中国科学院自动化研究所 模式识别实验室 New Laboratory of Pattern Recognition

中国科学院自动化研究所 Institute of Automation Chinese Academy of Sciences

達摩院 DAMO ACADEMY

[ICLR 2025]

MolSpectra: Pre-training 3D Molecular Representation with Multi-modal Energy Spectra

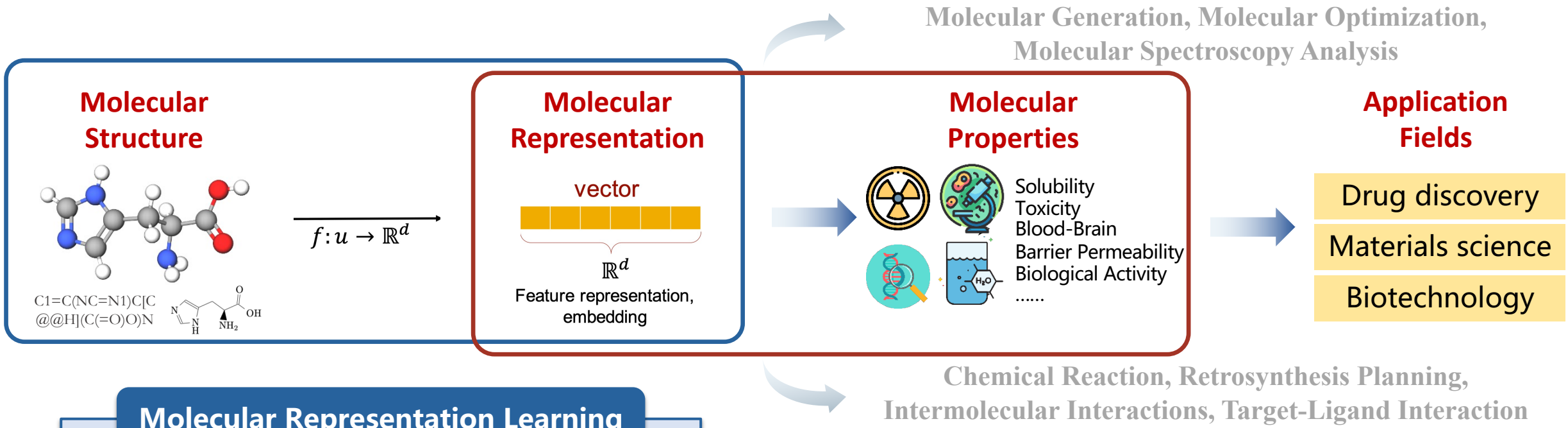
Liang Wang[1,2], Shaozhen Liu[1], Yu Rong[3], Deli Zhao[3], Qiang Liu[1,2], Shu Wu[1,2], Liang Wang[1,2]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]DAMO Academy, Alibaba Group

ICLR

12 March 2025

Research Background



Molecular Representation Learning

Translate the molecular structures into vectorized molecular representations to understand and predict various molecular properties, interactions, chemical reactions.

