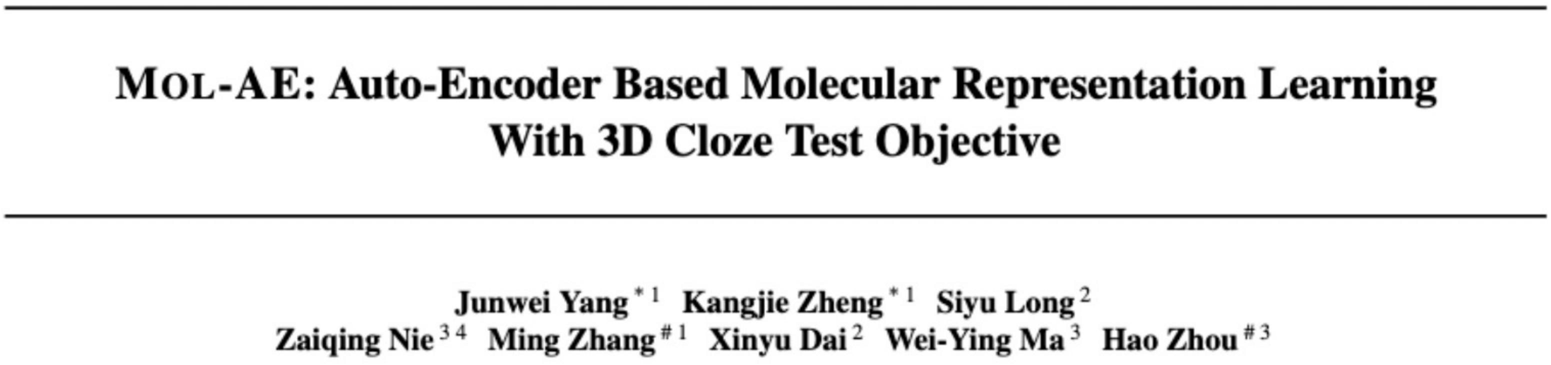
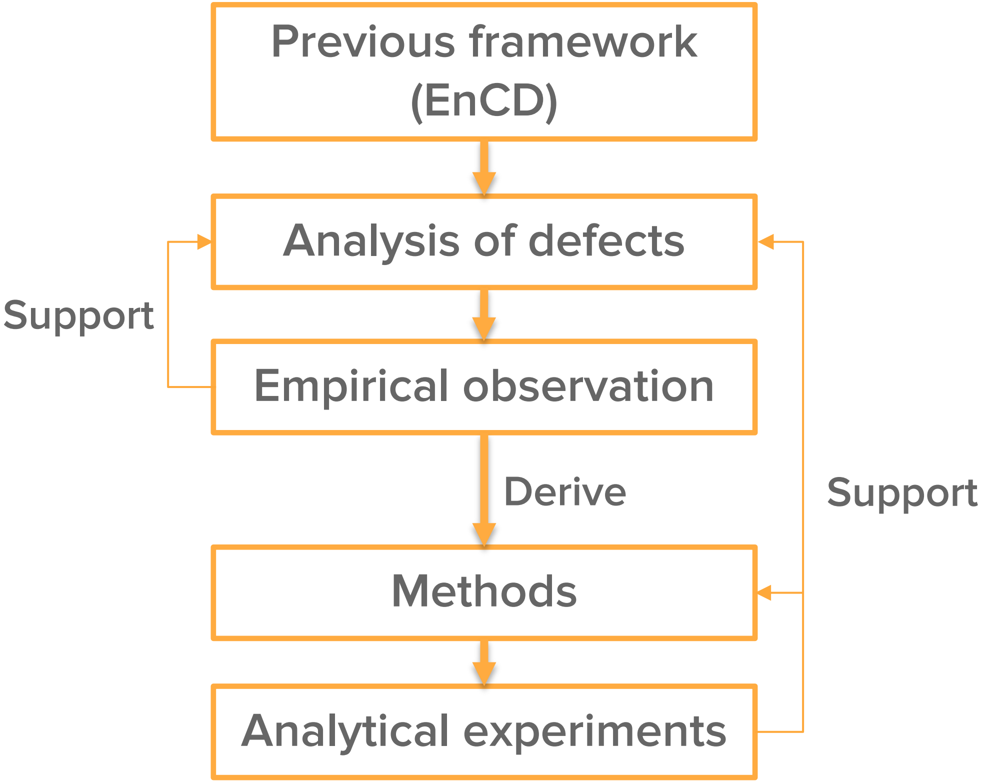


Mol-AE: Auto-Encoder Based Molecular Representation Learning With 3D Cloze Test Objective

Junwei Yang[*1] Kangjie Zheng[*1] Siyu Long[2]
Zaiqing Nie[3 4] Ming Zhang[#1] Xinyu Dai[2] Wei-Ying Ma[3] Hao Zhou[#3]

yjwtheonly@pku.edu.cn, kangjie.zheng@gmail.com

Outline



Previous framework (EnCD)

Analysis of defects

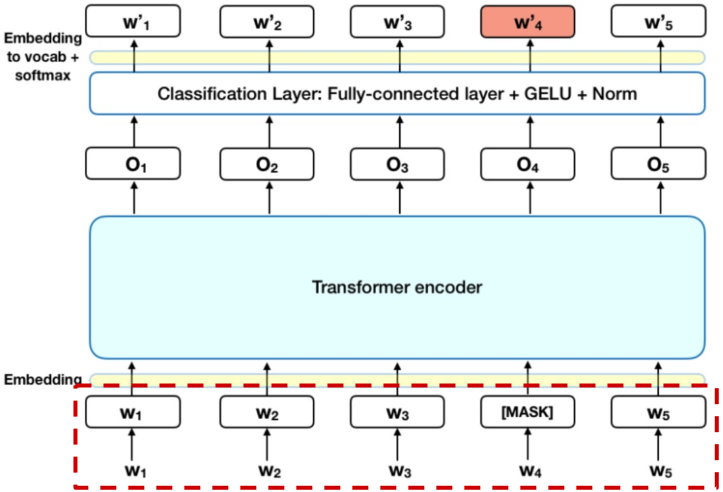
Empirical observation

Methods

Analytical experiments

Representation Learning

- Representation Learning in NLP
  - Transformer Encoder + Self-supervised Training: BERT

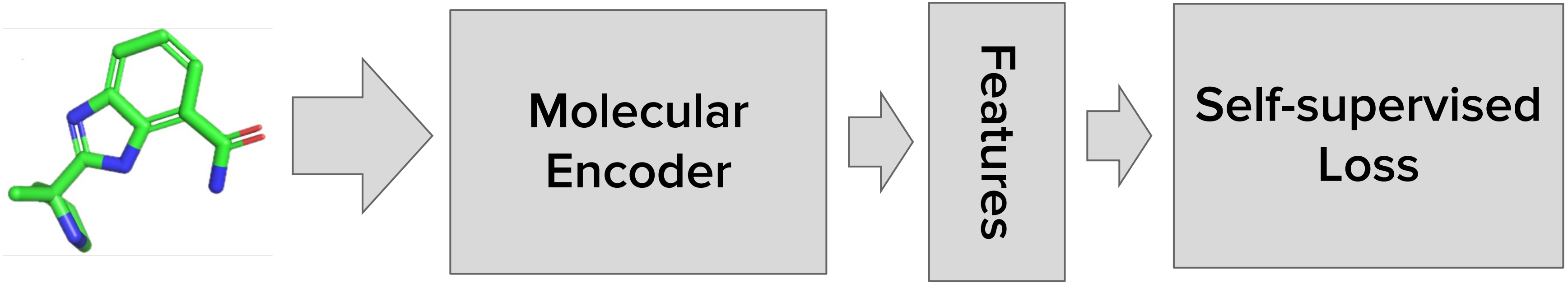

w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

随机MASK 15% Token

Pretraining Objective: $\mathcal{L}_{\mathrm{MLM}} = -\sum_{\hat{x}\in m(\mathbf{x})} \log p\left(\hat{x} \mid \mathbf{x}_{\backslash m(\mathbf{x})}\right)$

