PRESTO: Progressive Pretraining Enhances Synthetic Chemistry Outcomes

He Cao♠, Yanjun Shao♠, Zhiyuan Liu, Zijing Liu, Xiangru Tang, Yuan Yao, Yu Li

♠ equal contribution

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Yale

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY 粤港澳大湾区数字经济研究院

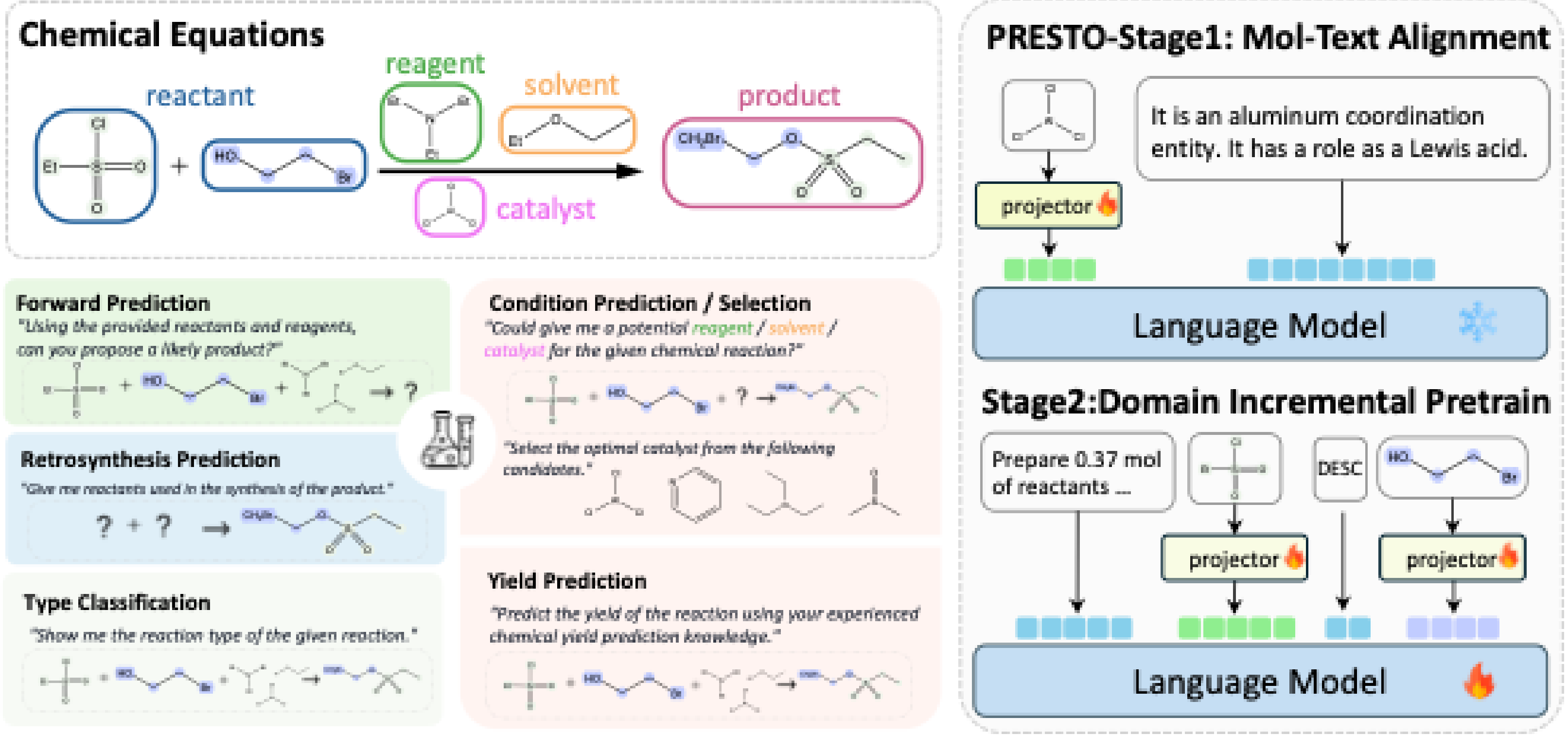
NUS National University of Singapore

Motivation

- ➢ Multimodal Large Language Models (MLLMs) in in biomolecular disciplines neglect **multiple molecular graph interactions** and lack clear downstream tasks to validate framework effectiveness.
- ➢ The effectiveness of MLLMs is influenced by inconsistent pretraining strategies, underscoring the necessity for systematic evaluation to optimize performance in synthetic chemistry.