$$h = f(molecule)$$

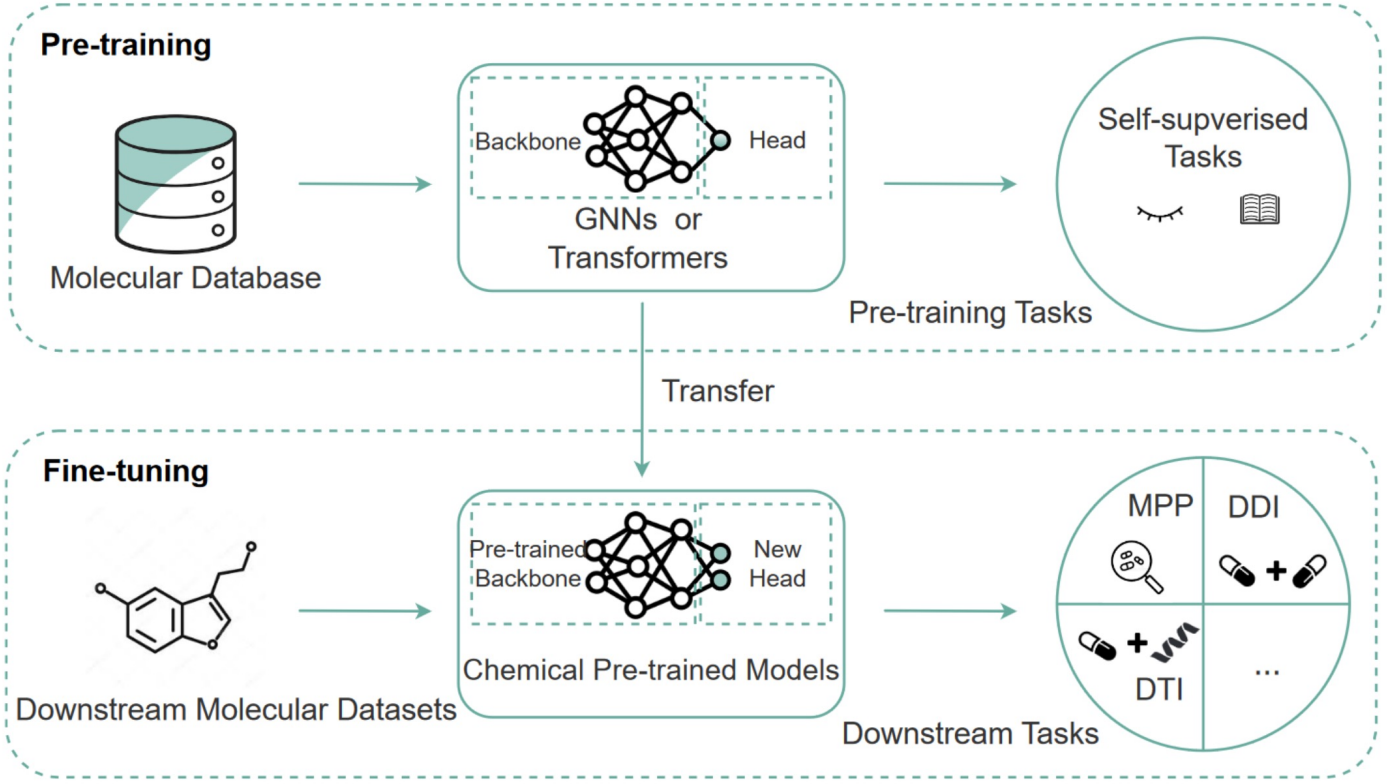
Research Background

Challenges of supervised molecular representation learning

(1) Scarcity of labeled data.

(2) Poor out-of-distribution generalization capability.

Pipeline of Molecular Representation Pre-training



- ✓ Pre-trained on large-scale unlabeled molecules.
- ✓ Fine-tuned on various downstream tasks.

Research Background

- **Denoising as learning a force field.**
  - It is not feasible to learn molecular force field directly, since it is either unknown or expensive to evaluate.
  - Alternative: approximate the data-generating force field with one that can be cheaply evaluated.
  - Prove that the denoising objective is equivalent to learning the molecular force field:
    - Molecular structure: $\mathbf{x} \in \mathbb{R}^{3N}$
    - The structure follows the Boltzmann distribution: $p_{\text{physical}}(\mathbf{x}) \propto \exp(-E(\mathbf{x}))$
    - Force field: $\nabla_{\mathbf{x}} \log p_{\text{physical}}(\mathbf{x}) = -\nabla_{\mathbf{x}} E(\mathbf{x})$
    - Approximate p_{physical} with a mixture of Gaussians centered at the known equilibrium structures

$$p_{\text{physical}}(\tilde{\mathbf{x}}) \approx q_\sigma(\tilde{\mathbf{x}}) := \frac{1}{n}\sum_{i=1}^{n} q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x_i})$$

where $q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x_i}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x_i}, \sigma^2 I_{3N})$.

[1] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, Jonathan Godwin. **"Pre-Training via Denoising for Molecular Property Prediction."** In *ICLR*, 2023

Research Background

- **Denoising as learning a force field. (Cont.)**
  - Learning the force field now yields a score-matching objective:

$$\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}[\| \mathrm{GNN}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}}) \|^2]$$

  - According to reference [1], minimizing the following two objectives is equivalent:

$$J_1(\theta) = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}[\| \mathrm{GNN}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}}) \|^2]$$

$$J_2(\theta) = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}[\| \mathrm{GNN}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) \|^2]$$

