

Sliced Denoising (Slide): Physics-Informed Molecular Pre-Training

Yuyan Ni, Shikun Feng, Weiying Ma, Zhiming Ma, Yanyan Lan



AMSS

Academy of Mathematics and Systems Science, CAS



AIR

清华大学 智能产业研究院

Institute for AI Industry Research, Tsinghua University



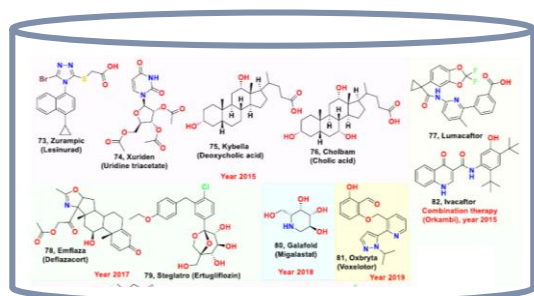
中国科学院大学

University of Chinese Academy of Sciences

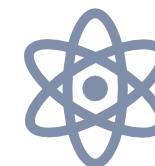
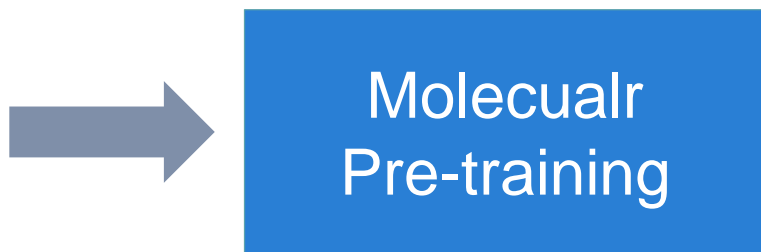
BAAI

北京智源人工智能研究院
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

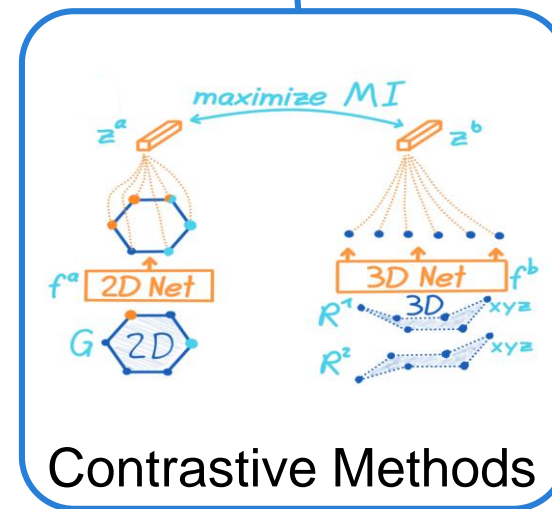
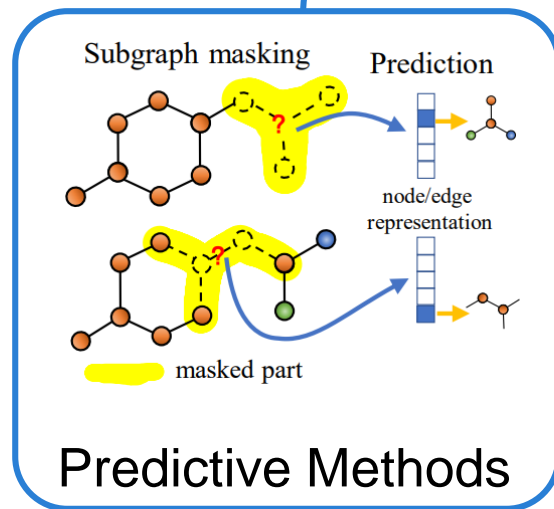
Background: Molecular Pre-training