3D Molecular Representation Learning

- What's 3D Molecular Representation Learning



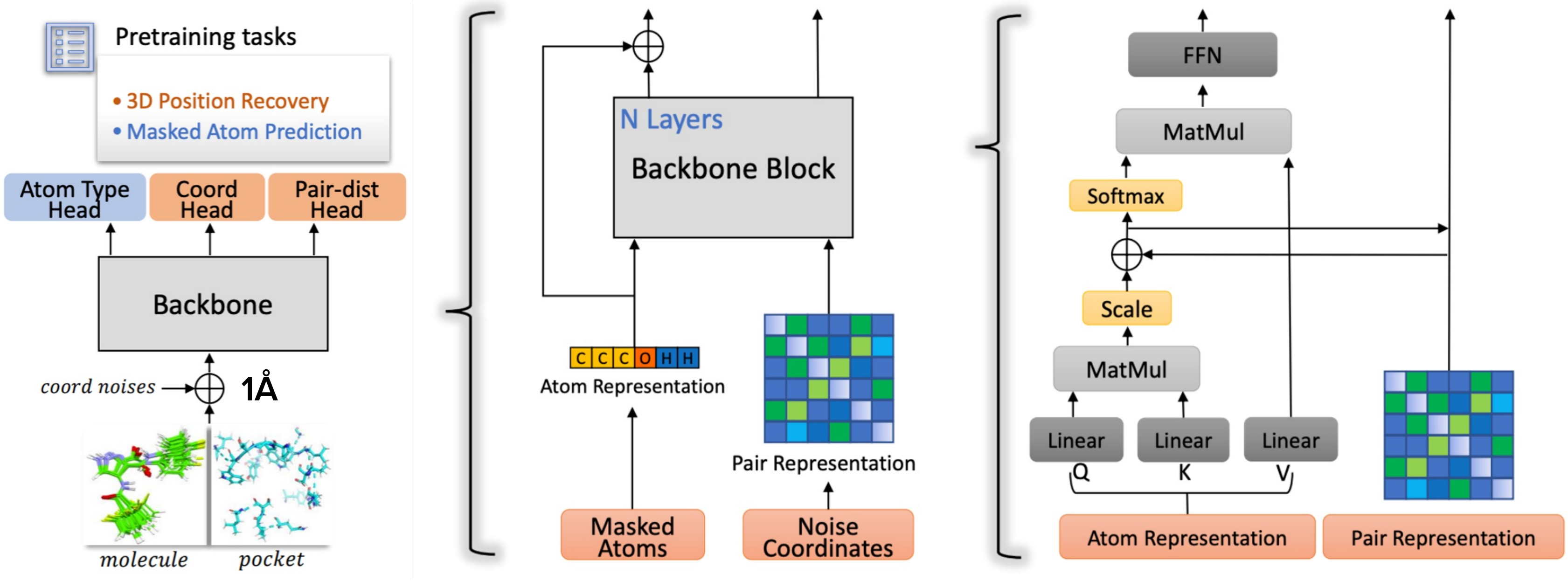
- 3D Molecular Information
  - **Atom Coordinates**
  - Distance between atoms
  - Dihedral Angel
  - ……

Transformer or GNN

- **Coordinates Denosing**
- Distance Prediction
- Dihedral Angel Prediction
- ……

3D Molecular Representation Learning

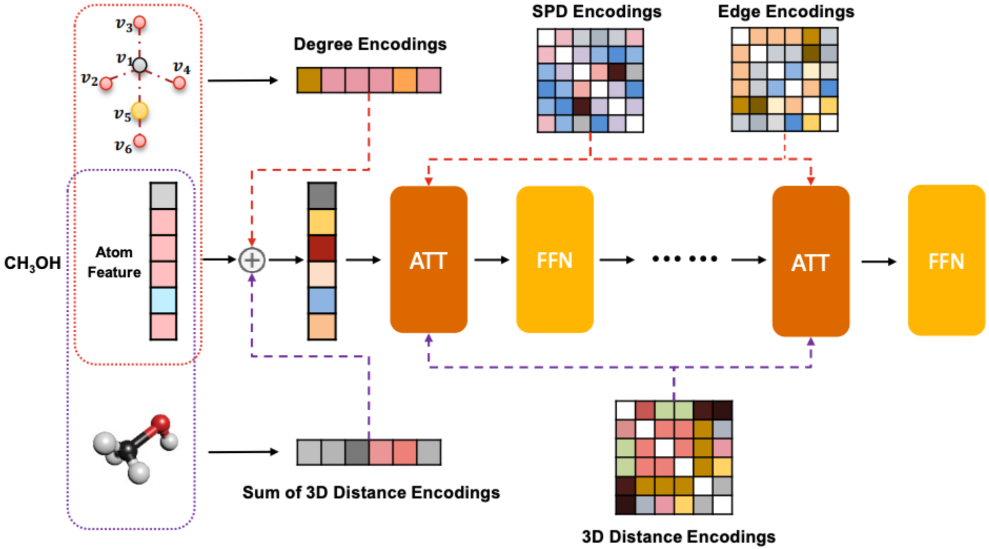
- Uni-Mol



Zhou, Gengmo, et al. "Uni-Mol: a universal 3D molecular representation learning framework." ICLR (2023).

3D Molecular Representation Learning

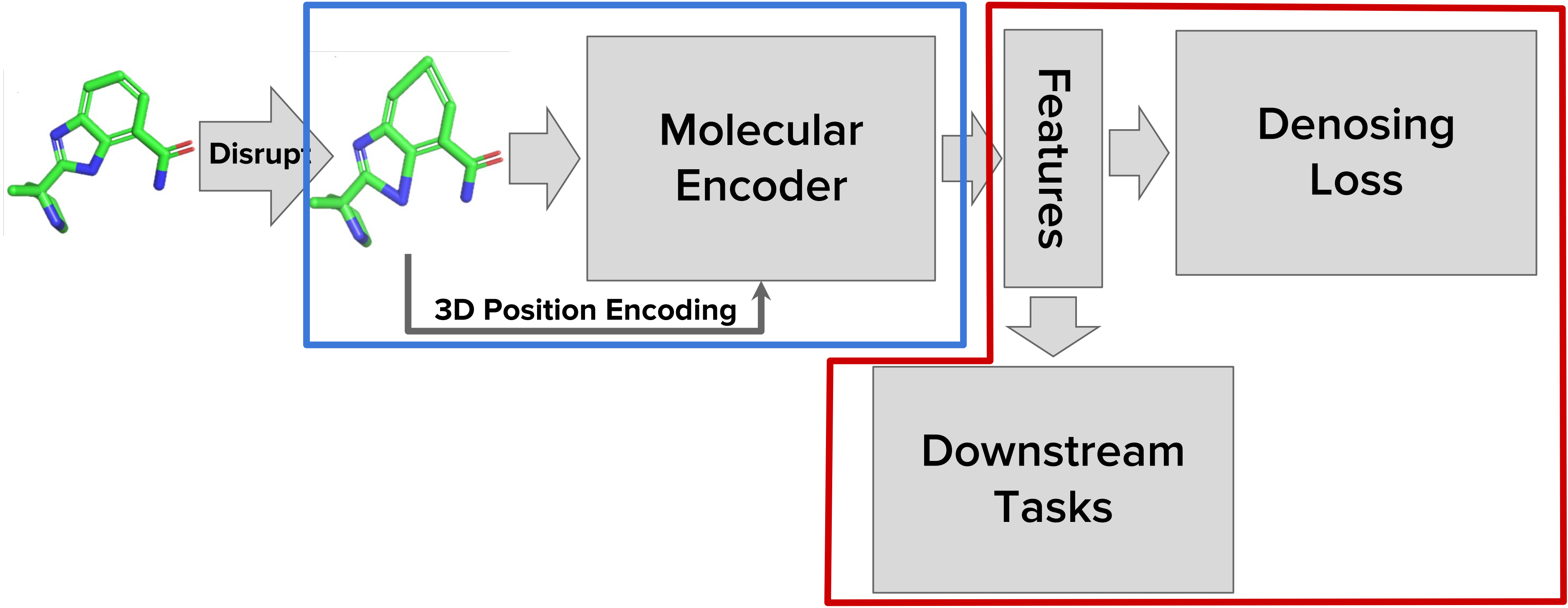
- Transformer-M



Shengjie Luo, et al. "One Transformer Can Understand Both 2D & 3D Molecular Data." ICLR (2023).