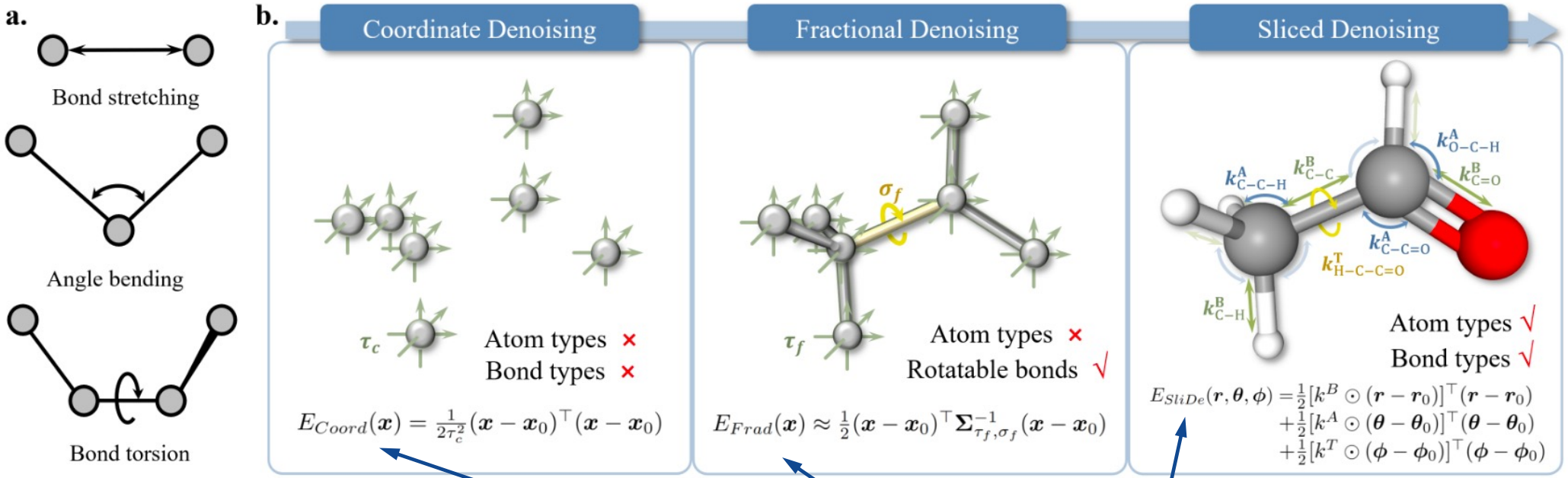
  - Thus, the objective in Eq. (1) is equivalent to:

$$\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}[\| \mathrm{GNN}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) \|^2] = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\| \mathrm{GNN}_\theta(\tilde{\mathbf{x}}) - \frac{\mathbf{x}-\tilde{\mathbf{x}}}{\sigma^2} \|^2\right]$$

Establishing the relationship between 3D geometries and the energy states of molecular systems is an effective pathway to learn 3D molecular representations.

[1] Pascal Vincent. **"A Connection Between Score Matching and Denoising Autoencoders."** *Neural Computation* July 2011