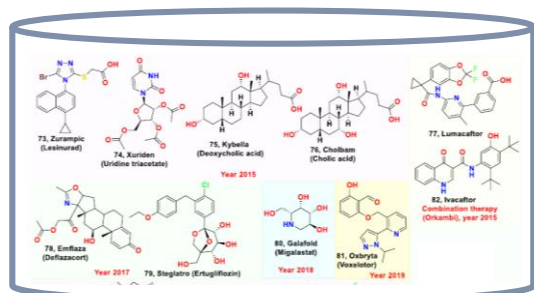
Large Scale Unlabeled Data



Molecular Property Prediction



Background: Molecular Pre-training



Large Scale Unlabeled Data

Molecular Pre-training

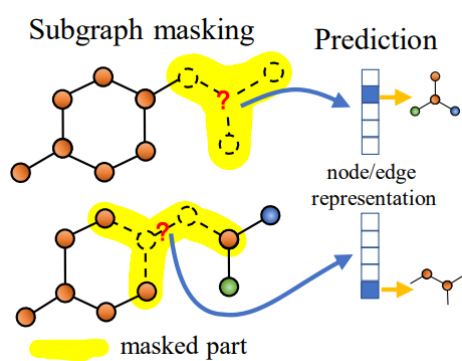


Molecular Property Prediction

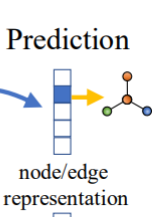
Explanatory Factors

How can we design a self-supervised task that truly adheres to physical principles?

Subgraph masking



Prediction

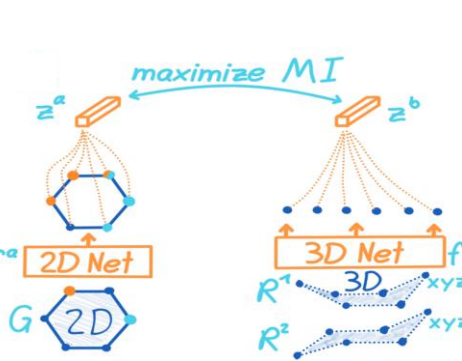


node/edge representation

masked part

Predictive Methods

maximize MI



2D Net

3D Net

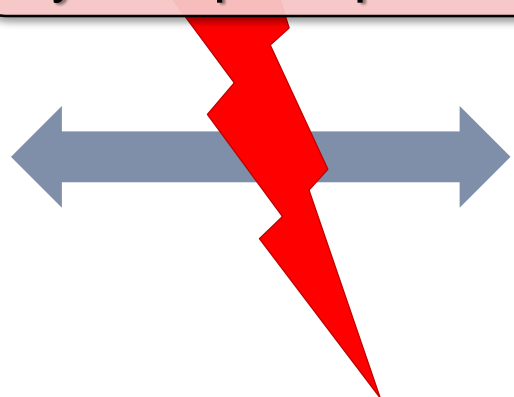
3D

2D

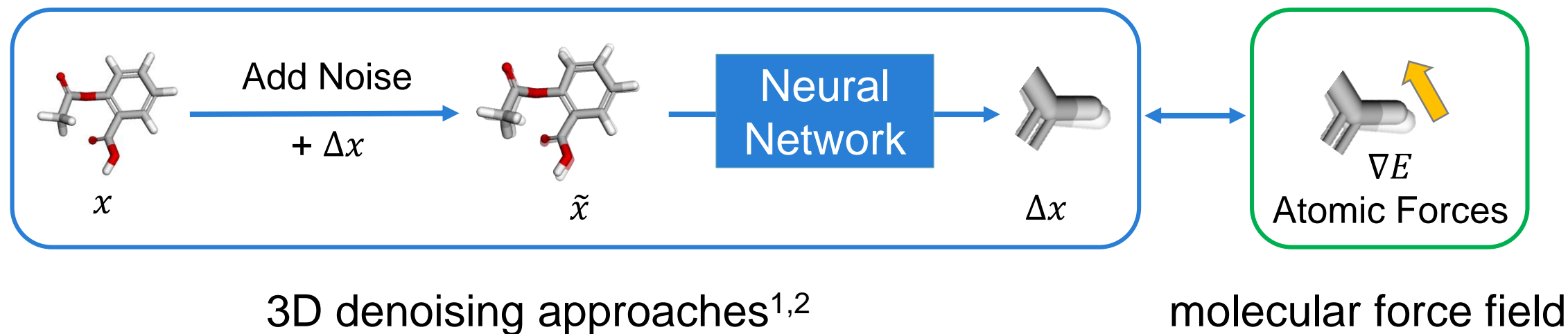
Contrastive Methods

$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond} (r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle} (\theta_i - \theta_0)^2}_{U_{angle}} + \underbrace{\sum_{dihedrals} k_i^{dihedral} [1 + \cos(n_i \phi_i + \delta_i)]}_{U_{dihedral}} + \underbrace{\sum_i \sum_{j \neq i} 4 \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]}_{U_{nonbond}} + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}$$

Physical laws



Background: Denoising Pre-training



Energy function is crucial in denoising:

$$\mathcal{L} \simeq E_{p(\tilde{x})} \|GNN_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x})\|^2$$