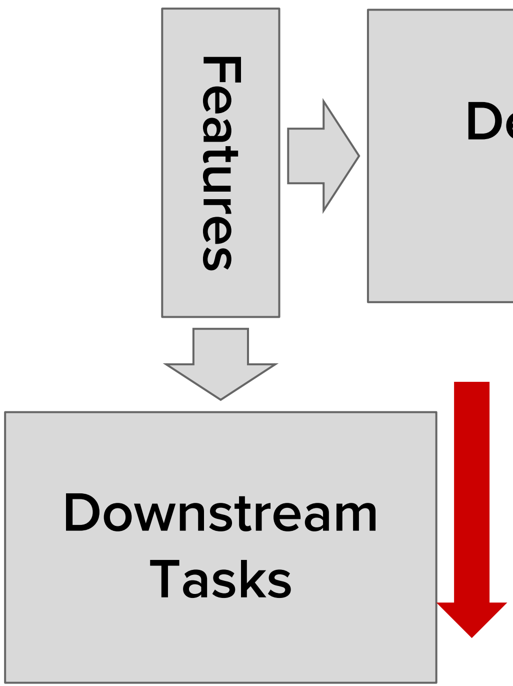
3D Molecular Representation Learning

- What's 3D Molecular Representation Learning
  - EnCD Framework: Encoder-only model with Coordinate Denoising objective



Analysis: Inconsistencies between Objectives

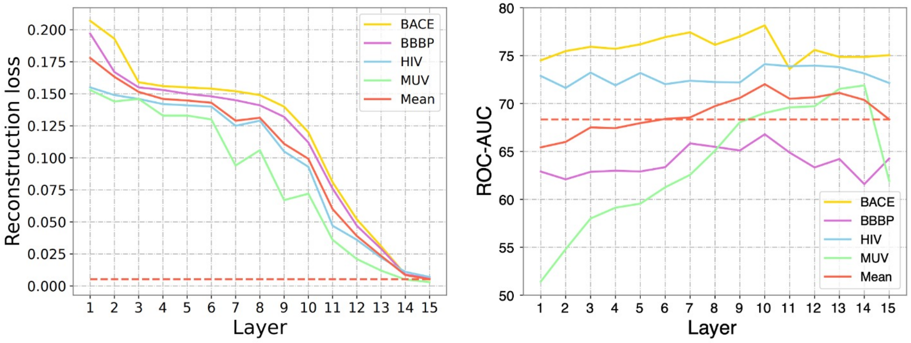
- Inconsistencies Between Pre-Training and Downstream Objectives



- **Inconsistencies:**
  - **Good Performance on Pre-training Tasks**
  - **Poor Performance on Downstream Tasks**

Analysis: Inconsistencies between Objectives

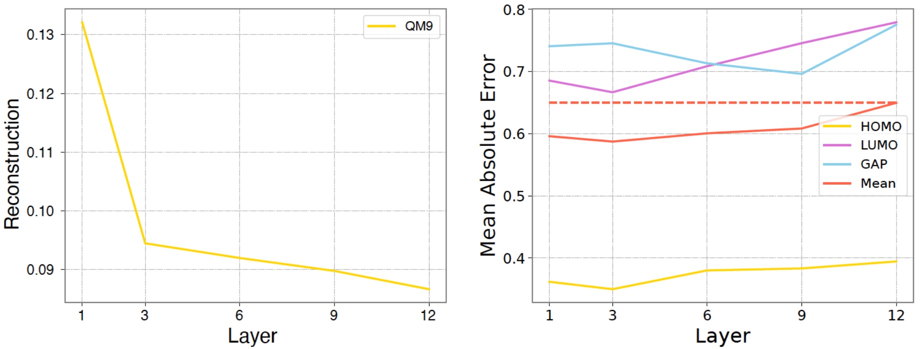
- Inconsistencies Between Pre-Training and Downstream Objectives
  - ○ Good Performance on Pre-training Tasks
  - ○ Poor Performance on Downstream Tasks



(a) Reconstruction probing.

(b) Downstream task probing.

UniMol



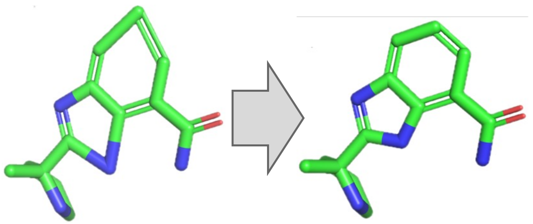
(a) Reconstruction probing.

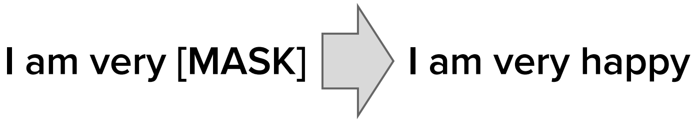
(b) Downstream task probing.

Transformer-M

Analysis: Inconsistencies between Objectives

- Inconsistencies Between Pre-Training and Downstream Objectives
  - Good Performance on Pre-training Tasks
  - Poor Performance on Downstream Tasks
- Differences between Molecules and Languages





I am very [MASK] → I am very happy

Molecules

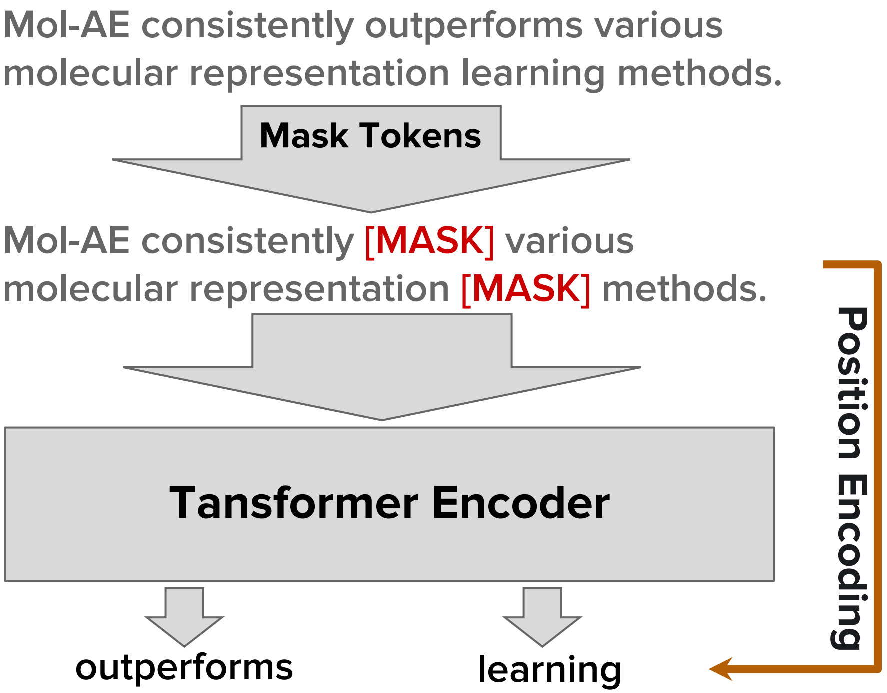
- Low-level Semantic
- Unordered information