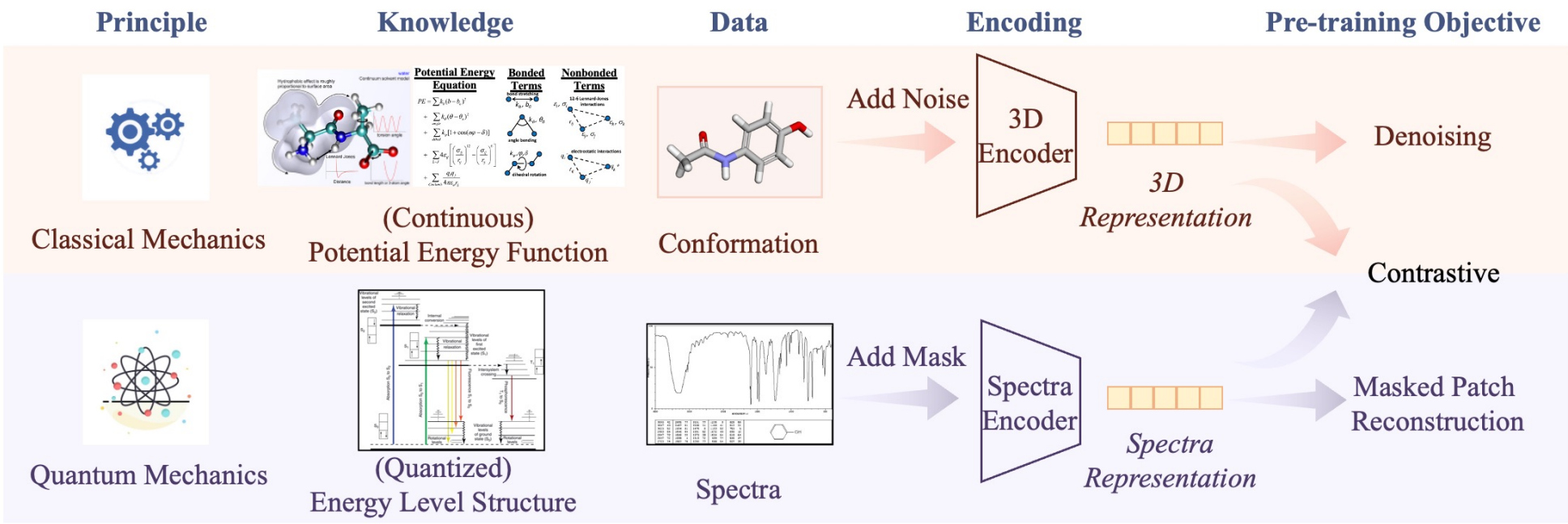
Research Background

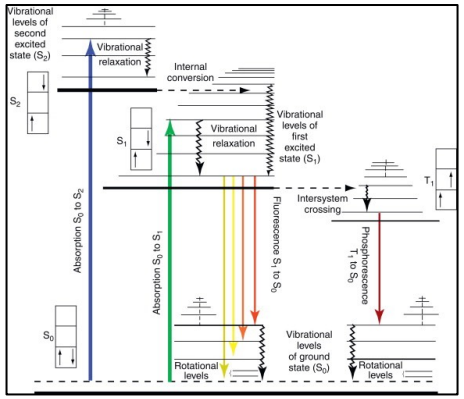


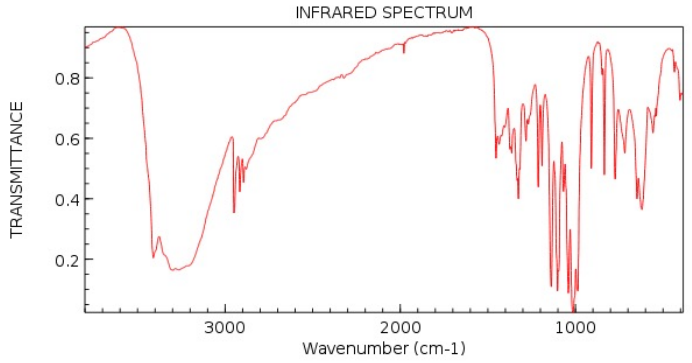
$$\mathcal{L}_{\text{Denoising}}(\mathcal{M}) = \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{x}_0)p(\boldsymbol{x}_0)}\|\text{GNN}_\theta(\boldsymbol{x}) - (\boldsymbol{x}-\boldsymbol{x}_0)\|^2$$
$$\simeq \mathbb{E}_{p(\boldsymbol{x})}\|\text{GNN}_\theta(\boldsymbol{x}) - (-\nabla_{\boldsymbol{x}} E(\boldsymbol{x}))\|^2,$$

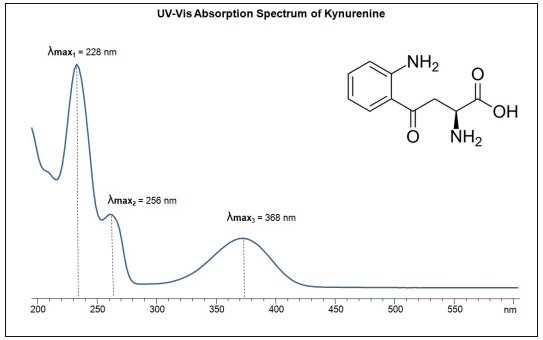
[1] Yuyan Ni, Shikun Feng, Wei-Ying Ma, Zhi-Ming Ma, Yanyan Lan. **"Sliced Denoising: A Physics-Informed Molecular Pre-Training Method."** In *ICLR*, 2024