Sampling Distribution

Approximate Force Field

[1] Zaidi et al., Pre-training via denoising for molecular property prediction, ICLR 2023

[2] Feng et al., Fractional denoising for 3D molecular pre-training, ICML 2023

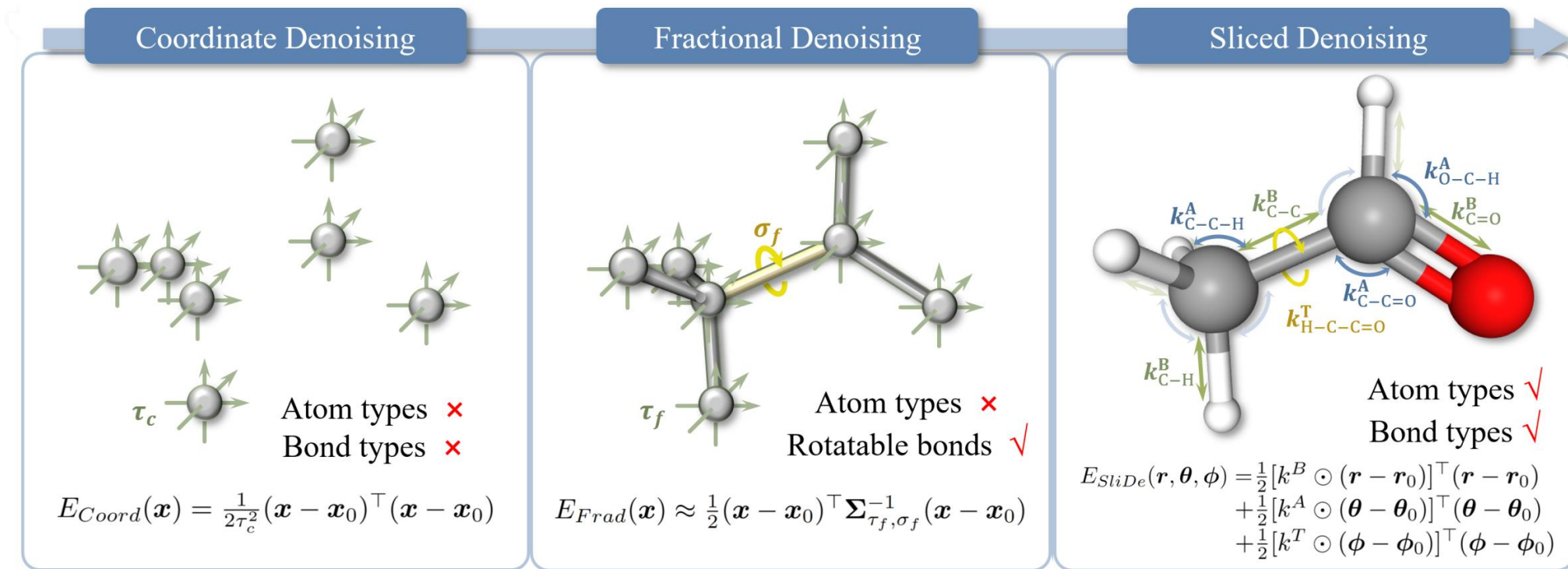
Background: Denoising Pre-training

Energy function is crucial in denoising:

$$E_{p(\tilde{x})} || GNN_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log p(\tilde{x}) ||^2$$