Language

- High-level Semantic
- Ordered information

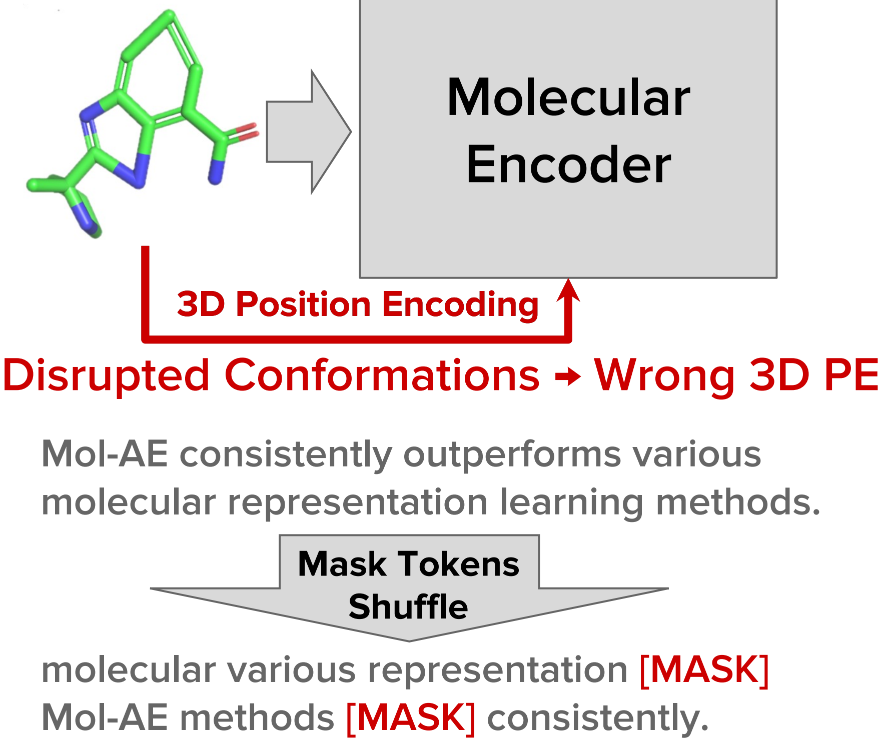
Analysis: Twisted Optimization of *Content* and *Identifier*

- **Differences** between Molecules and Languages



Mol-AE consistently outperforms various molecular representation learning methods.

Mol-AE consistently [MASK] various molecular representation [MASK] methods.



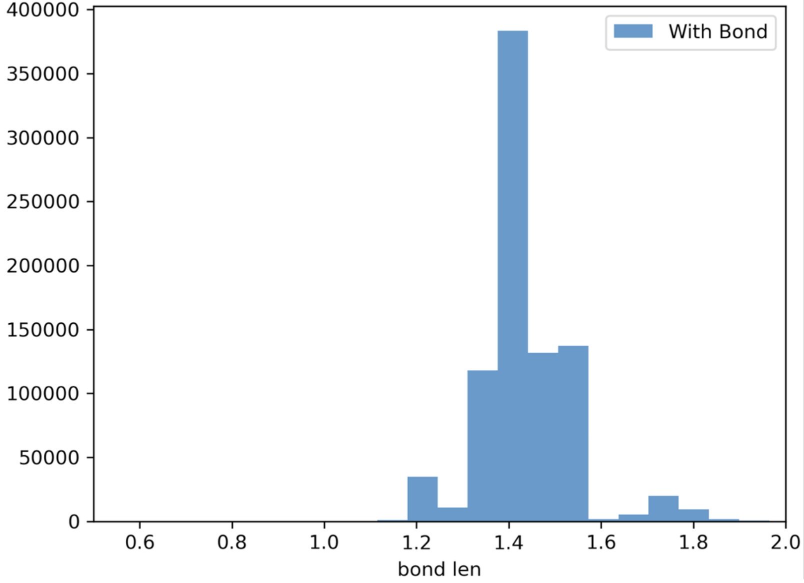
Disrupted Conformations → Wrong 3D PE

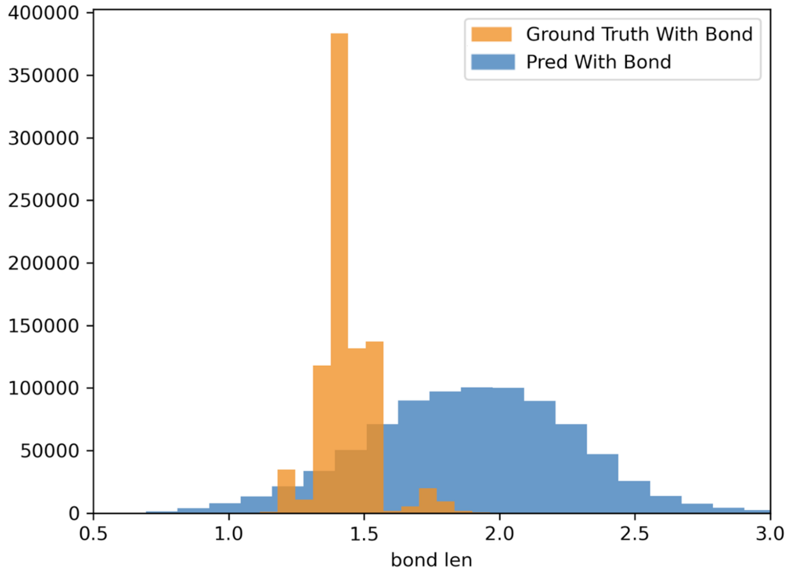
Mol-AE consistently outperforms various molecular representation learning methods.

molecular various representation [MASK] Mol-AE methods [MASK] consistently.

Analysis: Twisted Optimization of *Content* and *Identifier*

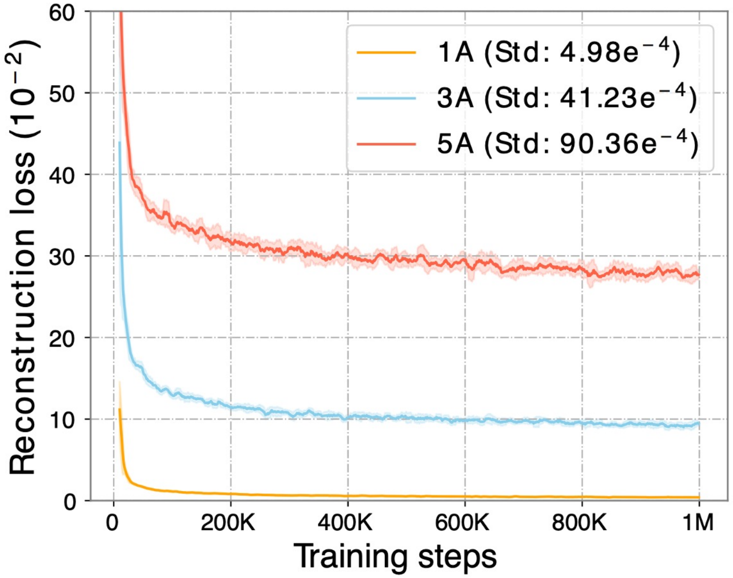
- How much impact does noise have on PE
  - Chemical bond length distribution: 1Å Uniform Noise is a strong noise.





