EFP-Final}} = \mathbb{E}_{\mathcal{G}\sim\mathbb{G}, l\sim U(1,L), \tilde{\boldsymbol{X}}\sim p_{\sigma_l}(\tilde{\boldsymbol{X}}|\boldsymbol{X})} \left[\sigma_l^2 \left\| \frac{1}{\sigma_l}\hat{\boldsymbol{F}}(\tilde{\boldsymbol{X}}) - \frac{1}{\alpha}\nabla_{\tilde{\boldsymbol{X}}} \log p_{\sigma_l}(\tilde{\boldsymbol{X}}|\boldsymbol{X})\right\|_F^2\right]$$

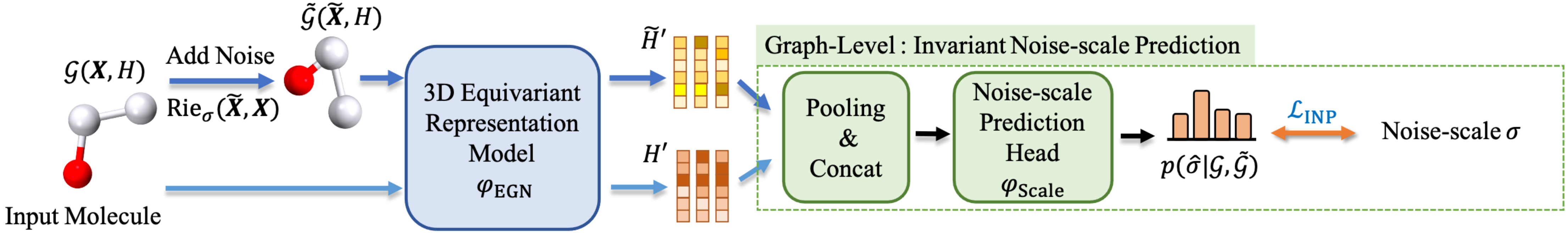
- $\alpha = \left(\left\|\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\right\|_F + \left\|\tilde{\boldsymbol{Y}}\boldsymbol{Y}^\top\right\|_F\right)/2$ for numerical stability


$\mathcal{G}(\boldsymbol{X}, H)$
Input Molecule
Add Noise
$\text{Rie}_\sigma(\tilde{\boldsymbol{X}}, \boldsymbol{X})$
$\tilde{\mathcal{G}}(\tilde{\boldsymbol{X}}, H)$
3D Equivariant Representation Model φ_{EGN}
$\tilde{H}'$
Node-Level: Equivariant Force Prediction
Energy Head φ_{Proj}
$\hat{E}(\tilde{\boldsymbol{X}})$
$\hat{\boldsymbol{F}} = -\nabla_{\tilde{\boldsymbol{X}}}\hat{E}(\tilde{\boldsymbol{X}})$
$\mathcal{L}_{\text{EFP}}$
$\boldsymbol{F} = \nabla_{\tilde{\boldsymbol{X}}}\text{logRie}_\sigma(\tilde{\boldsymbol{X}}, \boldsymbol{X})$

Methods

- Graph-level Invariant Noise-scale Prediction
- Discriminate the noise-scale given the original and perturbed graph
- Let $H', \widetilde{H}'$ be the output node representation of the original and perturbed graph
- The predicted probability is $\boldsymbol{p} \in \mathbb{R}^L = \varphi_{Scale}(\sum_{i=1}^{N} h_i', \sum_{i=1}^{N} \tilde{h}_i')$
- INP loss

$$\mathcal{L}_{\text{INP}} = \mathbb{E}_{\mathcal{G}\sim\mathbb{G}, l\sim U(1,L), \tilde{\boldsymbol{X}}\sim p_{\sigma_l}(\tilde{\boldsymbol{X}}|\boldsymbol{X})}\left[\mathcal{L}_{\text{CE}}\left(\mathbb{I}[l], \boldsymbol{p}\right)\right]$$