Progressive Pretraining Strategy: PRESTO

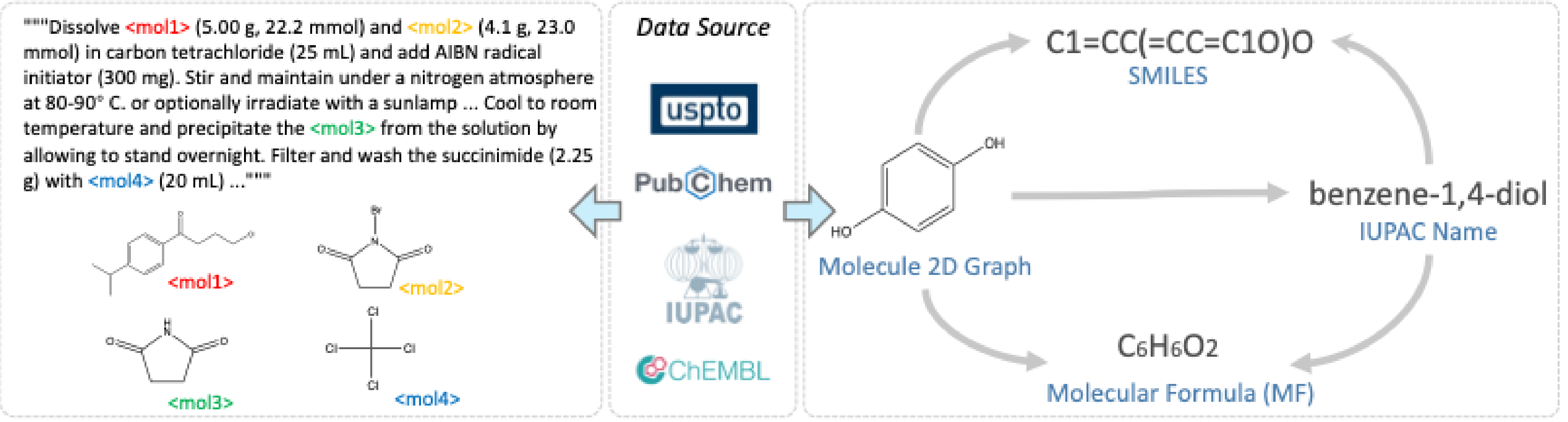
- ✓ Bridges **molecule-text** modality gap
- ✓ Enhances **multi-graph** understanding
- ✓ Tailored for synthetic chemistry tasks


Chemical Equations
reactant
reagent
solvent
product
catalyst
Forward Prediction
"Using the provided reactants and reagents, can you propose a likely product?"
Retrosynthesis Prediction
"Give me reactants used in the synthesis of the product."
Type Classification
"Show me the reaction type of the given reaction."
Condition Prediction / Selection
"Could give me a potential reagent / solvent / catalyst for the given chemical reaction?"
"Select the optimal catalyst from the following candidates."
Yield Prediction
"Predict the yield of the reaction using your experienced chemical yield prediction knowledge."
PRESTO-Stage1: Mol-Text Alignment
It is an aluminum coordination entity. It has a role as a Lewis acid.
projector
Language Model
Stage2:Domain Incremental Pretrain
Prepare 0.37 mol of reactants ...
DESC
projector
projector
Language Model

- **Stage1: Molecule-Text Alignment**
  - Cultivates cross-modal alignment ability
- **Stage2: Domain Incremental Pretraining**
  - Focuses on multi-graph understanding
  - Injects domain knowledge of synthetic chemistry
    - Interleaved text-molecule understanding
    - Molecule name/format conversion tasks