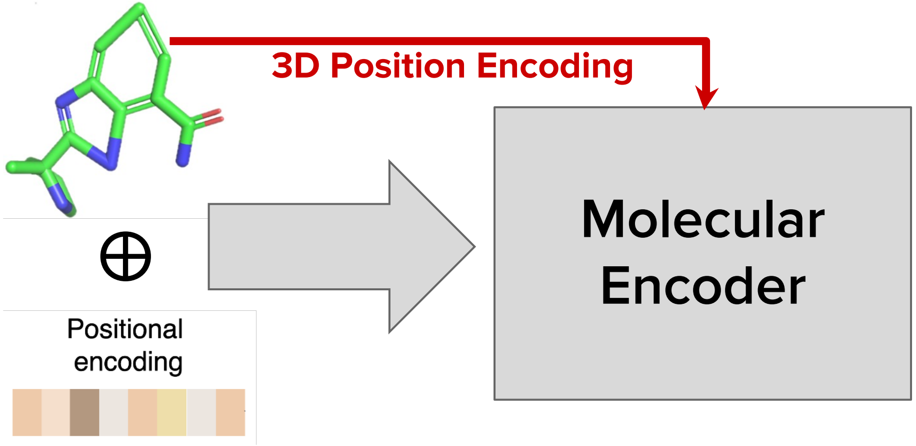
Analysis: Twisted Optimization of *Content* and *Identifier*

- Training Curve



(a) Uni-Mol

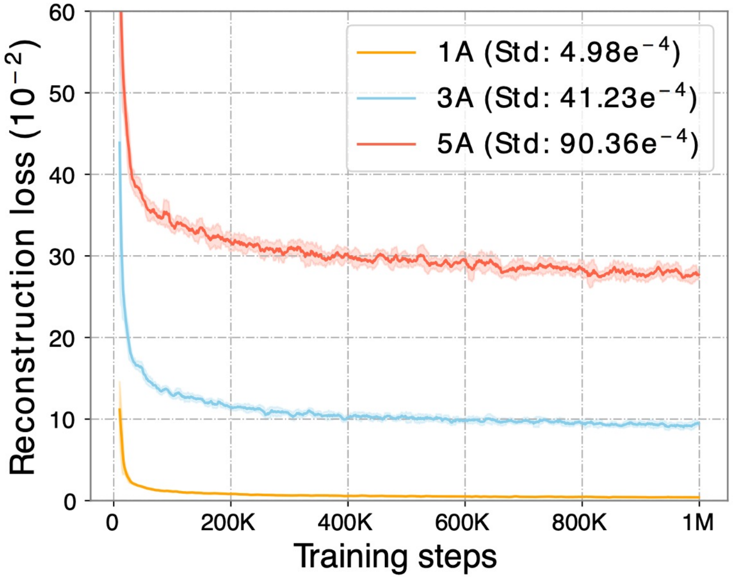
A Simple Solution: Adding PE into Encoder



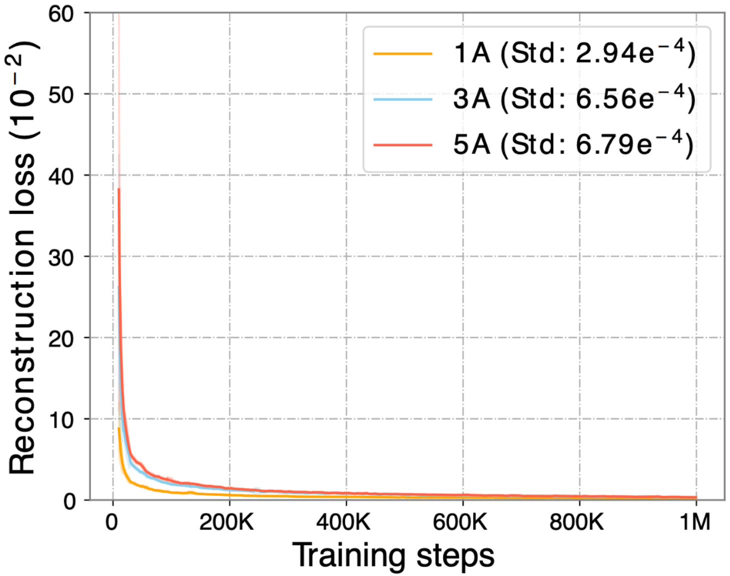
Order Information from SMILES

Analysis: Twisted Optimization of *Content* and *Identifier*

- Training Curve



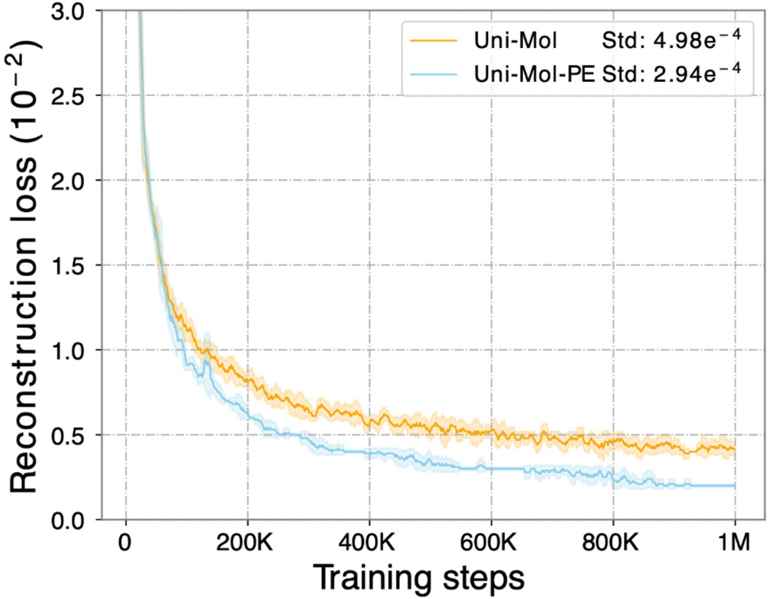
(a) Uni-Mol



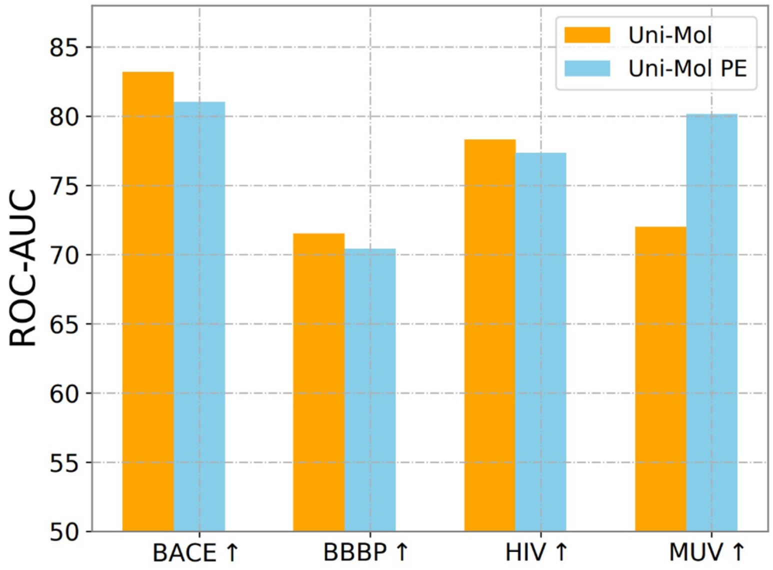
(b) Uni-Mol-PE

Analysis: Twisted Optimization of *Content* and *Identifier*

- But Uni-Mol-PE can't consistently outperform Uni-Mol on downstream tasks



(a) Training process (Noise intensity: 1Å)