Sampling Distribution

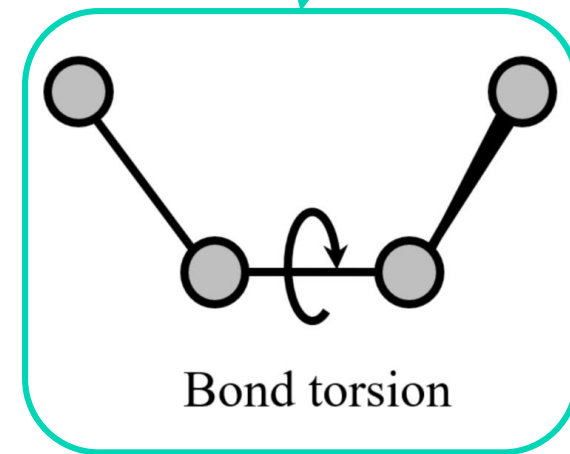
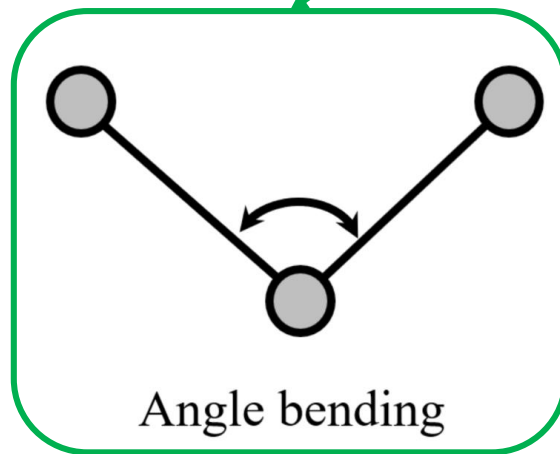
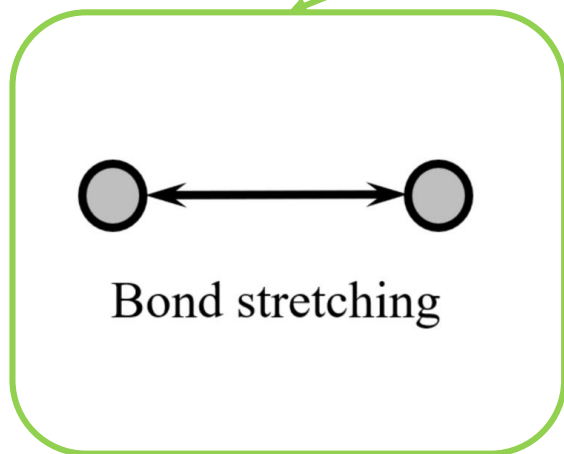
Approximate Force Field



Our Approach: Step 1. Energy Function

- Energy function:

$$E_{BAT}(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{2} \sum_{i \in \mathbb{B}} k_i^B (r_i - r_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{A}} k_i^A (\theta_i - \theta_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{T}} k_i^T \omega_i^2 (\phi_i - \phi_{i,0})^2.$$



Our Approach: Step 2. Noise Design

- Energy function:

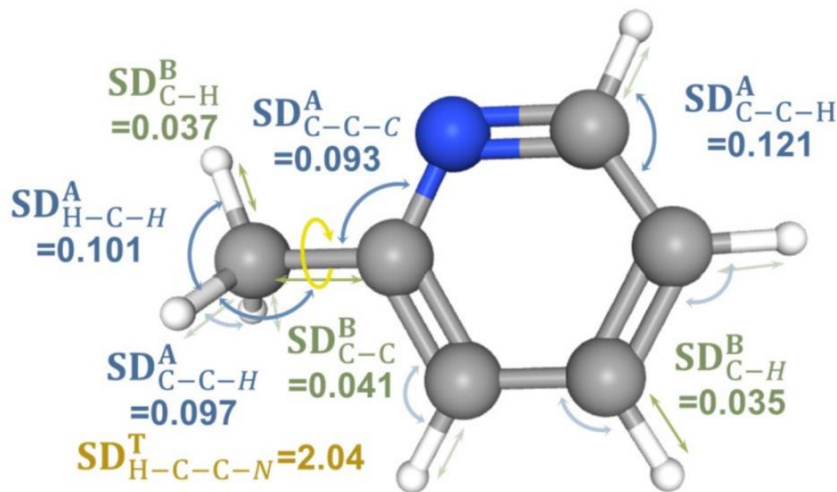
$$E_{BAT}(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{2} \sum_{i \in \mathbb{B}} k_i^B (r_i - r_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{A}} k_i^A (\theta_i - \theta_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{T}} k_i^T \omega_i^2 (\phi_i - \phi_{i,0})^2.$$

- Noise design:

$$p \propto \exp(-E)$$

Boltzmann Distribution

$$\mathbf{r} \sim \mathcal{N}(\mathbf{r}_0, \text{diag}(\frac{1}{\mathbf{k}^B})), \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \text{diag}(\frac{1}{\mathbf{k}^A})), \boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\phi}_0, \text{diag}(\frac{1}{\mathbf{k}^T \odot \boldsymbol{\omega}^2}))$$



Our Approach: Step 3. Force Field Learning

- Energy function:

$$E_{BAT}(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{2} \sum_{i \in \mathbb{B}} k_i^B (r_i - r_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{A}} k_i^A (\theta_i - \theta_{i,0})^2 + \frac{1}{2} \sum_{i \in \mathbb{T}} k_i^T \omega_i^2 (\phi_i - \phi_{i,0})^2.$$