Interleaved Dataset for Stage-2 Pretraining

- **Over 3 million** detailed synthetic procedures from USPTO-Patent [Lowe, 2017]
  - Use BERN2 [Sung et al., 2022] to extract molecule entities
- **Molecule name conversions:**
  - IUPAC [Favre and Powell, 2014] , chemical formulas [Hill,1900], and SMILES [Weininger, 1988]



"""Dissolve <mol1> (5.00 g, 22.2 mmol) and <mol2> (4.1 g, 23.0 mmol) in carbon tetrachloride (25 mL) and add AIBN radical initiator (300 mg). Stir and maintain under a nitrogen atmosphere at 80-90° C. or optionally irradiate with a sunlamp ... Cool to room temperature and precipitate the <mol3> from the solution by allowing to stand overnight. Filter and wash the succinimide (2.25 g) with <mol4> (20 mL) ..."""

Dataset for PRESTO Downstream Tasks

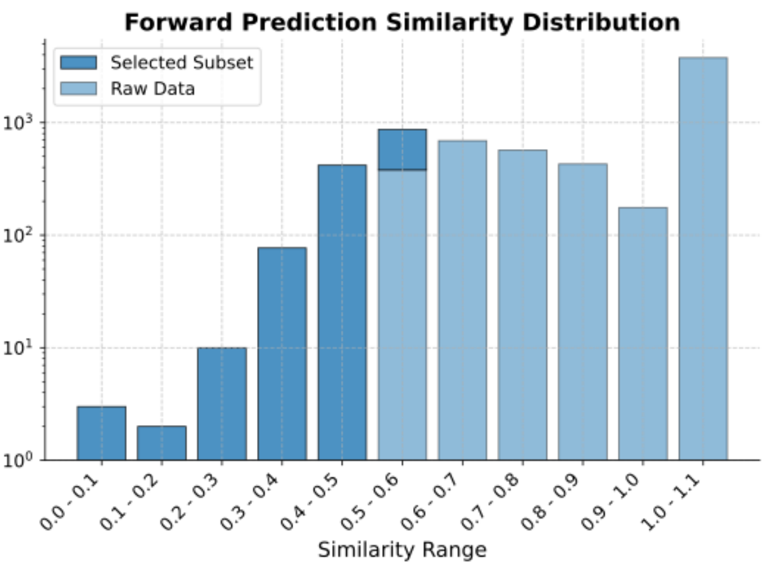
Task	# Train	# Valid	# Test	# All
Reaction Prediction				
Data Source: T5Chem, LLaSMol, Mol-Instruction				
Forward Prediction	124,384	-	1,000	125,384
Retrosynthesis Prediction	124,384	-	1,000	125,384
Reaction Condition Prediction				
Data Source: TextReact, ChemLLMBench, Mol-Instruction				
Reagent Prediction	57,162	6,216	6,378	69,756
Catalyst Prediction	10,232	1,059	1,015	12,306
Solvent Prediction	70,988	7,694	7,793	86,475
Reaction Condition Recommendation				
Data Source: ChemLLMBench [Guo et al., 2023]				
Reagent Selection	3,955	-	300	4,255
Reaction Type Classification				
Data Source: RXNFP [Schwaller et al., 2021]				
Reaction Type Classification	360,379	40,059	44,511	445,115
Yield Prediction				
Data Source: YieldBERT [Schwaller et al., 2021]				
Buchwald-Hartwig	3,855	-	100	3,955
Suzuki-Miyaura	5,660	-	100	5,760