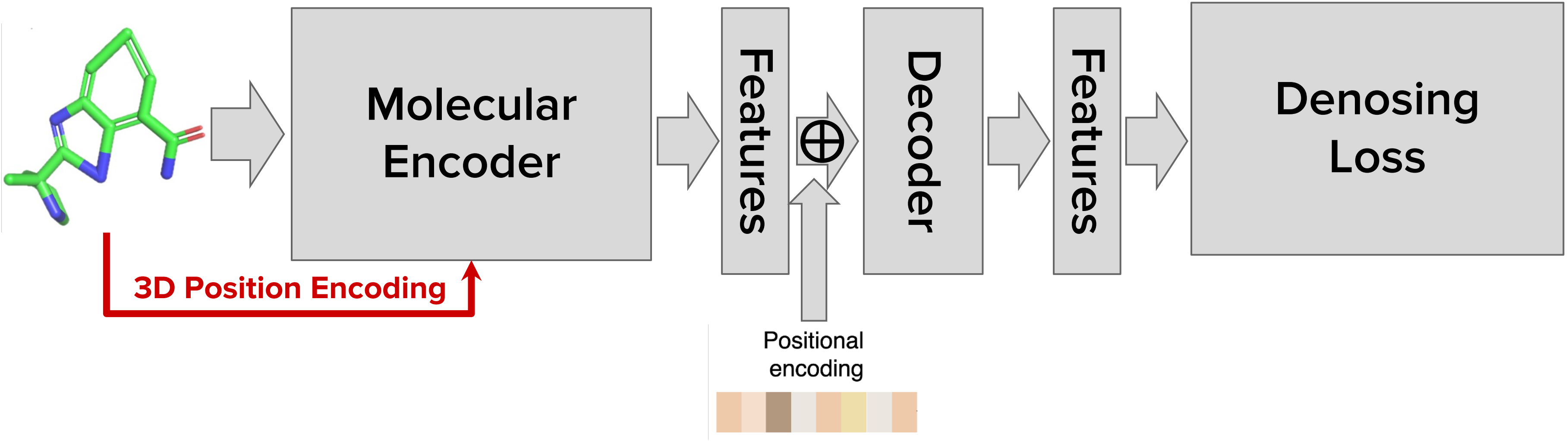
(b) Downstream performance

Positional information **does not help** the encoder learn better representations.

Analysis: Twisted Optimization of *Content* and *Identifier*

- Positional information **does not help** the encoder learn better representations.
- But it can help the model distinguish from different atoms.

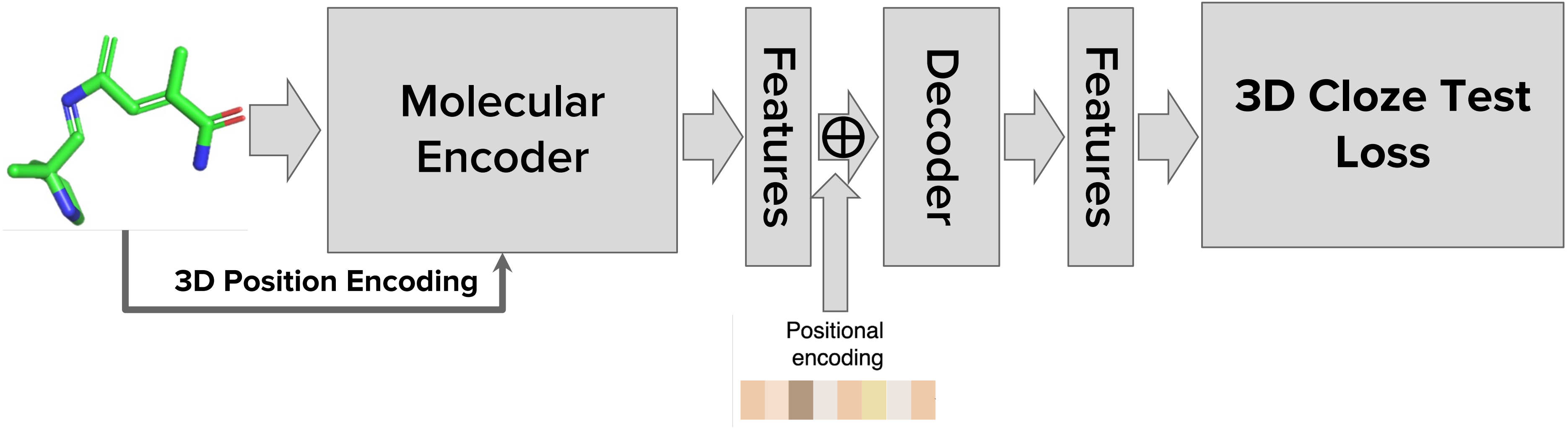
Separate the wheat from the chaff:



Analysis: Twisted Optimization of *Content* and *Identifier*

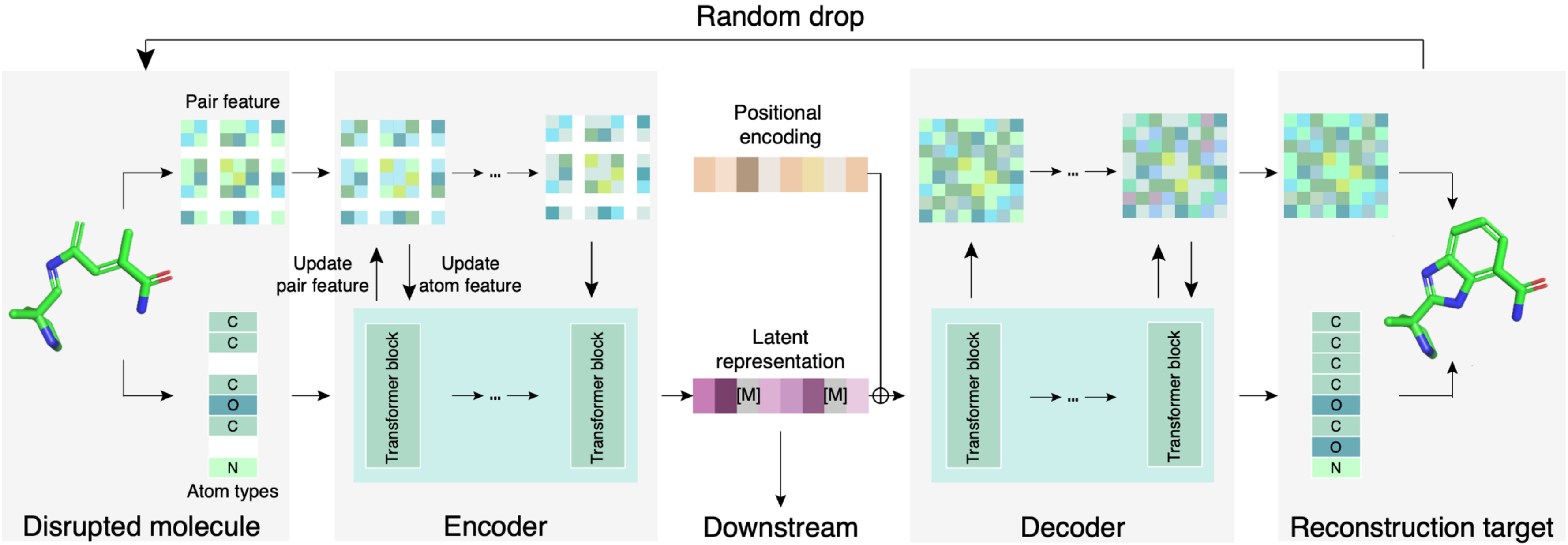
- Positional information **does not help** the encoder learn better representations.
- But it can help the model distinguish from different atoms.

Real but Partial Structure：



Method

- **Mol-AE: Auto-Encoder Based Molecular Representation Learning With 3D Cloze Test Objective**



Experiments

- The results on 9 molecule classification datasets.