Motivation

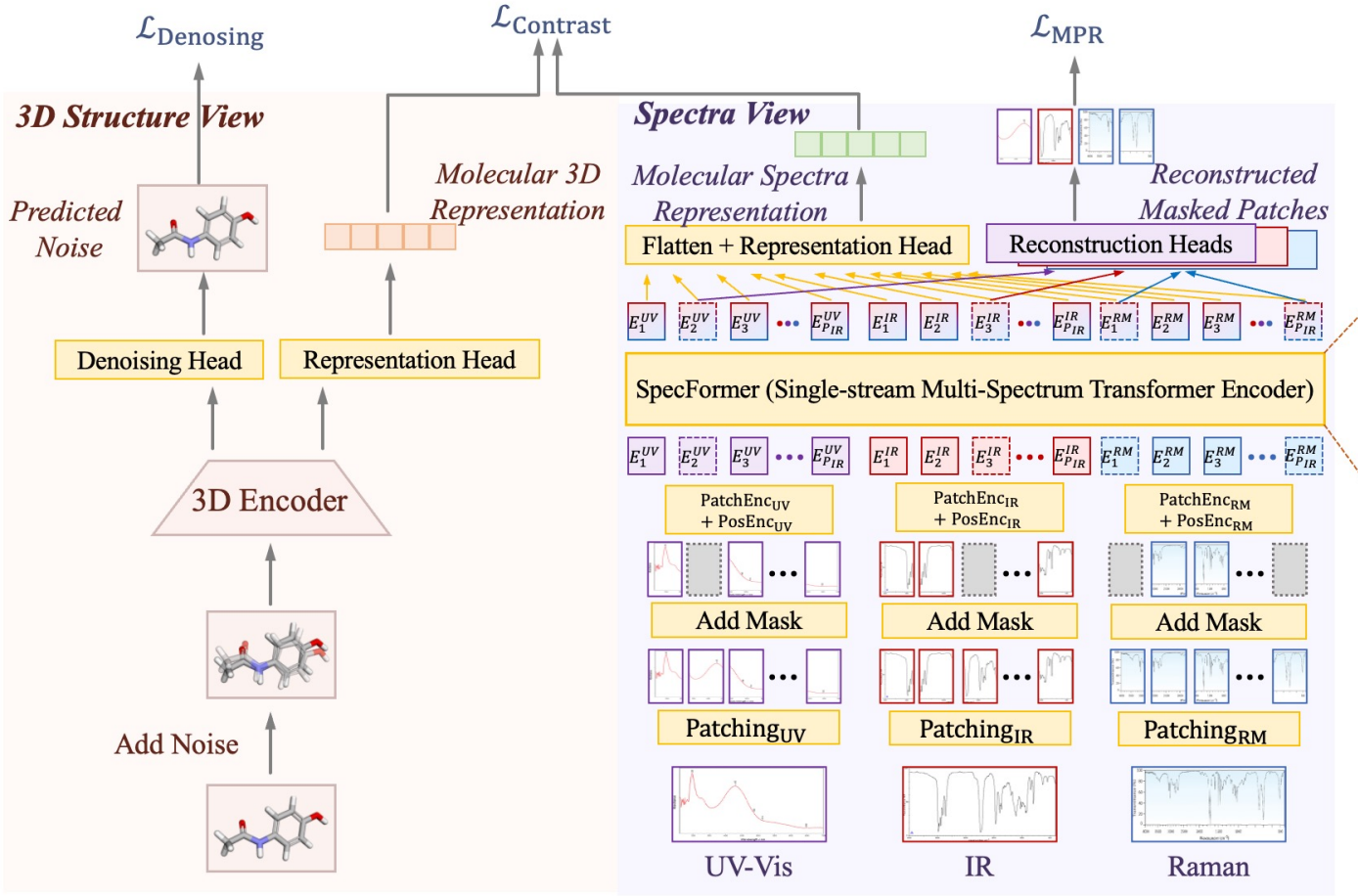

Principle
Knowledge
Data
Encoding
Pre-training Objective
Potential Energy Equation
Bonded Terms
Nonbonded Terms
Add Noise
3D Encoder
3D Representation
Denoising
Classical Mechanics
(Continuous) Potential Energy Function
Conformation
Contrastive
Quantum Mechanics
(Quantized) Energy Level Structure
Spectra
Add Mask
Spectra Encoder
Spectra Representation
Masked Patch Reconstruction

