中国人民大学
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Uni-Mol3: A Multi-Molecular Foundation Model for Advancing Organic Reaction Modeling

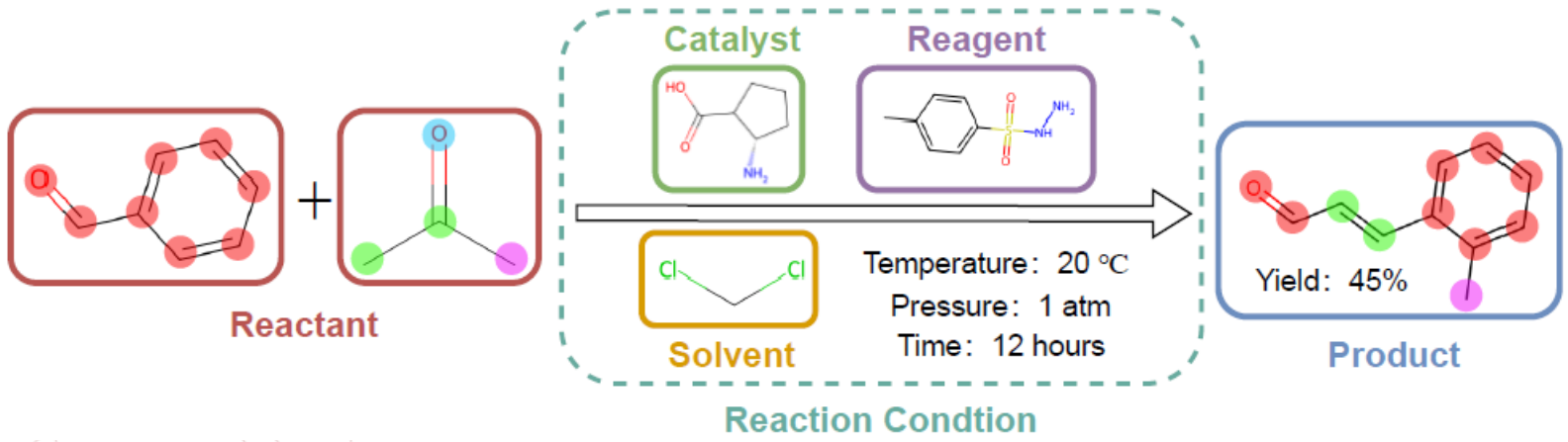
Lirong Wu, Junjie Wang, Zhifeng Gao, Xiaohong Ji, Rong Zhu,
Xinyu Li, Linfeng Zhang, Guolin Ke, Weinan E

Fanmeng Wang

2025-9-5

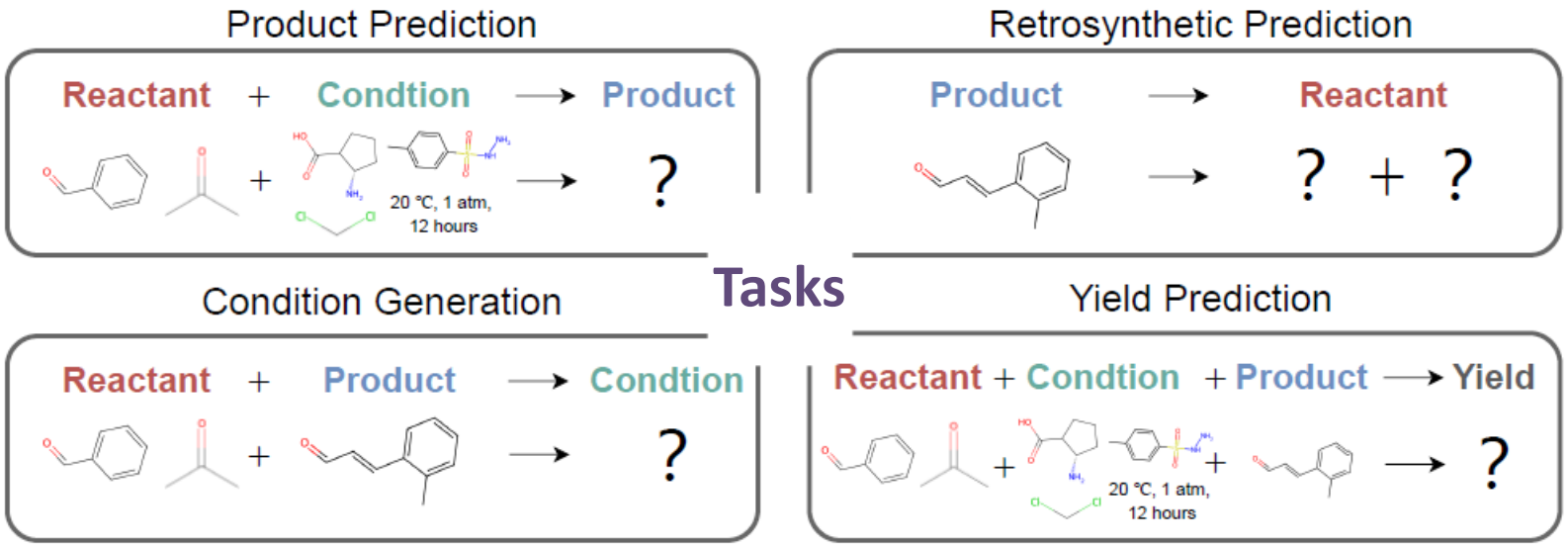
Introduction

- **Organic reaction**, as the cornerstone of the modern chemical industry, plays an irreplaceable role in new material development and drug discovery.
  - It is a process in which molecules are transformed by recombination of atoms.
  - Its cover three elements: reactants, reaction conditions, and products.
  - **Reaction conditions** are the key variables that determine the reaction path and product yields, including temperature, pressure, catalysts, solvents, reagents, etc.



Introduction

- While **single-molecular foundation models** like Uni-Mol and Uni-Mol2 have made remarkable progress in the representation learning of individual molecules, their extension to multi-molecular systems has been largely underexplored.
- In this context, various **organic reaction tasks** become particularly challenging, since they not only involve intermolecular interactions but are also heavily influenced by reaction conditions, which cannot be handled by single-molecule models.