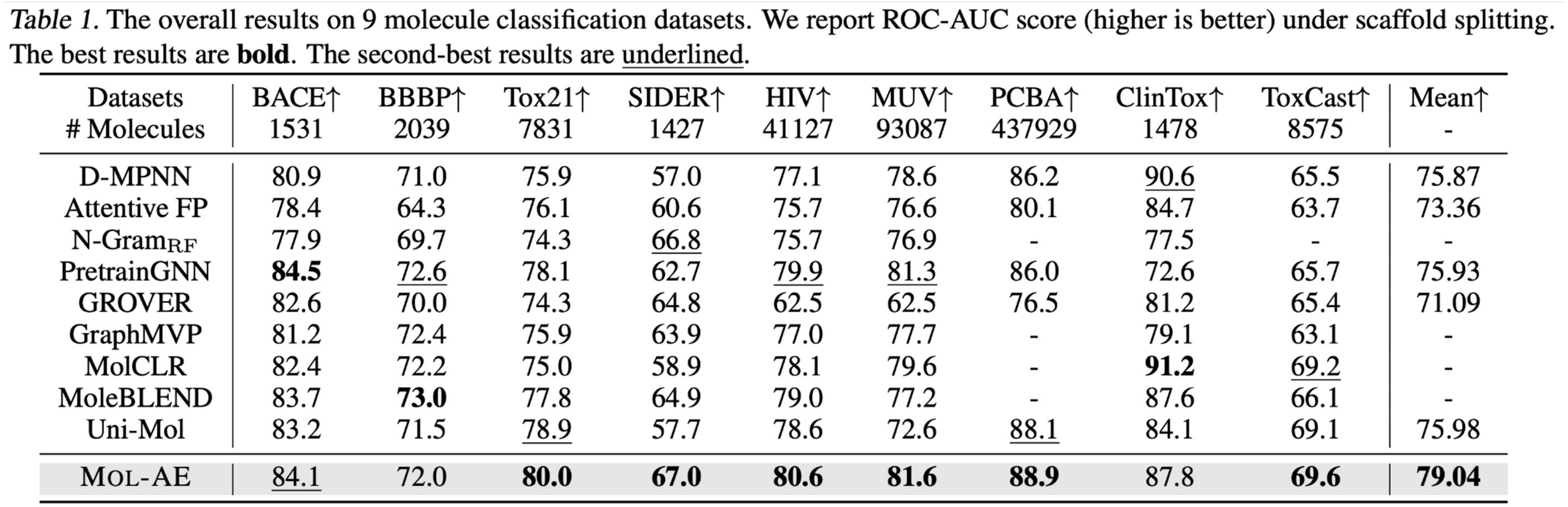
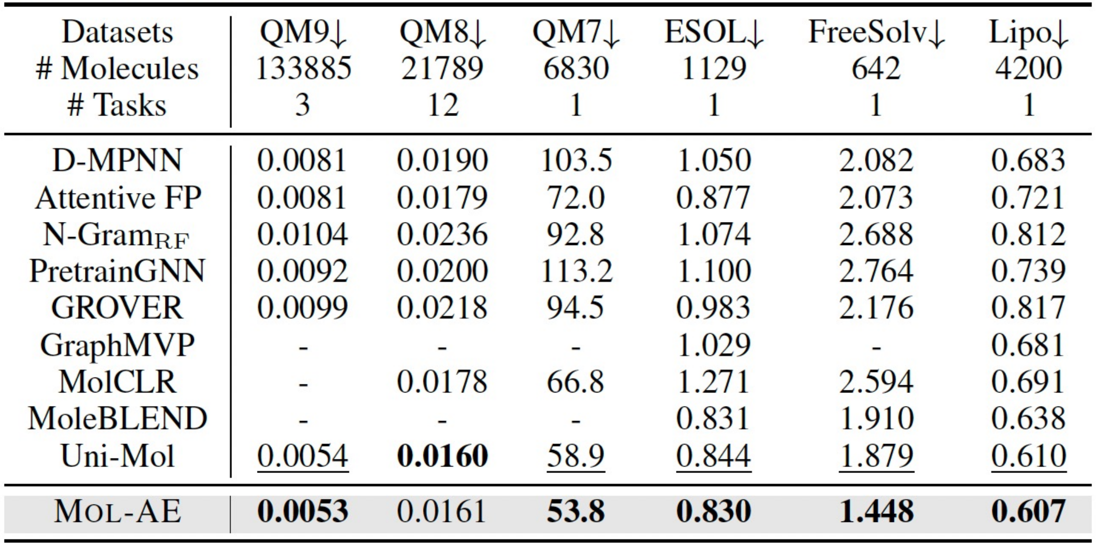


Table 1. The overall results on 9 molecule classification datasets. We report ROC-AUC score (higher is better) under scaffold splitting. The best results are **bold**. The second-best results are underlined.

Datasets # Molecules	BACE↑ 1531	BBBP↑ 2039	Tox21↑ 7831	SIDER↑ 1427	HIV↑ 41127	MUV↑ 93087	PCBA↑ 437929	ClinTox↑ 1478	ToxCast↑ 8575	Mean↑ -
D-MPNN	80.9	71.0	75.9	57.0	77.1	78.6	86.2	90.6	65.5	75.87
Attentive FP	78.4	64.3	76.1	60.6	75.7	76.6	80.1	84.7	63.7	73.36
N-Gram$_{\text{RF}}$	77.9	69.7	74.3	66.8	75.7	76.9	-	77.5	-	-
PretrainGNN	**84.5**	72.6	78.1	62.7	79.9	81.3	86.0	72.6	65.7	75.93
GROVER	82.6	70.0	74.3	64.8	62.5	62.5	76.5	81.2	65.4	71.09
GraphMVP	81.2	72.4	75.9	63.9	77.0	77.7	-	79.1	63.1	-
MolCLR	82.4	72.2	75.0	58.9	78.1	79.6	-	**91.2**	69.2	-
MoleBLEND	83.7	**73.0**	77.8	64.9	79.0	77.2	-	87.6	66.1	-
Uni-Mol	83.2	71.5	78.9	57.7	78.6	72.6	88.1	84.1	69.1	75.98
Mol-AE	84.1	72.0	**80.0**	**67.0**	**80.6**	**81.6**	**88.9**	87.8	**69.6**	**79.04**

Experiments

- The results on 6 molecule regression datasets.



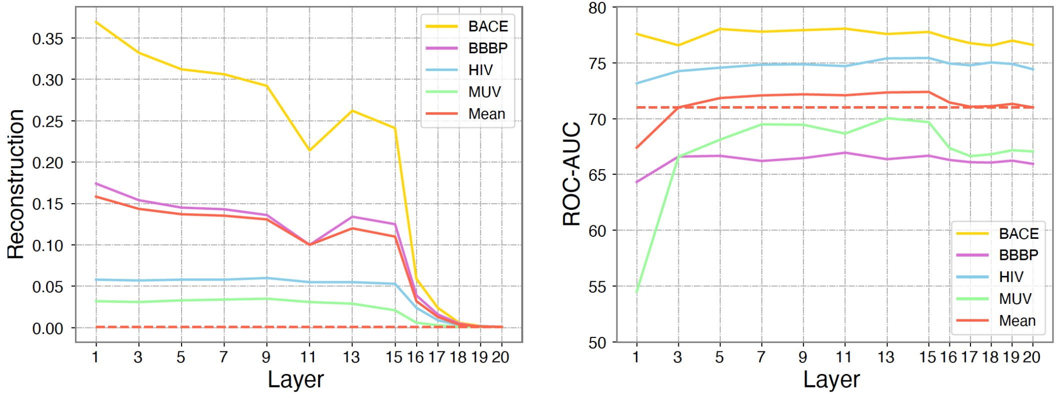
Datasets # Molecules # Tasks	QM9↓ 133885 3	QM8↓ 21789 12	QM7↓ 6830 1	ESOL↓ 1129 1	FreeSolv↓ 642 1	Lipo↓ 4200 1
D-MPNN	0.0081	0.0190	103.5	1.050	2.082	0.683
Attentive FP	0.0081	0.0179	72.0	0.877	2.073	0.721
N-Gram$_{\text{RF}}$	0.0104	0.0236	92.8	1.074	2.688	0.812
PretrainGNN	0.0092	0.0200	113.2	1.100	2.764	0.739
GROVER	0.0099	0.0218	94.5	0.983	2.176	0.817
GraphMVP	-	-	-	1.029	-	0.681
MolCLR	-	0.0178	66.8	1.271	2.594	0.691
MoleBLEND	-	-	-	0.831	1.910	0.638
Uni-Mol	0.0054	**0.0160**	58.9	0.844	1.879	0.610
MOL-AE	**0.0053**	0.0161	**53.8**	**0.830**	**1.448**	**0.607**

Analytical Experiments

- Why Auto-Encoder
- Why 3D cloze test
  - Why PE is added to the decoder
  - Why dropping

Analytical Experiments

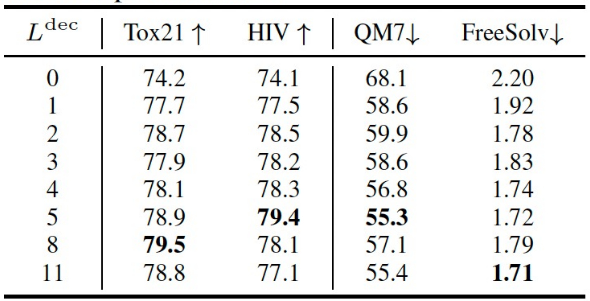
- Why Auto-Encoder
  - Shallow decoder is harmful
  - Clearer division of labor
  - Better performance



(a) Reconstruction probing.