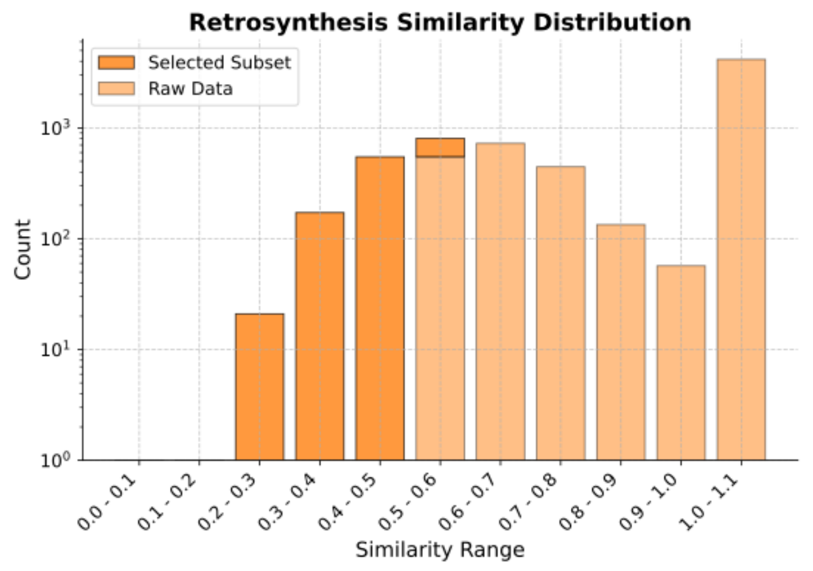
RESTO supports various **synthetic chemistry tasks**:

- Reaction prediction
  - Forward reaction prediction
  - Retrosynthesis prediction
- Reaction condition prediction
  - Reagent
  - Catalyst
  - Solvent
- Reaction condition recommendation
- Reaction type classification
- Yield prediction

Addressing Data Leakage in Benchmarks

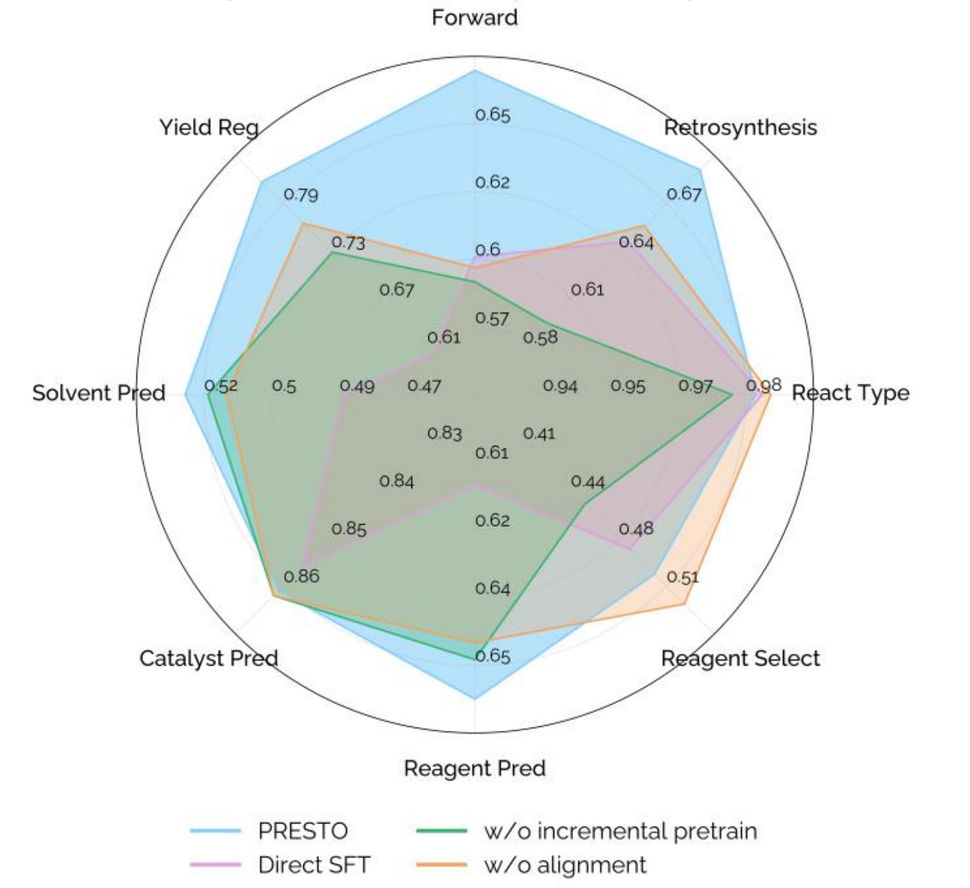
- **Issues:** Data leakage in Mol-Instruction [Fang et al., 2023] benchmark
  - Overlapping (i.e., high scaffold similarities, avg ≈ 0.8) between train and test sets)
  - **Consequences**: Overestimated performance, limited generalizability
- **Our solution:**
  - Non-overlapping, scaffold-based splits ⇝ more challenging test set
  - Unique test set reactions, molecular scaffold splitting (similarity threshold: 0.5-0.6)