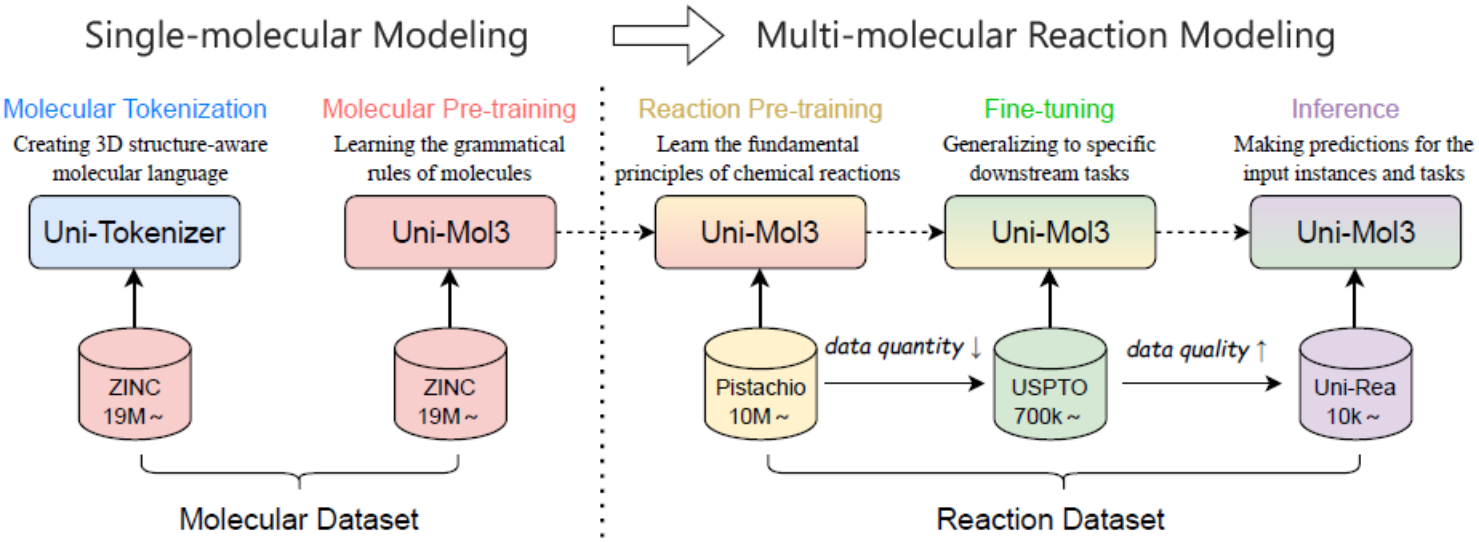
Introduction

- **Challenges**
  - Despite great progress in applying Transformer architectures with SMILES (i.e., framing organic reaction tasks as **language translation problems**), SMILES fail to encode full-atom 3D coordinates and stereochemical details.
  - The integration of single-molecular grammars with multi-molecular dependencies to build **a unified pre-training framework** remains a significant hurdle.
  - The severe scarcity of high-quality datasets, inconsistent annotation and standardized benchmarks.
- Building on Uni-Mol2's superior single-molecular representation capabilities, this work introduces **Uni-Mol3**, a novel deep learning framework that enables unified multi-molecular reaction modeling via a hierarchical pipeline.

Introduction

■ Uni-Mol3

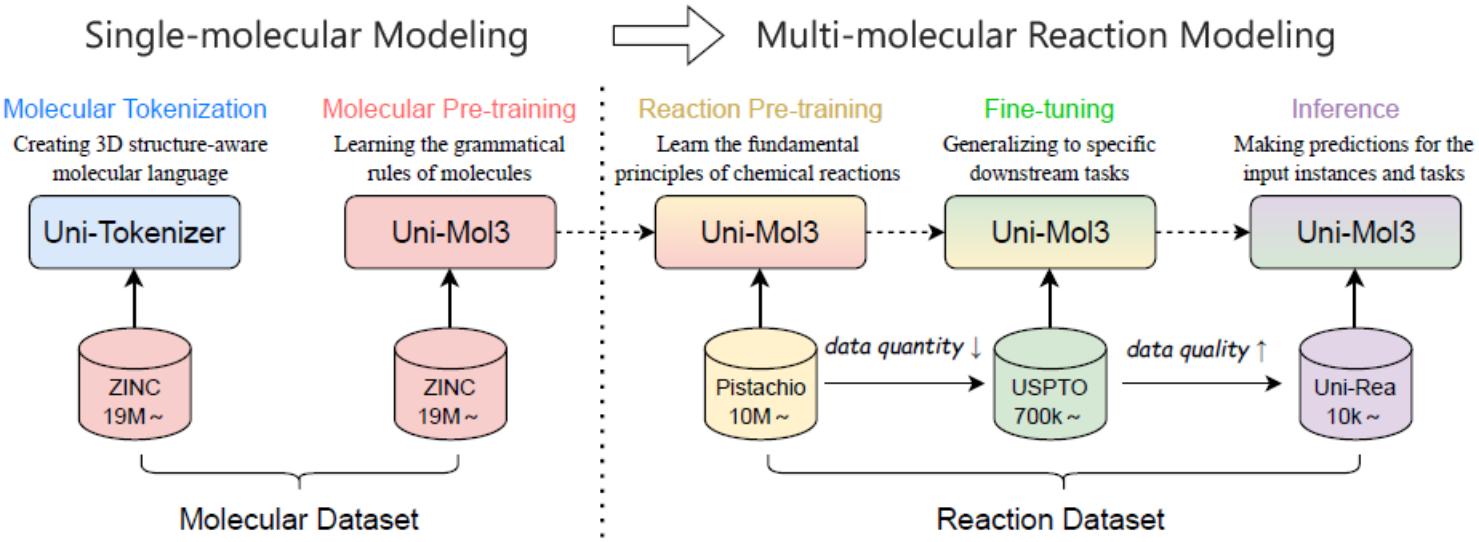
➢ To address the inherent limitation of SMILES in capturing spatial information, we propose a **3D structure-aware molecular language system**, where a multi-scale Uni-Tokenizer quantizes 1D atomic features, 2D graph structures, and 3D coordinates into discrete tokens.