- Noise design:

$$\mathbf{r} \sim \mathcal{N}(\mathbf{r}_0, \text{diag}(\frac{1}{\mathbf{k}^B})), \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \text{diag}(\frac{1}{\mathbf{k}^A})), \boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\phi}_0, \text{diag}(\frac{1}{\mathbf{k}^T \odot \boldsymbol{\omega}^2}))$$

- Force field learning:

$$E_{p(\mathbf{x}|\mathbf{x}_0)} \|\text{GNN}_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} E_{BAT}(\mathbf{d}(\mathbf{x}))\|^2$$

Molecular Force Field

Our Approach: Step 3. Force Field Learning

Goal

$$E_{p(\mathbf{x}|\mathbf{x}_0)} \left\| GNN_{\theta}(\mathbf{x}) - \nabla_{\mathbf{x}} E_{BAT}(\mathbf{d}(\mathbf{x})) \right\|^2$$

Cartesian coordinates

Relative coordinates $\mathbf{d} = (r, \theta, \phi)$

Variable Change

Efficient
Jacobian
Estimation

$$E_{p(\mathbf{x}|\mathbf{x}_0)} \left\| GNN_{\theta}(\mathbf{x}) - \nabla_{\mathbf{d}} E_{BAT}(\mathbf{d})^{\top} \cdot \mathbf{J}(\mathbf{x}) \right\|^2$$