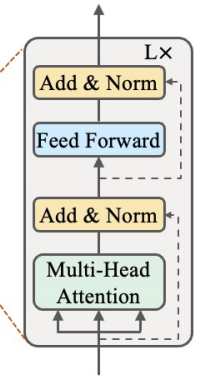

Vibrational levels of second excited state (S_2)
Vibrational relaxation
Internal conversion
Vibrational levels of first excited state (S_1)
Intersystem crossing
Absorption S_0 to S_2
Absorption S_0 to S_1
Fluorescence S_1 to S_0
Phosphorescence T_1 to S_0
Vibrational levels of ground state (S_0)
Rotational levels

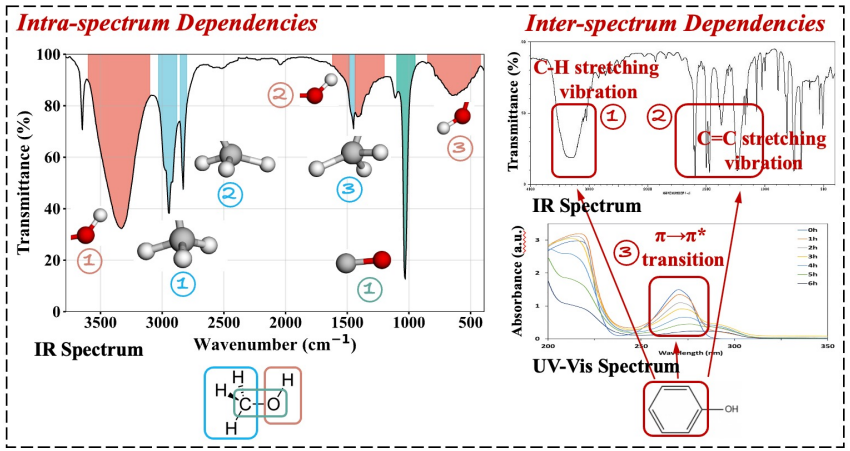

INFRARED SPECTRUM
TRANSMITTANCE
0.8
0.6
0.4
0.2
3000
2000
1000
Wavenumber (cm-1)


UV-Vis Absorption Spectrum of Kynurenine
λmax1 = 228 nm
λmax2 = 256 nm
λmax3 = 368 nm
NH2
O
OH
O
NH2
200
250
300
350
400
450
500
550
nm

MolSpectra







$$\mathcal{L} = \beta_{\text{Denoising}}\mathcal{L}_{\text{Denoising}} + \beta_{\text{MPR}}\mathcal{L}_{\text{MPR}} + \beta_{\text{Contrast}}\mathcal{L}_{\text{Contrast}}$$

Experiments

Effectiveness of Molecular Spectra in Training from Scratch

Table 1: Performance (MAE ↓) when training from scratch on QM9 dataset.

Task Units	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 ($\frac{cal}{mol \cdot K}$)
w/o spectra	0.029	0.071	29	25	48	0.106	1.55	11	12	12	12	0.031
w/ spectra	**0.027**	**0.049**	**28**	**24**	**43**	**0.084**	**1.45**	**10**	**11**	**10**	**10**	**0.030**

Experiments

Effectiveness of Molecular Spectra in Representation Pre-Training

Table 2: Performance (MAE↓) on QM9 dataset. The compared methods are divided into two groups training from scratch and pre-training then fine-tuning. The best results are highlighted in bold.

	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	0.070	1.70	14.00	19.00	14.00	14.00	0.033
EGNN	0.029	0.071	29.0	25.0	48.0	0.106	1.55	11.00	12.00	12.00	12.00	0.031
DimeNet++	0.030	0.044	24.6	19.5	32.6	0.330	1.21	6.32	6.28	6.53	7.56	0.023
PaiNN	0.012	0.045	27.6	20.4	45.7	0.070	1.28	5.85	5.83	5.98	7.35	0.024
SphereNet	0.025	0.045	22.8	18.9	31.1	0.270	**1.12**	6.26	6.36	6.33	7.78	0.022
TorchMD-Net	0.011	0.059	20.3	17.5	36.1	**0.033**	1.84	6.15	6.38	6.16	7.62	0.026
Transformer-M	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	0.016	0.052	17.7	14.7	31.8	0.450	1.71	6.57	6.11	6.45	6.91	**0.020**
MolSpectra	**0.011**	0.048	**15.5**	**13.1**	**26.8**	0.410	1.71	**5.67**	**5.45**	**5.87**	**6.18**	0.021

Experiments

Effectiveness of Molecular Spectra in Representation Pre-Training

Table 3: Performance (MAE↓) on MD17 force prediction (kcal/mol/ Å). The methods are divided into two groups: training from scratch and pre-training then fine-tuning. The best results are in bold.