Introduction

■ Uni-Mol3

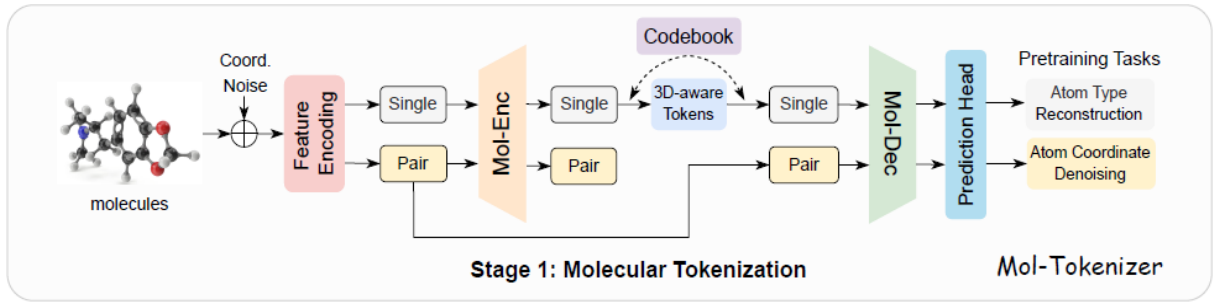
➢ Uni-Mol3 employs a **two-tier pre-training strategy**: molecular pre-training learns single-molecular grammars, reaction pre-training captures fundamental principles of multi-molecular reactions, thus forming a progressive learning framework from molecular grammars to reaction mechanisms.

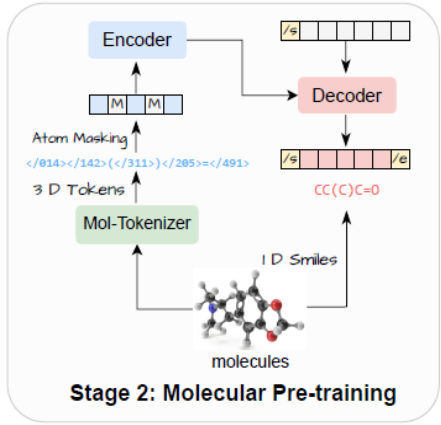


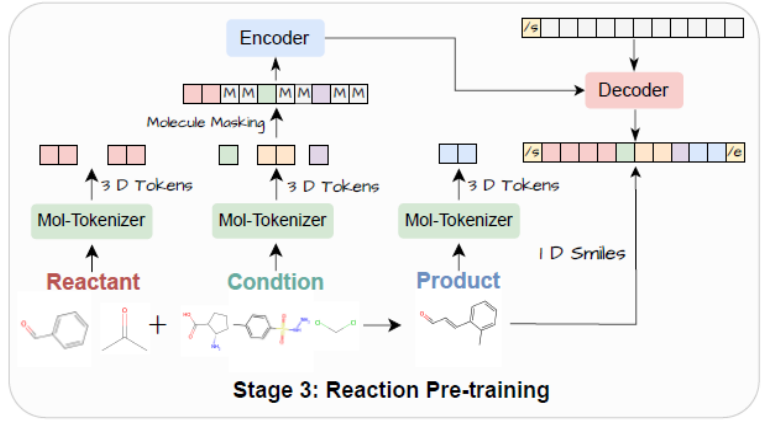
Method

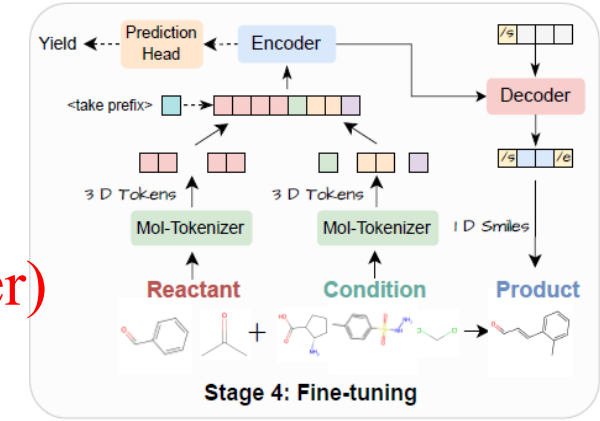
Overview

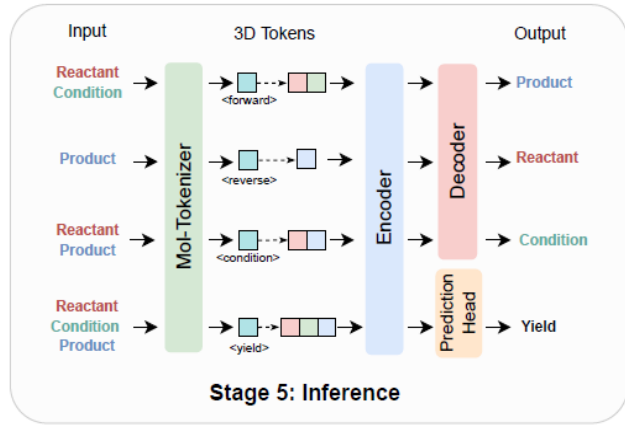
- **Mol-Tokenizer**: Initialized by the pre-trained Uni-Mol2, we further train a Mol-Tokenizer to quantize multi-scale molecular information into discrete tokens (Finite Scalar Quantization).
- **Molecular pre-training**: atom-level masked modeling
- **Reaction pre-training**: molecular-level masked modeling
- **Backbone**: Uni-Mol3 adopts T5 (Text-to-Text Transfer Transformer) as its backbone.











Method

■ Stage 1: Molecular Tokenization

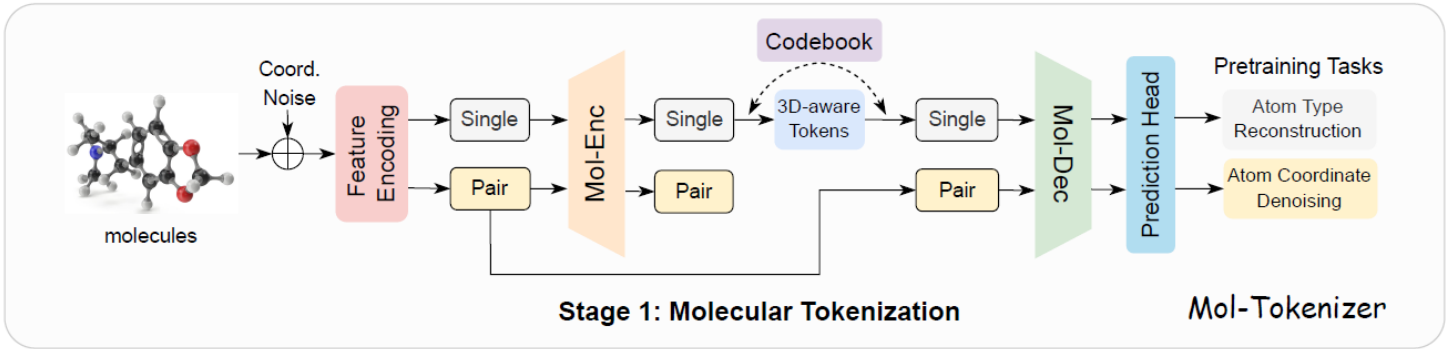
➢ Feature Encoding:

- We employ RDKit to obtain atom token, atom degree, and atom types, then the single representation of atom i is initialized as follows:

$$x^i_{\text{single}} = \text{Embedding}\left(x^i_{\text{token}}\right) + \text{Embedding}\left(x^i_{\text{degree}}\right) + \text{Embedding}\left(x^i_{\text{type}}\right)$$

- The pair representation between atom i and atom j is initialized through bond type, the shortest path distance, and the Euclidean distance encoded by the Gaussian kernel approach with pair type, i.e.,

$$x^{i,j}_{\text{pair}} = \text{Embedding}\left(e^{i,j}\right) + \text{Embedding}\left(x^{i,j}_{\text{SPD}}\right) + x^{i,j}_{\text{dis}}$$



Method

■ Stage 1: Molecular Tokenization

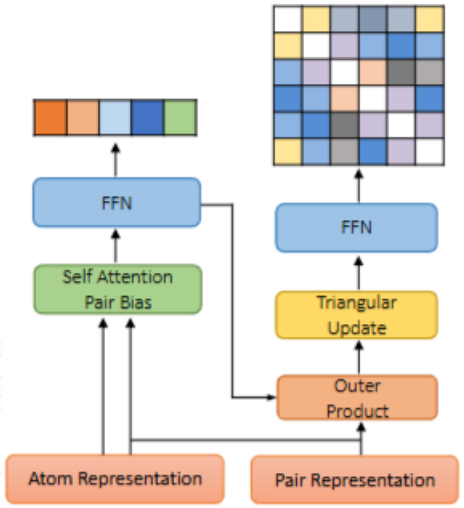
➢ Encoder and Decoder:

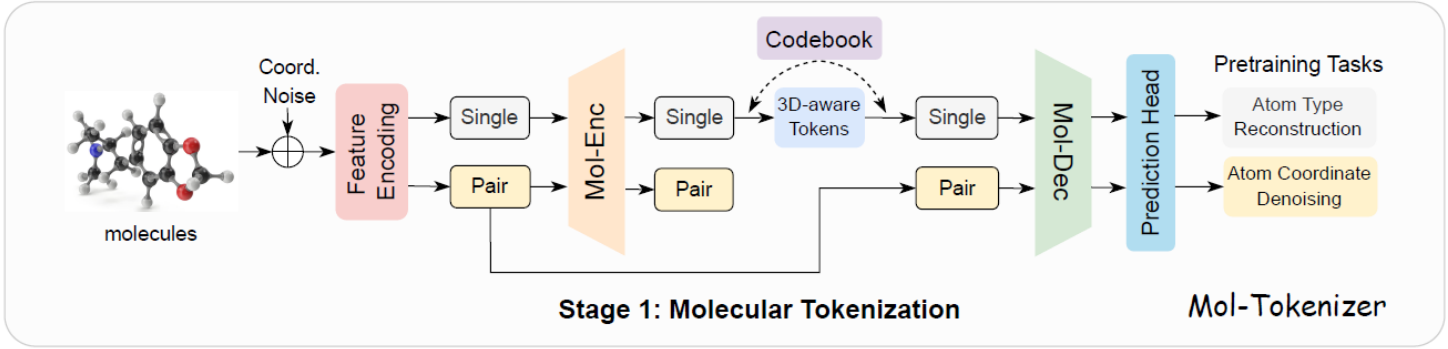
- The encoder and decoder in Mol-Tokenizer adopt the same backbone as Uni-Mol2, each containing several two-track transformer layers.
- We initialize atom and pair embeddings of the first layer as follows:

$$h^{(0)}_{\text{single}}, h^{(0)}_{\text{pair}} = x_{\text{single}}, x_{\text{pair}}$$

- Then each layer iteratively updates single and pair representations, i.e.,

$$h^{(l)}_{\text{single}}, h^{(l)}_{\text{pair}} = \psi^{(l)}(h^{(l-1)}_{\text{single}}, h^{(l-1)}_{\text{pair}})$$





Method

■ Stage 1: Molecular Tokenization

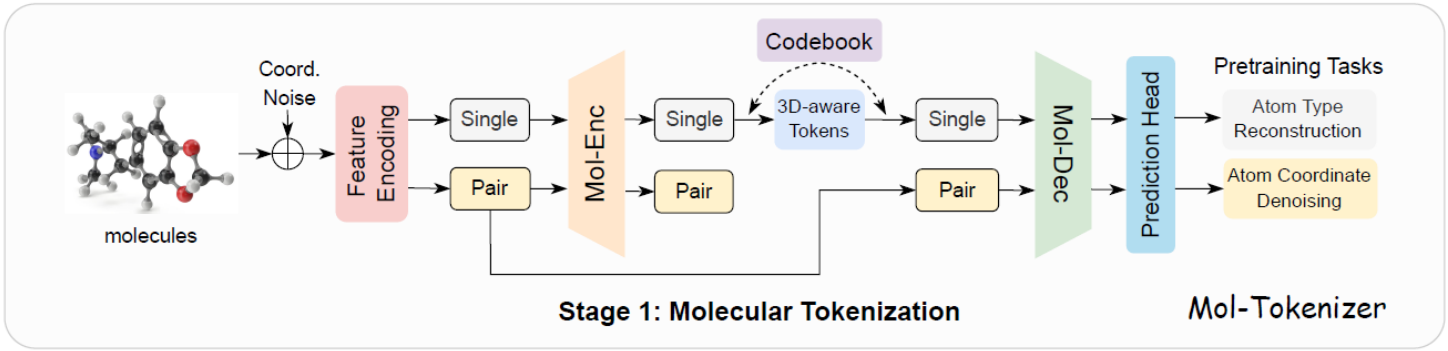
➢ Quantization:

- We use Finite Scalar Quantization to quantize the single representation of each atom i from the encoder into a finite set of codewords.

$$f(h^i_{\text{single}}) = \lfloor L/2 \rfloor \tanh(h^i_{\text{single}})$$

$$s_i = \text{round}\left(f(h^i_{\text{single}})\right) \in \mathbb{R}^d$$

where each channel in s_i takes one of L unique values. Thereby, we have a codebook $s_i \in \mathcal{A}$ ($|\mathcal{A}| = L^d$) that is the product of d per-channel codebook sets. The vectors in $\mathcal{A}$ can be enumerated by a simple bijection from any s_i to an integer z in $\{1, \cdots, L^d\}$.