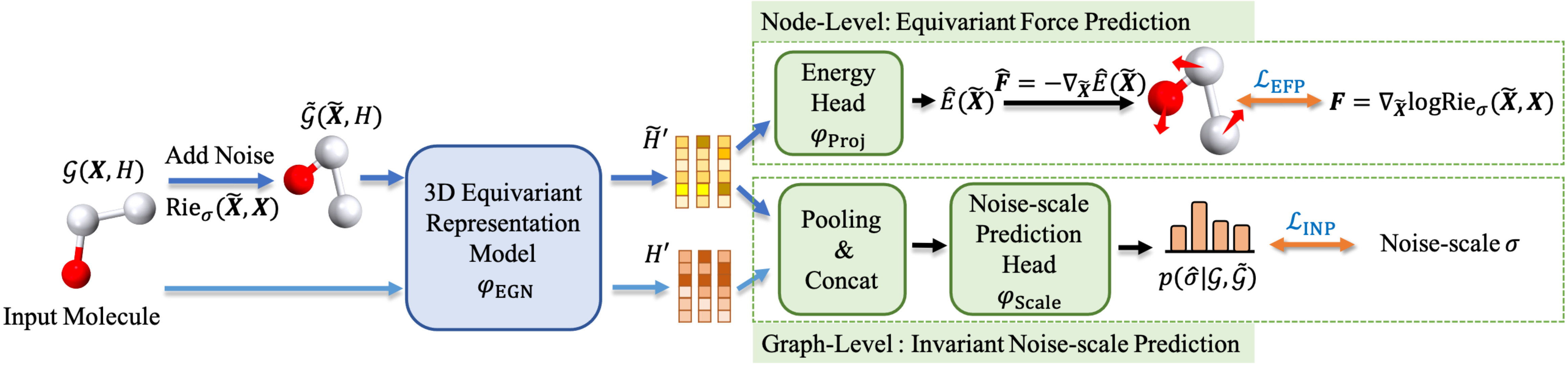


$\mathcal{G}(\boldsymbol{X}, H)$
Add Noise
$\text{Rie}_\sigma(\widetilde{\boldsymbol{X}}, \boldsymbol{X})$
$\tilde{\mathcal{G}}(\widetilde{\boldsymbol{X}}, H)$
Input Molecule
3D Equivariant Representation Model φ_{EGN}
$\widetilde{H}'$
H'
Graph-Level : Invariant Noise-scale Prediction
Pooling & Concat
Noise-scale Prediction Head φ_{Scale}
$p(\hat{\sigma}|\mathcal{G}, \tilde{\mathcal{G}})$
$\mathcal{L}_{\text{INP}}$
Noise-scale σ