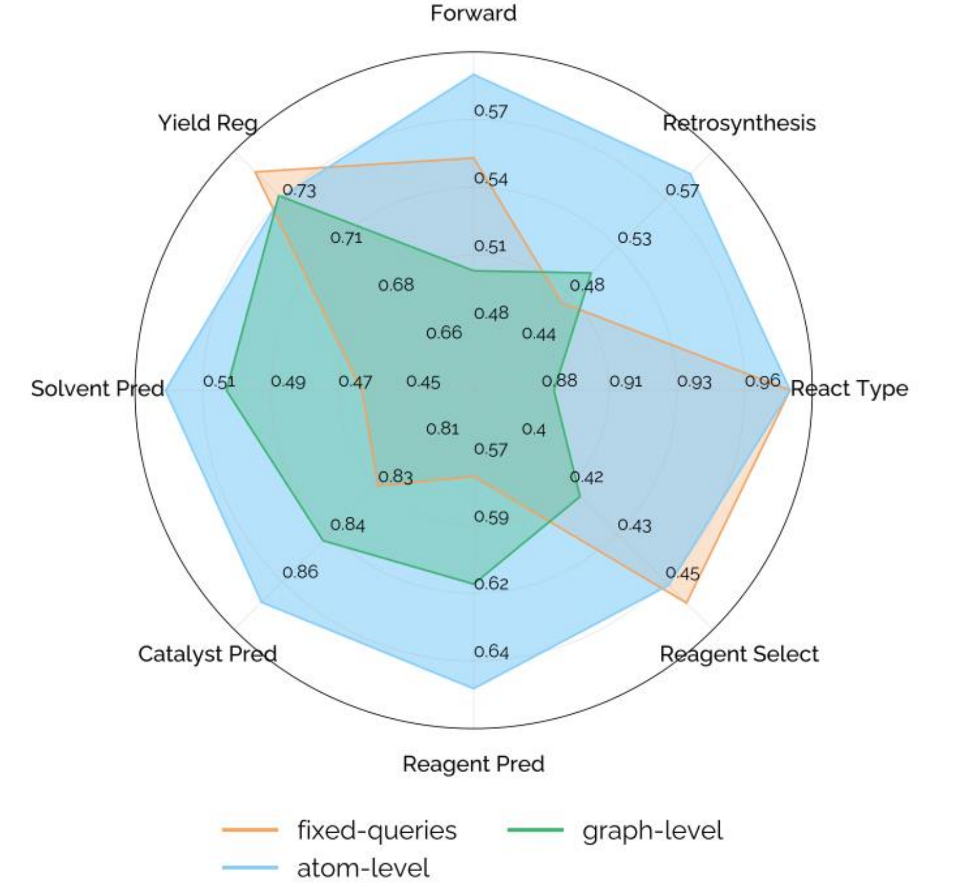


Experiments: Key Findings

* Multi-stage Pretraining Strategy Ablations



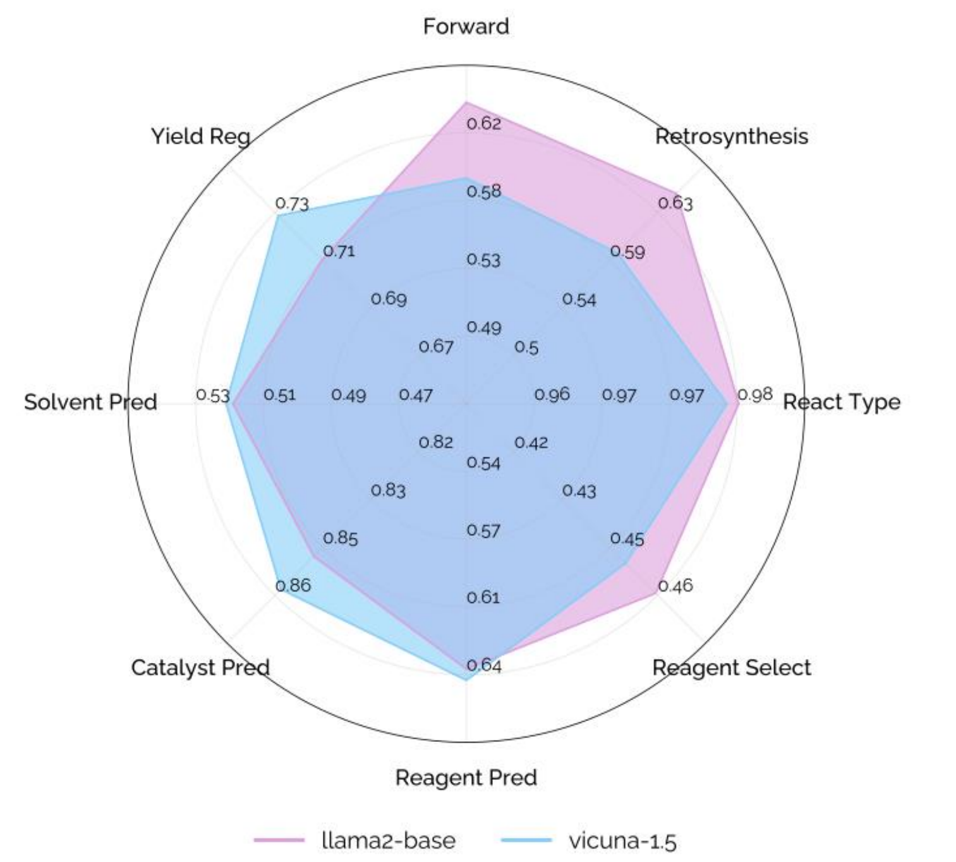
* Molecular Token Granularity Ablations