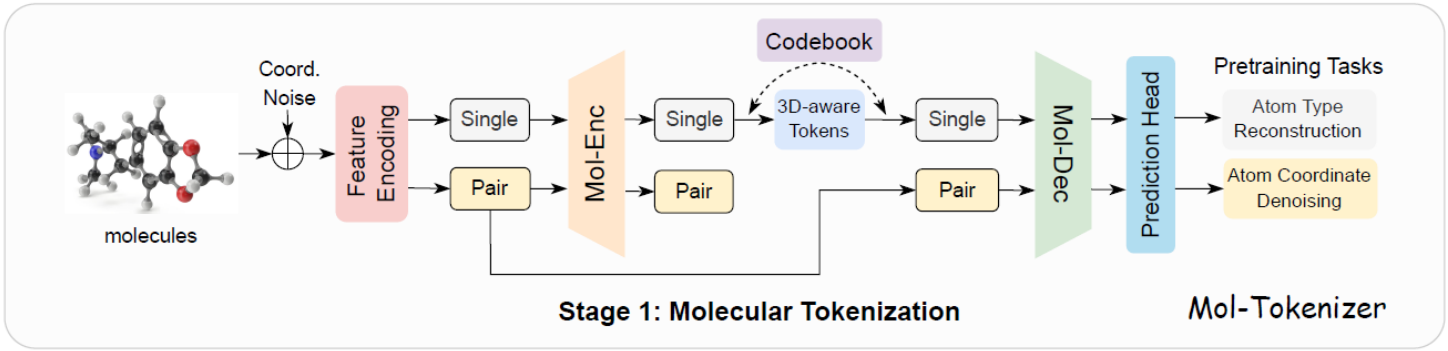
Method

■ Stage 1: Molecular Tokenization

➢ Trarining:

- We train the Mol-Tokenizer with two complementary tasks, i.e., atom type reconstruction and atom coordinate denoising.
- For the **decoder**, the discrete 3D token z is used to initialize the single representation through Embedding(z). To prevent leakage of atom type information, we initialize the pair representations as x_{pair} rather than using the output pair representations of the encoder.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{type}} + \mathcal{L}_{\text{coor}} + \mathcal{L}_{\text{distance}}$$

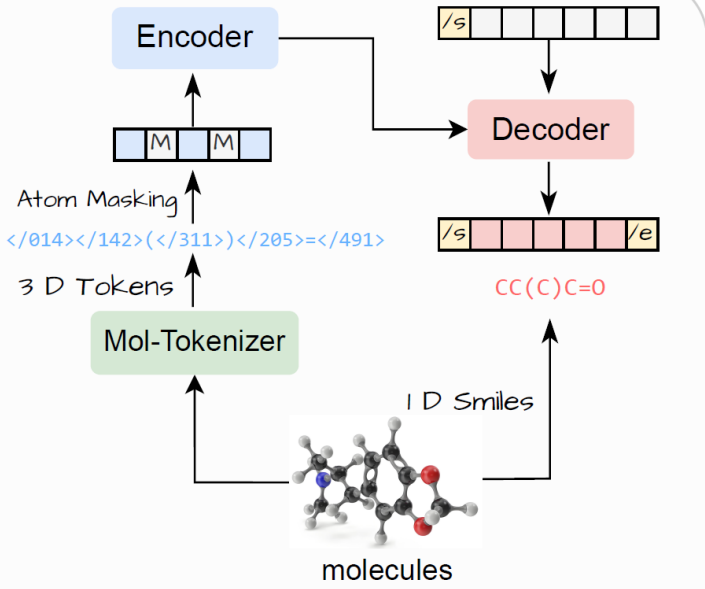


Method

■ Stage 2: Molecular Pre-training

- To enable the model to learn the molecular grammars and chemical semantic space, we first pre-train Uni-Mol3 at the single-molecular level.
- We first transform each input molecule into **corresponding 3D tokens** using the trained Uni-Tokenizer, i.e.,

$$Z = \{z_1, z_2, \ldots, z_N\} = \text{Uni-Tokenizer}(M)$$



Stage 2: Molecular Pre-training

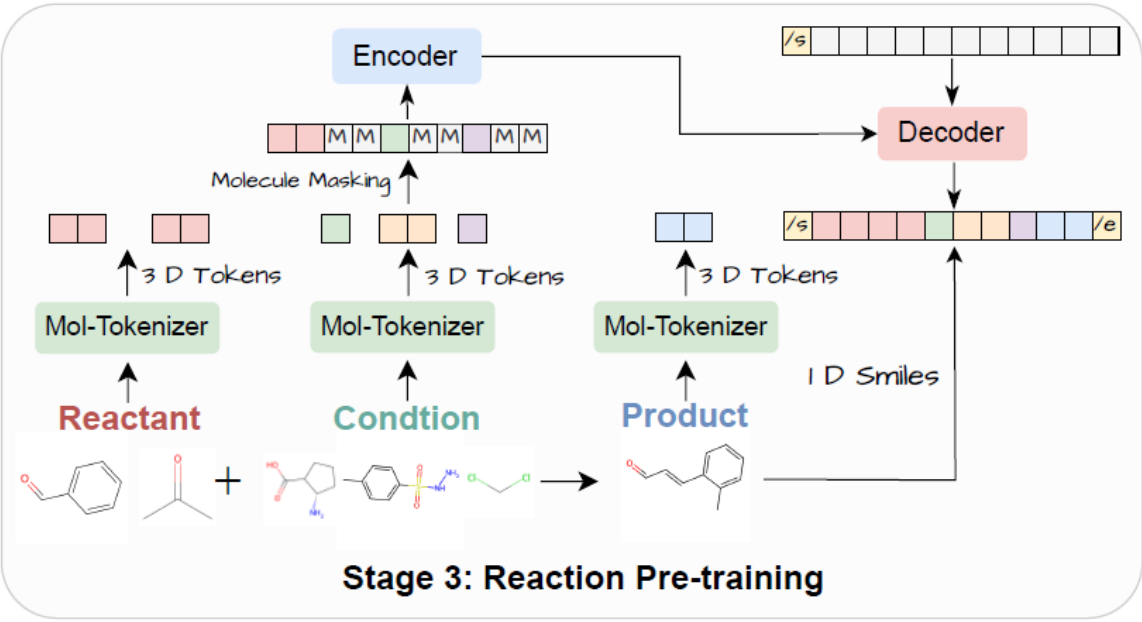
- The **encoder** takes the masked 3D tokens $\hat{Z}$ as input to generate a conditional embedding c
- The **decoder** generates 1D smiles X autoregressively under the condition c with the following optimization objective:

$$\mathcal{L}_{\text{Mol-Pre}} = -\sum_{i=1}^{N} \log p\left(x_i \mid x_1, x_2, \ldots, x_{i-1}, \mathbf{c}\right)$$

Method

■ Stage 3: Reaction Pre-training

- To enable the model to learn the syntax (reaction rules) and semantics (chemical meaning) of the chemical reaction, we further extends the pre-training from the single-molecular level to the multi-molecular level.
- We use the **same pretraining pipeline** as in molecular pretraining, but replace atom-level masking modeling with molecule-level masking modeling.