Methods

- Combination of the two tasks

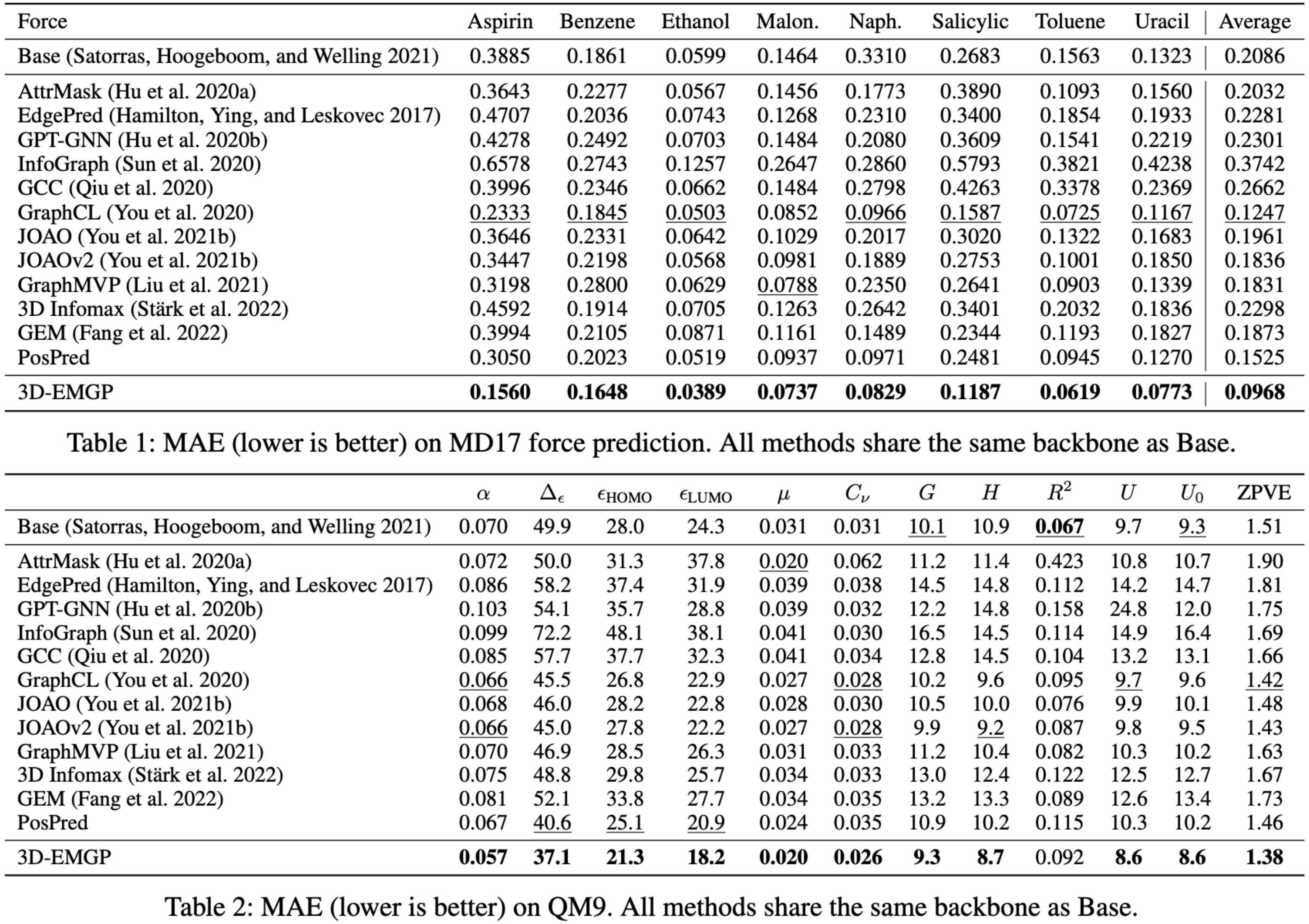
$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{EFP-Final}} + \lambda_2 \mathcal{L}_{\text{INP}}$$


Node-Level: Equivariant Force Prediction
$\mathcal{G}(\boldsymbol{X}, H)$
Input Molecule
Add Noise
$\text{Rie}_\sigma(\widetilde{\boldsymbol{X}}, \boldsymbol{X})$
$\tilde{\mathcal{G}}(\widetilde{\boldsymbol{X}}, H)$
3D Equivariant Representation Model
φ_{EGN}
$\widetilde{H}'$
H'
Energy Head
φ_{Proj}
$\hat{E}(\widetilde{\boldsymbol{X}})$
$\widehat{\boldsymbol{F}} = -\nabla_{\widetilde{\boldsymbol{X}}} \hat{E}(\widetilde{\boldsymbol{X}})$
$\mathcal{L}_{\text{EFP}}$
$\boldsymbol{F} = \nabla_{\widetilde{\boldsymbol{X}}} \log \text{Rie}_\sigma(\widetilde{\boldsymbol{X}}, \boldsymbol{X})$
Pooling & Concat
Noise-scale Prediction Head
φ_{Scale}
$p(\hat{\sigma} | \mathcal{G}, \tilde{\mathcal{G}})$
$\mathcal{L}_{\text{INP}}$
Noise-scale σ
Graph-Level : Invariant Noise-scale Prediction

Experiments

- Datasets
- Pretraining
  - 100,000 conformations from GEOM
  - Filter molecules in downstream task
- Downstream tasks
  - QM9, 100k/18k/13k for training/validation/testing
  - MD17, 9500/500/others for training/validation/testing
- Backbone model
  - EGNN (main), SchNet, TorchMD-Net (analysis)
- Baselines
  - 2D pretraining
    - Contrastive : InfoGraph, GCC, GraphCL, JOAO, JOAOv2
    - Generative : AttrMask, EdgePred, GPT-GNN
  - 2D + 3D pretraining
    - GraphMVP, 3D Infomax
  - 3D pretraining
    - ChemRL-GEM, PosPred