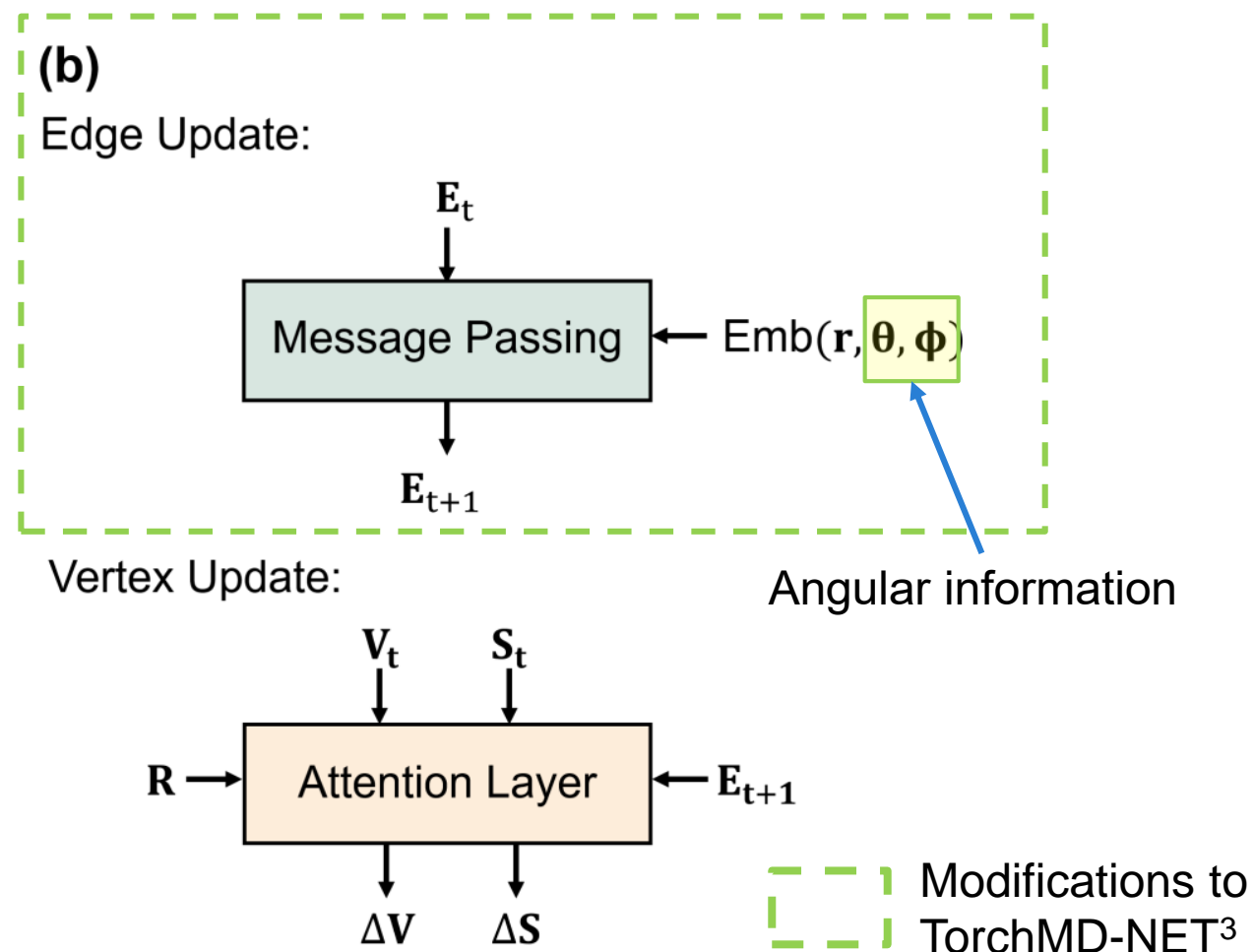
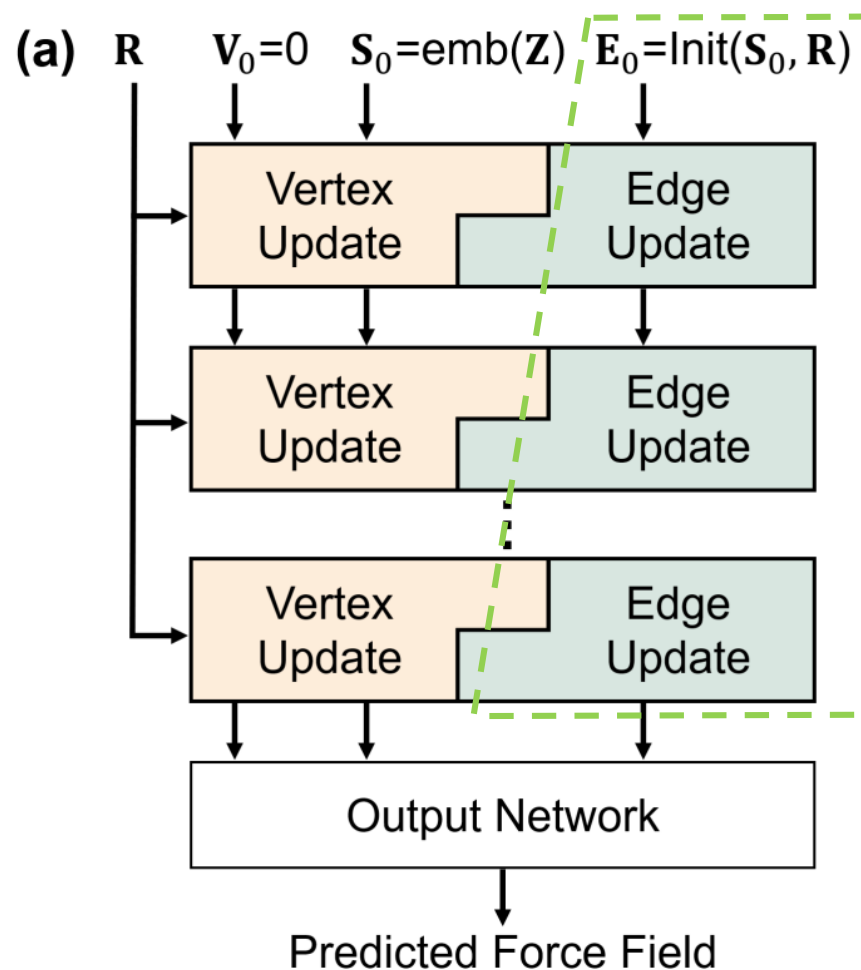
Random Slicing $\mathbf{v}_i \sim \mathcal{N}(\mathbf{0}, I_{3N})$

$$E_{p(\mathbf{x}|\mathbf{x}_0)} \frac{1}{N_v} \sum_{i=1}^{N_v} \left[GNN_{\theta}(\mathbf{x})^{\top} \cdot \mathbf{v}_i - \frac{1}{\sigma} \nabla_{\mathbf{d}} E_{BAT}(\mathbf{d})^{\top} \cdot \mathbf{J}(\mathbf{x}) \cdot \mathbf{v}_i \right]^2$$