Stage 3: Reaction Pre-training

$$Z_i = \text{Uni-Tokenizer}(M_i), \qquad \forall M_i \in \mathcal{R}.$$

$$Z_{\text{Reac}} = \left[Z_1, Z_2, Z_3, \cdots, Z_{|\mathcal{R}|}\right]$$

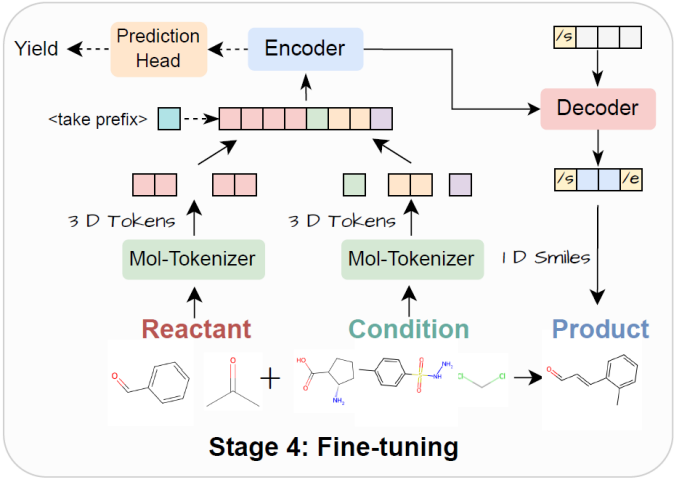
$$\widehat{Z}_{\text{Reac}} = \left[[\text{M}], Z_2, [\text{M}], \cdots, Z_{|\mathcal{R}|}\right]$$

$$\mathcal{L}_{\text{Reac-Pre}} = -\sum_{i=1}^{|\mathcal{R}|}\sum_{j=1}^{N_i} \log p\left(x_{i,j} \mid x_{i,1}, x_{i,2}, \ldots, x_{i,j-1}, \{x_{i-1,k}\}_{k=1}^{N_{i-1}}, \ldots, \{x_{1,k}\}_{k=1}^{N_1}, \mathbf{c}\right)$$

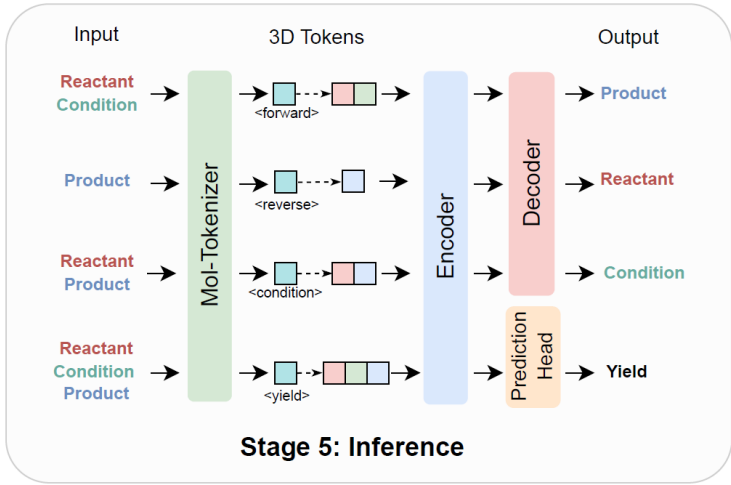
Method

■ Stage 4&5: Fine-tuning and Inference

- For **generative tasks**, such as product prediction, retrosynthetic prediction, and condition generation, the decoder will be directly used to generate the target molecules in an autoregressive manner.
- For **regression or classification tasks**, a separate prediction head is used to predict the targets from the encoder's output.
- To **distinguish between different tasks**, a task-specific prefix token is added as a prompt to the front of the 3D token sequence output by the Uni-Tokenizer.