Experiments



Force	Aspirin	Benzene	Ethanol	Malon.	Naph.	Salicylic	Toluene	Uracil	Average
Base (Satorras, Hoogeboom, and Welling 2021)	0.3885	0.1861	0.0599	0.1464	0.3310	0.2683	0.1563	0.1323	0.2086
AttrMask (Hu et al. 2020a)	0.3643	0.2277	0.0567	0.1456	0.1773	0.3890	0.1093	0.1560	0.2032
EdgePred (Hamilton, Ying, and Leskovec 2017)	0.4707	0.2036	0.0743	0.1268	0.2310	0.3400	0.1854	0.1933	0.2281
GPT-GNN (Hu et al. 2020b)	0.4278	0.2492	0.0703	0.1484	0.2080	0.3609	0.1541	0.2219	0.2301
InfoGraph (Sun et al. 2020)	0.6578	0.2743	0.1257	0.2647	0.2860	0.5793	0.3821	0.4238	0.3742
GCC (Qiu et al. 2020)	0.3996	0.2346	0.0662	0.1484	0.2798	0.4263	0.3378	0.2369	0.2662
GraphCL (You et al. 2020)	0.2333	0.1845	0.0503	0.0852	0.0966	0.1587	0.0725	0.1167	0.1247
JOAO (You et al. 2021b)	0.3646	0.2331	0.0642	0.1029	0.2017	0.3020	0.1322	0.1683	0.1961
JOAOv2 (You et al. 2021b)	0.3447	0.2198	0.0568	0.0981	0.1889	0.2753	0.1001	0.1850	0.1836
GraphMVP (Liu et al. 2021)	0.3198	0.2800	0.0629	0.0788	0.2350	0.2641	0.0903	0.1339	0.1831
3D Infomax (Stärk et al. 2022)	0.4592	0.1914	0.0705	0.1263	0.2642	0.3401	0.2032	0.1836	0.2298
GEM (Fang et al. 2022)	0.3994	0.2105	0.0871	0.1161	0.1489	0.2344	0.1193	0.1827	0.1873
PosPred	0.3050	0.2023	0.0519	0.0937	0.0971	0.2481	0.0945	0.1270	0.1525
3D-EMGP	**0.1560**	**0.1648**	**0.0389**	**0.0737**	**0.0829**	**0.1187**	**0.0619**	**0.0773**	**0.0968**

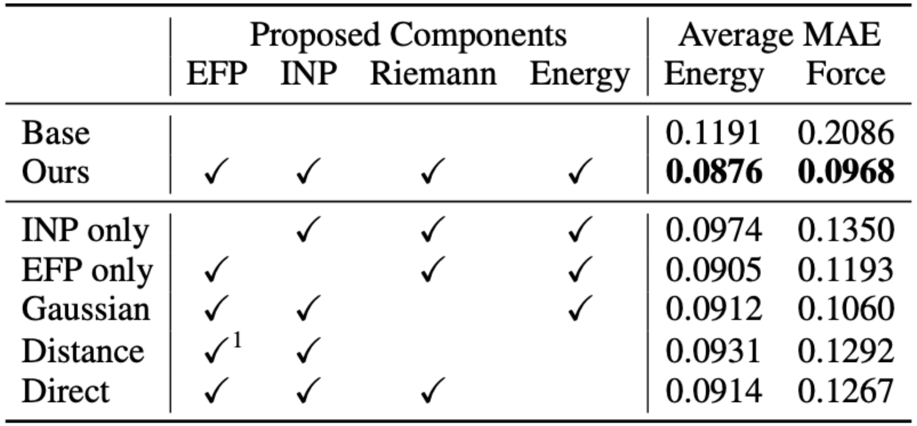
Table 1: MAE (lower is better) on MD17 force prediction. All methods share the same backbone as Base.