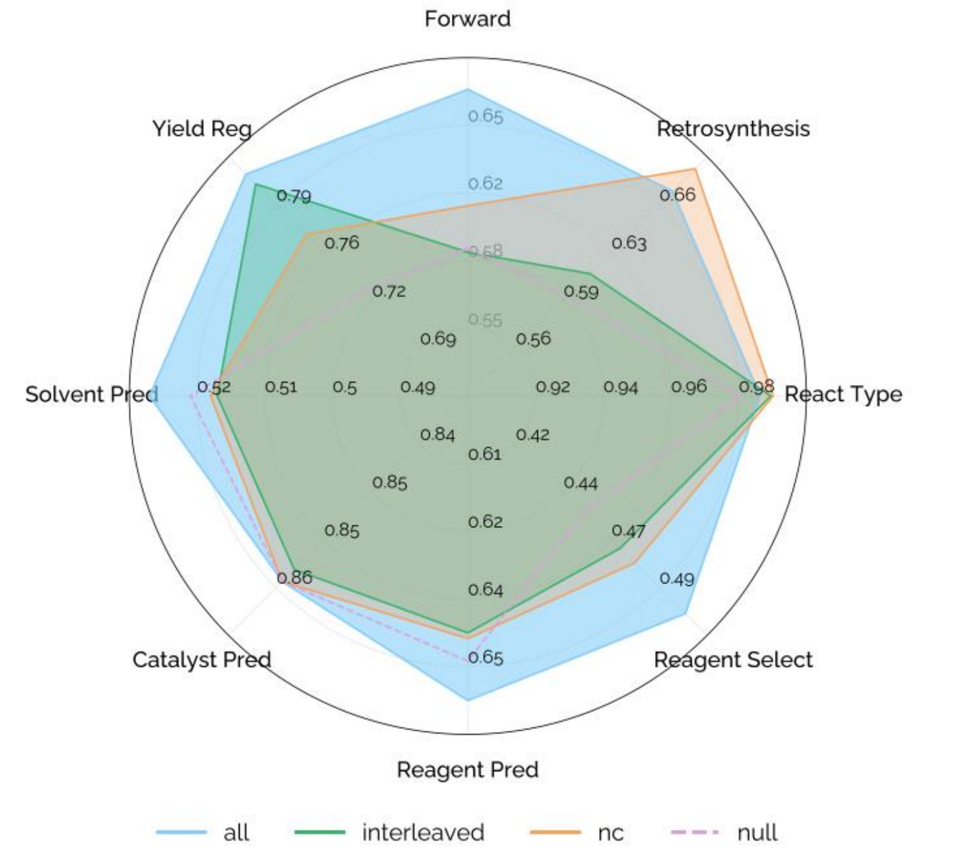
- ➢ Progressive pretraining strategy enhances performance
  - ✓ Align modalities → Domain incremental pretrain → Downstream SFT
- ➢ Molecular representation granularity matters