Stage 4: Fine-tuning

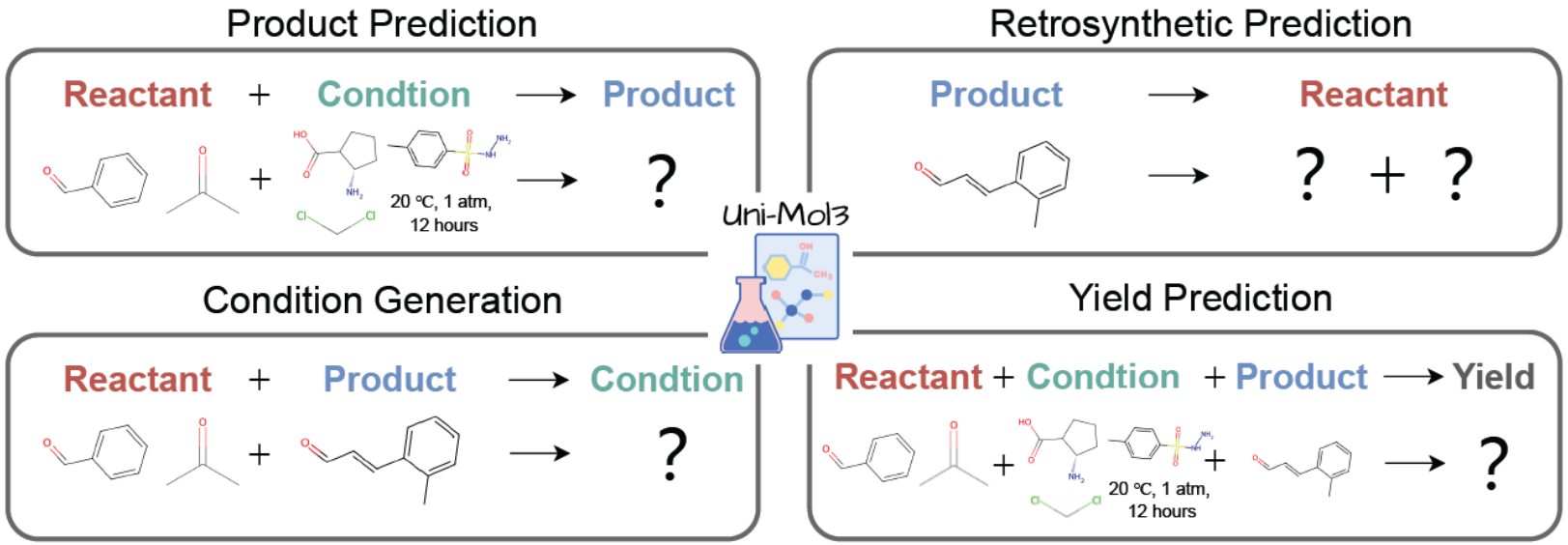


Stage 5: Inference

Method

■ Stage 4&5: Fine-tuning and Inference

➢ We mainly focuses on four representative reaction tasks:



Here, **Product Prediction** comprises two subtasks. One is labeled with the prompt token < forward-sep >, where reactants and conditions are input as separate entities. The other is labeled with the prompt token <forward-mixed>, indicating that reactants and conditions are mixed in input.

Experiment

■ Datasets

- ➢ **Single-molecular modeling** (i.e., Mol-Tokenizer and molecular pre-training)
  - Uni-Mol dataset, containing ~19M molecules, primarily from the ZINC and Pubmed databases.
- ➢ **Reaction Pretraining**
  - We use a large-scale reaction dataset derived from the Pistachio database developed by NextMove Software.
  - After preprocessing (removing reactions involving invalid molecules and those containing molecules with more than 80 atoms), we obtain a new Pistachio-full dataset containing 11,973,789 reactions.
  - We further split 10,000 reactions individually for downstream task testing, i.e., Pistachio-FP, Pistachio-RS, Pistachio-CG.

Experiment

■ Datasets

➢ Downstream Tasks

Table 1: The statistical information of datasets in this work.

Dataset	# Train	# Valid	# Test	# All	Downstream Task
USPTO-MIT [25]	407,791	29,915	39,876	477,582	Product Prediction
SMol-Reactions-FP [36]	116,360	-	943	117,303	Product Prediction
Pistachio-FP	11,963,789	-	10,000	11,973,789	Product Prediction
USPTO-50k [40]	40,022	5,004	5,004	50,030	Retrosynthesis
SMol-Reactions-RS [36]	128,684	-	1,000	129,684	Retrosynthesis
Pistachio-RS	11,963,789	-	10,000	11,973,789	Retrosynthesis
USPTO-500-MT [15]	116,360	12,937	14,238	143,535	Condition Generation
USPTO-Condition [41]	543,854	67,964	67,992	679,810	Condition Generation
Pistachio-CG	9,668,808	-	7,997	9,676,805	Condition Generation
Buchwald-Hartwig Test1	3,057	-	898	3,955	Reaction Yield Prediction
Buchwald-Hartwig Test2	3,055	-	900	3,955	Reaction Yield Prediction
Buchwald-Hartwig Test3	3,058	-	897	3,955	Reaction Yield Prediction
Buchwald-Hartwig Test4	3,055	-	900	3,955	Reaction Yield Prediction

For the Pistachio-CG dataset, we further filter out reactions whose conditions are unavailable or unknown, resulting in fewer samples.

Experiment

■ Metrics

- **Regression Tasks** (i.e., Reaction Yield Prediction)
  - Mean Absolute Error (MAE) & Mean Squared Error (MSE) & Coefficient of Determination (R^2)
- **Generative Tasks** (i.e., Product Prediction, Retrosynthesis and Condition Generation)
  - Top-1 Accuracy
  - Levenshtein Distance (LEV), measuring the minimum edits (insert, delete, substitute) to align two SMILES strings
  - Tanimoto coefficient of molecular molar fingerprinting (MFP-TC) between predicted and ground-truth SMILES
  - Invalidity Rate

Experiment

■ Task 1: (Forward) Product Prediction

Table 3: Performance comparison for product prediction on the USPTO-MIT dataset, where reactant-condition separated and mixed are separately evaluated. The best and second results are marked as **bold** and underline. (same for all the tables below)

Model	USPTO-MIT (Mixed)				USPTO-MIT (Seperated)			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	88.3	0.543	0.971	0.32	89.0	0.445	0.975	0.26
Chemformer [27]	88.6	0.514	0.976	0.25	89.8	0.428	0.979	0.17
T5Chem [15]	88.9	0.527	0.974	0.20	90.2	0.414	0.981	**0.10**
Uni-Mol3 (ours)	**89.6**	**0.485**	**0.979**	**0.15**	**90.8**	**0.387**	**0.983**	0.15

Table 4: Performance comparison for product prediction on the Pistachio-FP dataset, where reactant-condition separated and mixed are separately evaluated.

Model	Pistachio-FP (Mixed)				Pistachio-FP (Seperated)			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	88.3	0.637	0.977	0.36	90.3	0.529	0.982	0.33
Chemformer [27]	90.3	0.575	0.980	0.18	91.8	0.447	0.985	0.14
T5Chem [15]	90.9	0.560	0.982	0.15	92.1	0.428	0.986	0.18
Uni-Mol3 (ours)	**91.7**	**0.462**	**0.985**	**0.09**	**93.0**	**0.374**	**0.988**	**0.07**

Table 5: Results for product prediction on the SMol-Reactions-FP dataset.

Model	SMol-Reactions-FP			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	32.8	10.314	0.646	4.54
Chemformer [27]	36.9	7.849	0.718	0.72
T5Chem [15]	37.2	8.030	**0.735**	0.69
PRESTO [36]	35.4	9.582	0.685	1.65
Uni-Mol3 (ours)	**38.7**	**7.014**	0.732	**0.64**

Uni-Mol3 outperforms existing baselines significantly—particularly in the LEV and invalidity rate metrics

Experiment

■ Task 2: Retrosynthetic Prediction

Table 6: Retrosynthesis results on the Pistachio-RS and USPTO-50k datasets.