	α	Δ_ϵ	ϵ_{HOMO}	ϵ_{LUMO}	μ	C_ν	G	H	R^2	U	U_0	ZPVE
Base (Satorras, Hoogeboom, and Welling 2021)	0.070	49.9	28.0	24.3	0.031	0.031	10.1	10.9	**0.067**	9.7	9.3	1.51
AttrMask (Hu et al. 2020a)	0.072	50.0	31.3	37.8	0.020	0.062	11.2	11.4	0.423	10.8	10.7	1.90
EdgePred (Hamilton, Ying, and Leskovec 2017)	0.086	58.2	37.4	31.9	0.039	0.038	14.5	14.8	0.112	14.2	14.7	1.81
GPT-GNN (Hu et al. 2020b)	0.103	54.1	35.7	28.8	0.039	0.032	12.2	14.8	0.158	24.8	12.0	1.75
InfoGraph (Sun et al. 2020)	0.099	72.2	48.1	38.1	0.041	0.030	16.5	14.5	0.114	14.9	16.4	1.69
GCC (Qiu et al. 2020)	0.085	57.7	37.7	32.3	0.041	0.034	12.8	14.5	0.104	13.2	13.1	1.66
GraphCL (You et al. 2020)	0.066	45.5	26.8	22.9	0.027	0.028	10.2	9.6	0.095	9.7	9.6	1.42
JOAO (You et al. 2021b)	0.068	46.0	28.2	22.8	0.028	0.030	10.5	10.0	0.076	9.9	10.1	1.48
JOAOv2 (You et al. 2021b)	0.066	45.0	27.8	22.2	0.027	0.028	9.9	9.2	0.087	9.8	9.5	1.43
GraphMVP (Liu et al. 2021)	0.070	46.9	28.5	26.3	0.031	0.033	11.2	10.4	0.082	10.3	10.2	1.63
3D Infomax (Stärk et al. 2022)	0.075	48.8	29.8	25.7	0.034	0.033	13.0	12.4	0.122	12.5	12.7	1.67
GEM (Fang et al. 2022)	0.081	52.1	33.8	27.7	0.034	0.035	13.2	13.3	0.089	12.6	13.4	1.73
PosPred	0.067	40.6	25.1	20.9	0.024	0.035	10.9	10.2	0.115	10.3	10.2	1.46
3D-EMGP	**0.057**	**37.1**	**21.3**	**18.2**	**0.020**	**0.026**	**9.3**	**8.7**	0.092	**8.6**	**8.6**	**1.38**

Table 2: MAE (lower is better) on QM9. All methods share the same backbone as Base.

Experiments

- **Ablation studies on each components**



	Proposed Components				Average MAE	
	EFP	INP	Riemann	Energy	Energy	Force
Base					0.1191	0.2086
Ours	✓	✓	✓	✓	**0.0876**	**0.0968**
INP only		✓	✓	✓	0.0974	0.1350
EFP only	✓		✓	✓	0.0905	0.1193
Gaussian	✓	✓		✓	0.0912	0.1060
Distance	✓[1]	✓			0.0931	0.1292
Direct	✓	✓	✓		0.0914	0.1267

Table 3: Ablation studies on MD17. [1]Denoising on distance.