Experiments: Key Findings

* Base v.s. Instruction-Tuned LLMs



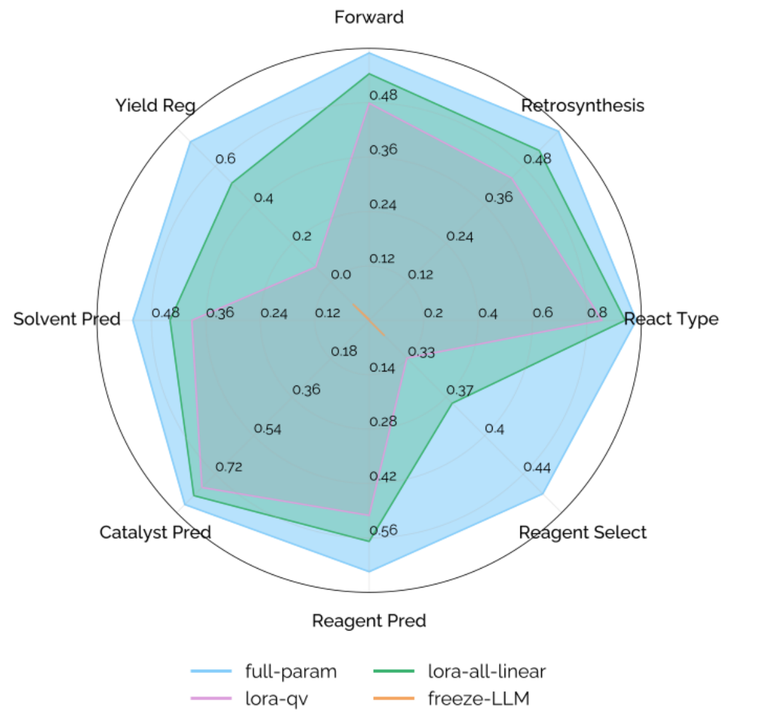
* Pretrain Stage-2 Dataset Configuration