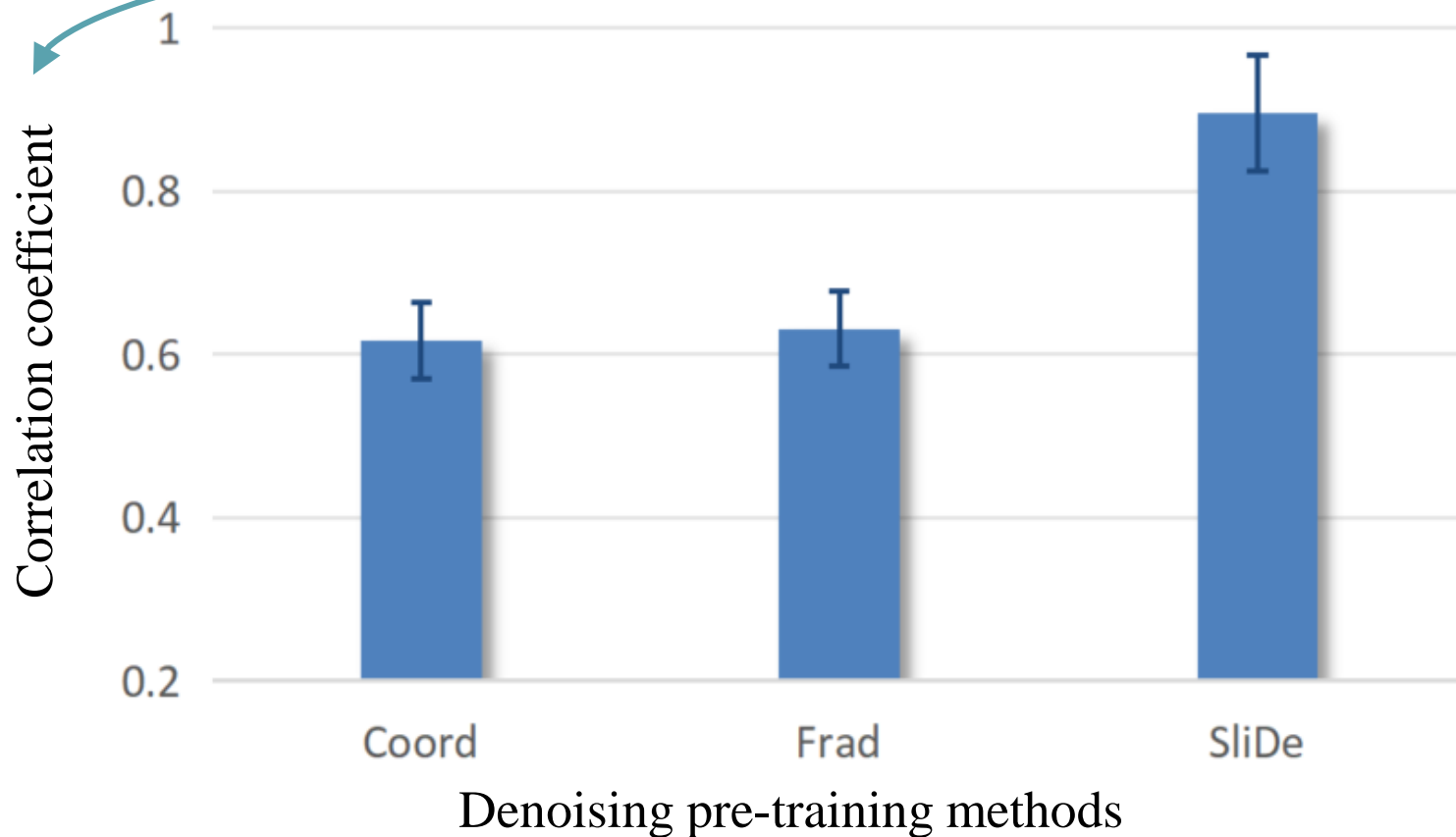
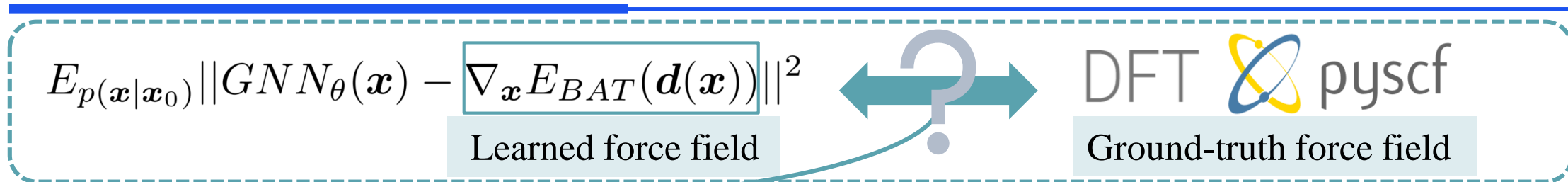
$$E_{p(\mathbf{x}|\mathbf{x}_0)} \frac{1}{N_v} \sum_{i=1}^{N_v} \left[GNN_{\theta}(\mathbf{x})^{\top} \cdot \mathbf{v}_i - \frac{1}{\sigma} \nabla_{\mathbf{d}} E_{BAT}(\mathbf{d})^{\top} \cdot \left(f^{\mathcal{M}}(\mathbf{x} + \sigma \mathbf{v}_i) - f^{\mathcal{M}}(\mathbf{x}) \right) \right]^2$$