- **Generalization on different backbones**

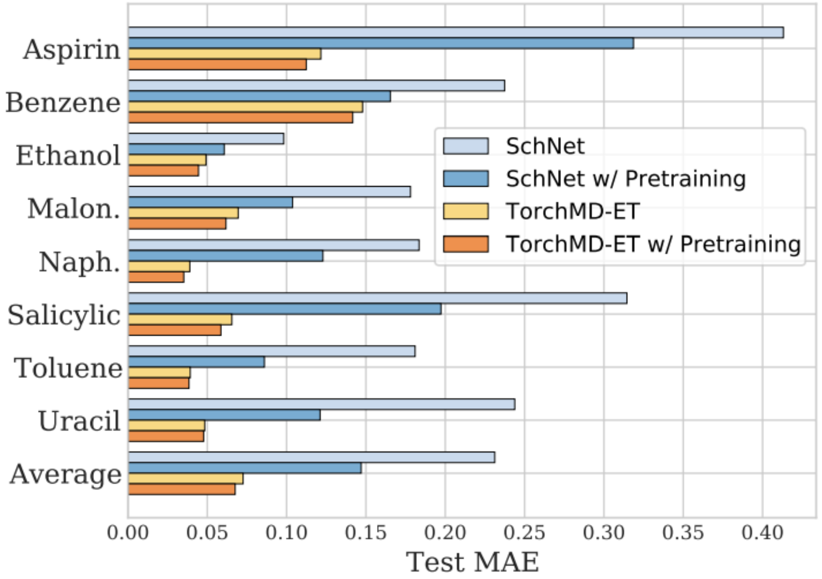


Figure 3: MAE on MD17 with different backbones.

Experiments

- **Energy landscape visualization**

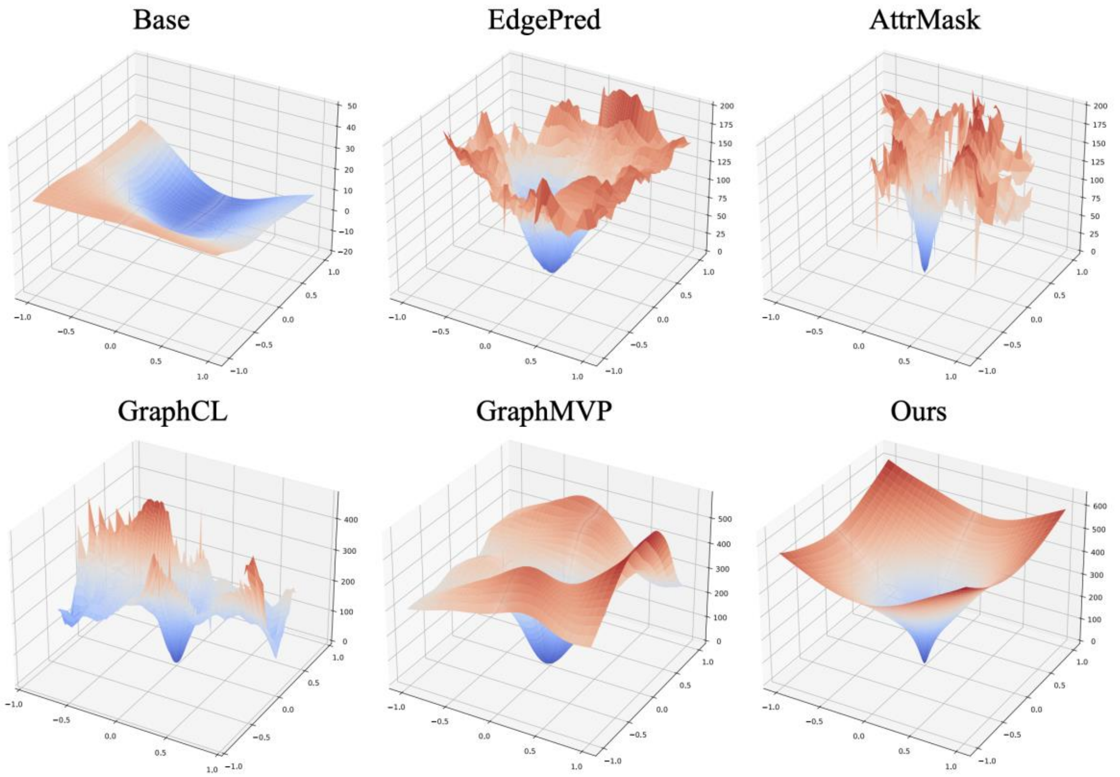


Figure 4: Energy landscape of different pretrained models.

Conclusion

- We propose a general 3D pretraining framework
    - Node-level equivariant force prediction via energy-based modeling and Riemann-Gaussian distribution
    - Graph-level invariant noise-scale prediction as a surrogate task
- We conduct experiments on QM9 and MD17, showcasing the superiority of our method
- We provide necessary analyses and visualizations to verify and explain the effectiveness of our method

Thanks！