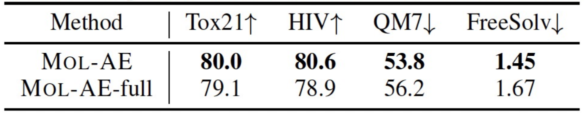
(b) Downstream task probing.

Table 3. **Decoder capacity.** Using an overly shallow decoder can harm the model's performance.



L^{dec}	Tox21 ↑	HIV ↑	QM7↓	FreeSolv↓
0	74.2	74.1	68.1	2.20
1	77.7	77.5	58.6	1.92
2	78.7	78.5	59.9	1.78
3	77.9	78.2	58.6	1.83
4	78.1	78.3	56.8	1.74
5	78.9	**79.4**	**55.3**	1.72
8	**79.5**	78.1	57.1	1.79
11	78.8	77.1	55.4	**1.71**

Table 9. Performance comparison of MOL-AE and MOL-AE-full on downstream tasks.



Method	Tox21↑	HIV↑	QM7↓	FreeSolv↓
MOL-AE	**80.0**	**80.6**	**53.8**	**1.45**
MOL-AE-full	79.1	78.9	56.2	1.67

Analytical Experiments

- Why PE is added to the decoder
  - Better convergence
  - Sequential order may be harmful

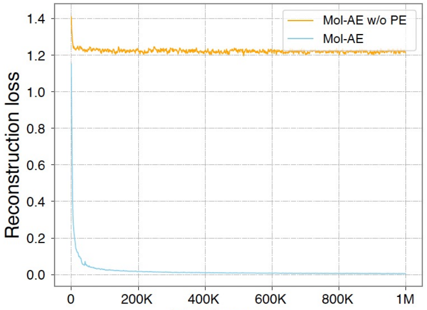
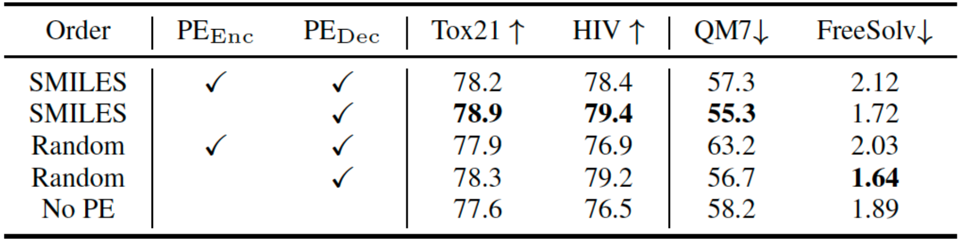
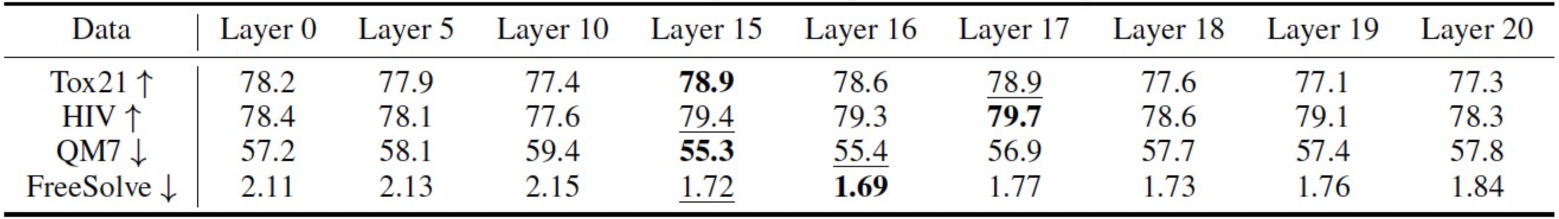


Table 4. **Sequential order information in PE.** Introducing PE in encoder will potentially harm the capacity for 3D molecular understanding.



Order	PE_{Enc}	PE_{Dec}	Tox21 ↑	HIV ↑	QM7↓	FreeSolv↓
SMILES	✓	✓	78.2	78.4	57.3	2.12
SMILES		✓	**78.9**	**79.4**	**55.3**	1.72
Random	✓	✓	77.9	76.9	63.2	2.03
Random		✓	78.3	79.2	56.7	**1.64**
No PE			77.6	76.5	58.2	1.89

Table 10. Ablation study on adding PE to different layers.



Data	Layer 0	Layer 5	Layer 10	Layer 15	Layer 16	Layer 17	Layer 18	Layer 19	Layer 20
Tox21 ↑	78.2	77.9	77.4	**78.9**	78.6	78.9	77.6	77.1	77.3
HIV ↑	78.4	78.1	77.6	79.4	79.3	**79.7**	78.6	79.1	78.3
QM7 ↓	57.2	58.1	59.4	**55.3**	55.4	56.9	57.7	57.4	57.8
FreeSolve ↓	2.11	2.13	2.15	1.72	**1.69**	1.77	1.73	1.76	1.84

Analytical Experiments

- Why dropping
  - Better performance

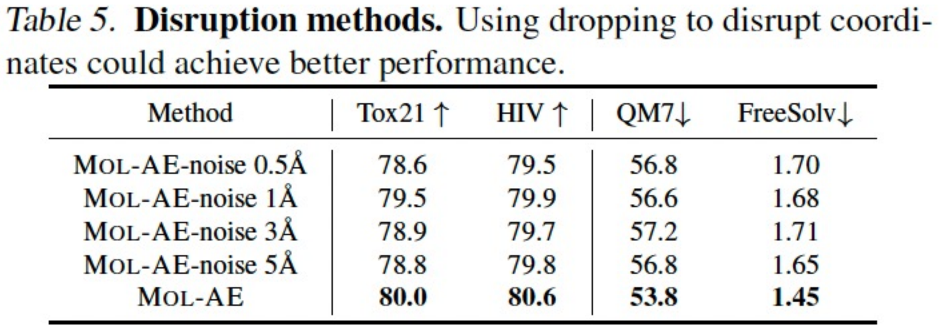


Table 5. **Disruption methods.** Using dropping to disrupt coordinates could achieve better performance.

Method	Tox21 ↑	HIV ↑	QM7↓	FreeSolv↓
Mol-AE-noise 0.5Å	78.6	79.5	56.8	1.70
Mol-AE-noise 1Å	79.5	79.9	56.6	1.68
Mol-AE-noise 3Å	78.9	79.7	57.2	1.71
Mol-AE-noise 5Å	78.8	79.8	56.8	1.65
Mol-AE	**80.0**	**80.6**	**53.8**	**1.45**

For a more theoretical explanation, please refer to
Yu Meng, et al. "Representation Deficiency in Masked Language Modeling" ICLR (2024).

Thanks

Paper: https://www.biorxiv.org/content/10.1101/2024.04.13.589331v1
Project: https://github.com/yjwtheonly/MolAE