Sliced Denoising (SliDe) Loss

Geometric Equivariant Transformer (GET)



Is SliDe More “Physical Consistent”?



Molecular property prediction – QM9

Table 1: Performance (MAE ↓) on 12 quantum chemistry property prediction in QM9.

	μ (D)	α (a_0^3)	homo (meV)	lumo (meV)	gap (meV)	R^2 (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{mol \cdot K}$)
SchNet	0.033	0.235	41.0	34.0	63.0	0.07	1.70	14.00	19.00	14.00	14.00	0.033
E(n)-GNN	0.029	0.071	29.0	25.0	48.0	0.11	1.55	11.00	12.00	12.00	12.00	0.031
DimeNet++	0.030	0.044	24.6	19.5	32.6	0.33	1.21	6.32	6.28	6.53	7.56	0.023
PaiNN	0.012	0.045	27.6	20.4	45.7	0.07	1.28	5.85	5.83	5.98	7.35	0.024
SphereNet	0.025	0.045	22.8	18.9	31.1	0.27	1.120	6.26	6.36	6.33	7.78	0.022
ET	0.011	0.059	20.3	17.5	36.1	0.033	1.840	6.15	6.38	6.16	7.62	0.026
TM	0.037	0.041	17.5	16.2	27.4	0.075	1.18	9.37	9.41	9.39	9.63	0.022
SE(3)-DDM	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026
Coord	0.012	0.0517	17.7	14.3	31.8	0.4496	1.71	6.57	6.11	6.45	6.91	0.020
Frad	0.010	0.0374	15.3	13.7	27.8	0.3419	1.418	5.33	5.62	5.55	6.19	0.020
SliDe	0.0087	0.0366	13.6	12.3	26.2	0.3405	1.521	4.28	4.29	4.26	5.37	0.019

Molecular property prediction – MD17, ANI-1x

Table 2: Performance (MAE ↓) on MD17 force prediction (kcal/mol/Å).

	Aspirin	Benzene	Ethanol	Malonal -dehyde	Naphtha -lene	Salicy -lic Acid	Toluene	Uracil
SphereNet	0.430	0.178	0.208	0.340	0.178	0.360	0.155	0.267
SchNet	1.35	0.31	0.39	0.66	0.58	0.85	0.57	0.56
DimeNet	0.499	0.187	0.230	0.383	0.215	0.374	0.216	0.301
PaiNN*	0.338	-	0.224	0.319	0.077	0.195	0.094	0.139
ET	0.2450	0.2187	0.1067	0.1667	0.0593	0.1284	0.0644	0.0887
SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
Coord	0.2108	0.1692	0.0959	0.1392	0.0529	0.1087	0.0582	0.0742
Frad	0.2087	0.1994	0.0910	0.1415	0.0530	0.1081	0.0540	0.0760
SliDe	0.1740	0.1691	0.0882	0.1538	0.0483	0.1006	0.0540	0.0825

	Noneq	SliDe
w/o pre-train	1.50	1.362
pre-train	1.01	0.786
pre-train improvement	32.7%	42.3%

Table 3: Performance (MAE ↓) on ANI-1x energy prediction (kcal/mol).

Take Home Message

Physical consistency is important for molecular representation (especially for quantum downstream tasks), **setting a new paradigm of explainable SSL:**

- Self-supervised tasks can be designed directly from physical quantities.
- The quality of physically Interpretable representation can be quantified by physical consistency e.g. force field in Slide

More interesting results in paper:

- Physical consistency as a new hyperparameter tuning approach.
- Good data scaling and robustness properties.
- Regularization term contributes to downstream transfer.

Waiting for you @ Poster Session 1, Tue 7 May 10:45 -12:45 !