Model	Pistachio-RS				USPTO-50k			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	72.6	2.554	0.953	0.56	42.6	4.486	0.911	0.79
Chemformer [27]	74.6	2.374	0.957	0.24	**52.3**	4.218	0.908	0.16
T5Chem [15]	75.2	2.247	0.959	0.32	46.2	3.959	0.916	0.32
Uni-Mol3 (ours)	**76.9**	**2.145**	**0.963**	**0.18**	49.0	**3.653**	**0.924**	**0.06**

Table 7: Results for retrosynthetic prediction on the SMol-Reactions-RS dataset.

Model	SMol-Reactions-RS			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	23.9	15.382	0.659	2.73
Chemformer [27]	26.5	12.017	0.714	0.24
T5Chem [15]	28.0	10.593	0.758	0.26
PRESTO [36]	27.7	11.229	0.745	1.15
Uni-Mol3 (ours)	**29.1**	**9.933**	**0.786**	**0.20**

Uni-Mol3 demonstrates remarkable adaptability in the retrosynthetic prediction task, particularly excelling on the SMol-Reactions-RS dataset

Experiment

■ Task 3: Condition Generation

Table 8: Results for condition generation on the Pistachio-CG dataset.

Model	Pistachio-CG			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	40.4	7.272	0.810	0.11
Chemformer [27]	42.1	6.947	0.819	0.04
T5Chem [15]	43.3	6.705	0.823	0.06
Uni-Mol3 (ours)	**44.4**	**6.482**	**0.827**	**0.03**

Table 9: Results for condition generation on USPTO-500-MT and USPTO-Condition.

Model	USPTO-500-MT				USPTO-Condition			
	Top-1 (%)	LEV	MFP-TC	Invalid (%)	Top-1 (%)	LEV	MFP-TC	Invalid (%)
Molformer [13]	19.9	9.773	0.694	0.068	25.6	5.439	0.739	0.031
Chemformer [27]	24.1	8.655	0.707	0.027	29.3	5.215	0.744	0.012
T5Chem [15]	**24.9**	8.541	0.712	0.012	29.8	**5.087**	0.747	0.004
Uni-Mol3 (ours)	24.5	**8.523**	**0.715**	**0.007**	**30.5**	5.157	**0.748**	**0.001**

Uni-Mol3 outperforms all baseline models across all four evaluation metrics, highlighting its comprehensive superiority in handling open-ended condition generation.

Experiment

■ Task 4: Reaction Yield Prediction

Table 10: Results for yield prediction on 4 test sets of the Buchwald-Hartwig dataset.

Model	Test1			Test2		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
DRFP	8.224	12.048	0.810	7.906	11.749	0.828
Chemprop	8.531	12.406	0.798	9.444	12.710	0.780
YieldBert	6.705	10.849	0.838	7.457	10.631	0.842
T5Chem	8.145	11.837	0.815	6.075	8.784	0.895
Uni-Mol3 (ours)	**5.867**	**9.680**	**0.874**	**5.420**	**8.170**	**0.909**

Model	Test1			Test2		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
DRFP	9.525	14.880	0.719	13.240	19.037	0.496
Chemprop	10.340	15.280	0.708	15.783	20.155	0.429
YieldBert	9.109	14.136	0.746	13.045	18.639	0.503
T5Chem	8.977	13.892	0.765	12.952	18.711	**0.610**
Uni-Mol3 (ours)	**8.856**	**13.506**	**0.769**	**12.740**	**18.245**	0.525

Uni-Mol3 demonstrates overall superiority over baseline models, with the top performance on 11 out of 12 metrics

Experiment

■ Ablation Study and Analysis

- We conduct analysis of the first three stages in the hierarchical, including Uni-Tokenizer, molecular pre-training, and reaction pre-training.
- Here, for the "**w/o Uni-Tokenizer**" set of experiments, we directly use SMILES strings as inputs.

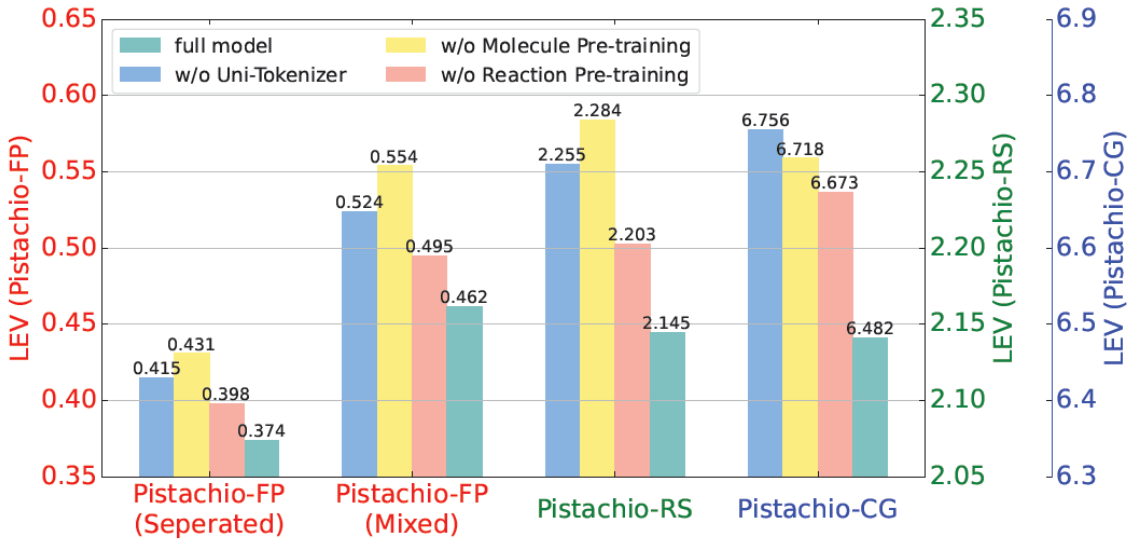


Fig. 5: Ablation study on Uni-Tokenizer, molecular pre-training, and reaction pre-training on three Pistachio datasets with the Levenshtein Distance (LEV) as a metric.

Conclusion

- We propose **Uni-Mol3**, a deep learning framework for multi-molecular organic reaction modeling.
- By integrating 3D structure-aware molecular tokenization, hierarchical pre-training, and prompt-aware fine-tuning, the model achieves **state-of-the-art performance** across diverse reaction tasks.
- By **unifying single and multi-molecular modeling**, Uni-Mol3 defines a versatile framework for intelligent reactions, that promises to advance data-driven innovation in organic synthesis and accelerate its translation to industrial applications.

Thank You for listening!

Fanmeng Wang

2025-9-5