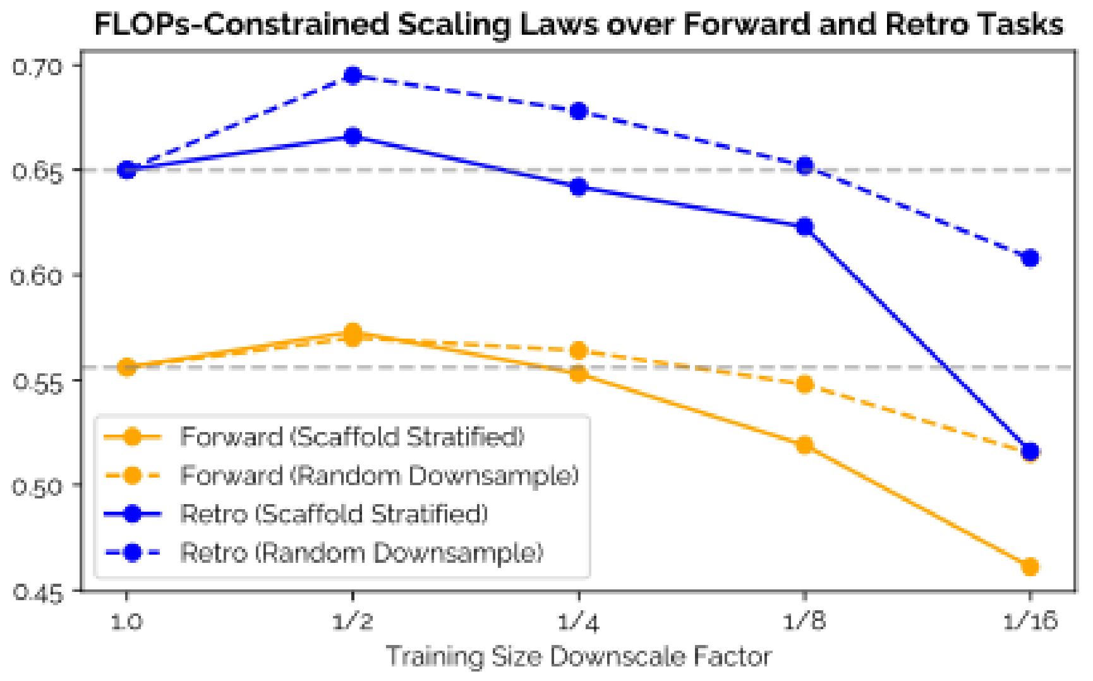
- ➢ Base and instruction-tuned LLMs show similar capabilities
- ➢ Interleaved data and name-conversion data is crucial for domain knowledge injection.

Additional Findings

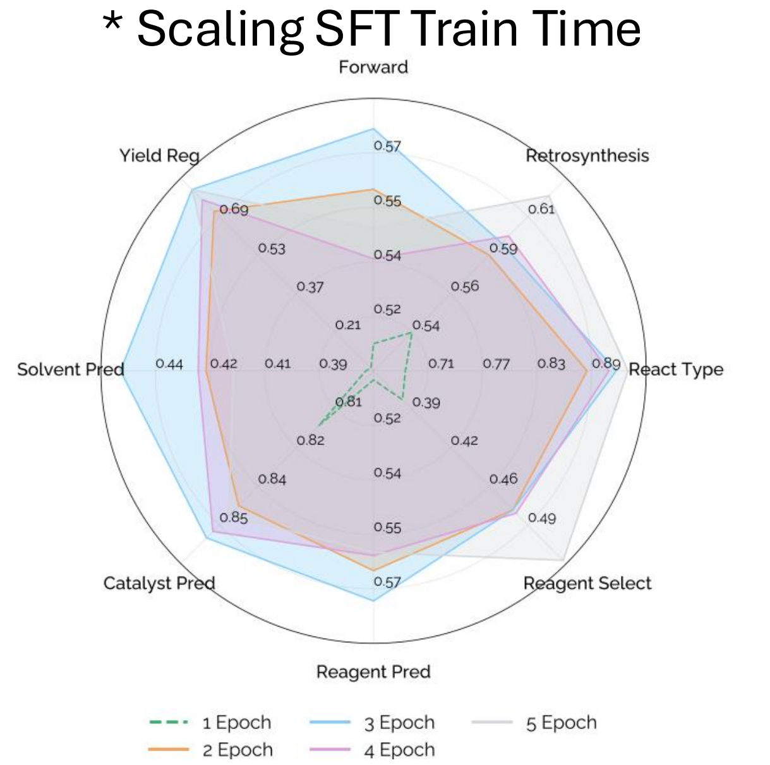
* Ablation on # Trainable Params



* SFT training dataset coverage and diversity



* Scaling SFT Train Time



- ➢ Updating LLMs is essential
  - full-finetune > PEFT to get better results on domain tasks
- ➢ Balancing SFT training time optimizes downstream task performance
  - 3-epoch is OK
- ➢ Coverage and diversity of the SFT dataset are critical for better results

Conclusions and Future Directions

- ➢ PRESTO: A versatile framework for synthetic chemistry. Bridges modality gap and enhances LLM multi-graphs understanding. Potential to advance synthetic chemistry and drug discovery.
- ➢ Future Work:
  - ➢ Expand to including 3D molecular representations.
  - ➢ Enhance dialogue capabilities. Develop larger domain-specific LLMs.