	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	1.350	0.310	0.390	0.660	0.580	0.850	0.570	0.560
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
TorchMD-Net	0.245	0.219	0.107	0.167	0.059	0.128	0.064	0.089
SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
Coord	0.211	0.169	0.096	0.139	**0.053**	0.109	**0.058**	**0.074**
MolSpectra	**0.099**	**0.097**	**0.052**	**0.077**	0.085	**0.093**	0.075	0.095

Experiments

Sensitivity Analysis of Patch Length, Stride, and Mask Ratio

Table 4: Sensitivity of patch length and stride.

patch length	stride	overlap ratio	homo	lumo	gap
20	5	75%	15.9	13.7	28.0
20	10	50%	**15.5**	**13.1**	**26.8**
20	15	25%	16.1	13.6	28.1
20	20	0%	15.7	13.5	27.5
16	8	50%	16.0	13.4	27.6
30	15	50%	15.9	14.0	28.1

Table 5: Sensitivity of mask ratio.

mask ratio	homo	lumo	gap
0.05	15.7	13.4	29.7
0.10	**15.5**	**13.1**	**26.8**
0.15	15.7	13.5	28.0
0.20	16.0	13.6	28.1
0.25	16.3	13.5	28.0
0.30	16.2	13.7	29.0

Ablation Study of Spectral Modalities

Table 7: Ablation of spectral modalities.

UV-Vis	IR	Raman	homo	lumo	gap
✓	✓	✓	15.5	13.1	26.8
-	✓	✓	15.8	13.3	27.1
✓	-	✓	16.6	14.1	28.9
✓	✓	-	16.1	13.9	28.3

Experiments

Visualization of Attention Patterns and Learned Spectra Representations in SpecFormer

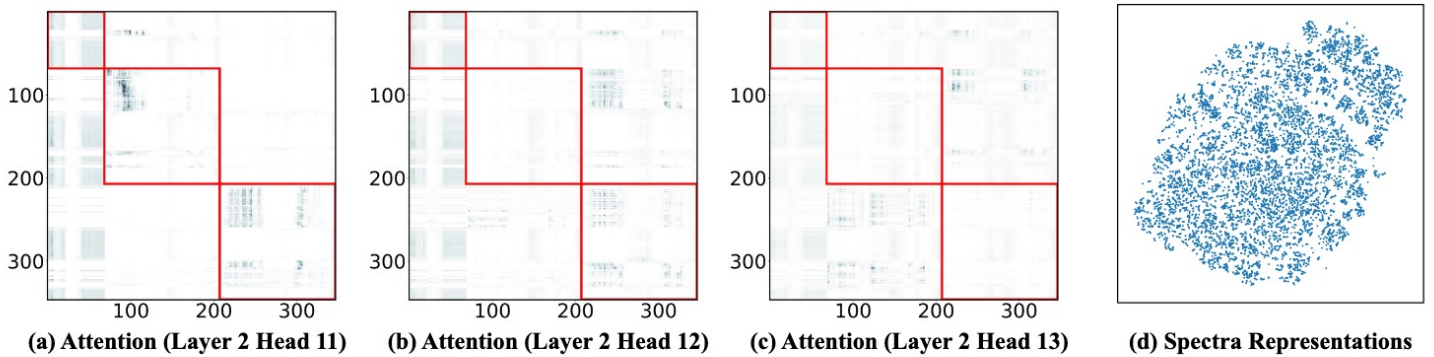


Figure A2: (a-c) Attention maps from three attention heads in SpecFormer. Different heads model distinct dependencies. (d) t-SNE visualization of the spectra representations produced by SpecFormer.

中国科学院自动化研究所
模式识别实验室
New Laboratory of Pattern Recognition

中国科学院自动化研究所
Institute of Automation
Chinese Academy of Sciences

達摩院
DAMO ACADEMY

Thank you for your attention!

Contact： liang.wang@cripac.ia.ac.cn