

# From Token to Discovery: A New Paradigm in Material Science

Tong Xie, Bram Hoex

Master AI(UNSW AI4Science)  
Green Dynamics



Intelligent **MATERIALS**, Invented **TODAY**

Homepage: <https://unswai4s.notion.site/>



UNSW  
SYDNEY

 **GreenDynamics**



# Our Mission

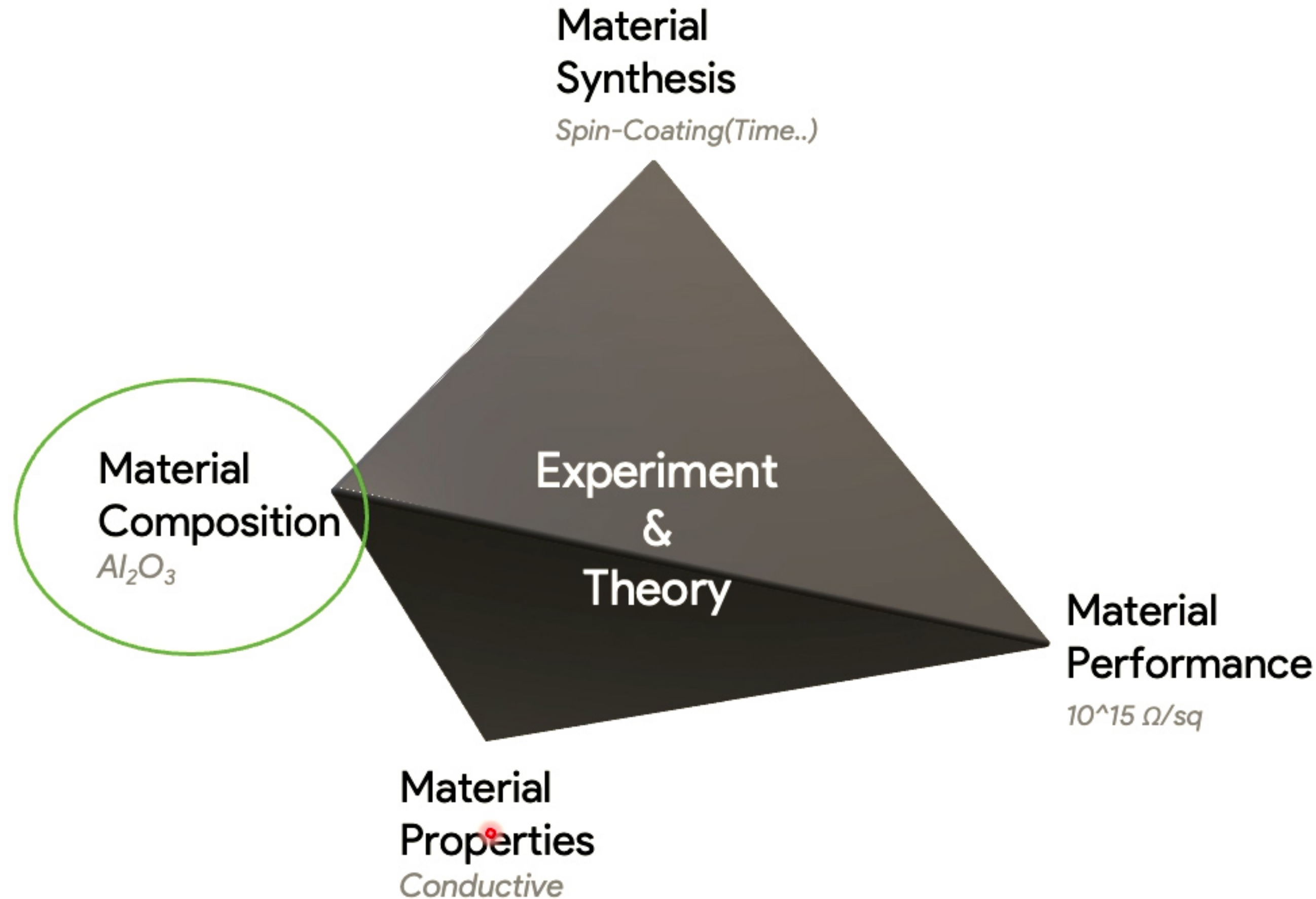
Shape the material  
future through “AI  
Scientists” dedicated to  
solving the grand  
challenges of our time.





# From Decades to Months: Supercharging Material Innovation

Design a Single Material we need to consider:



SCIENCE

Millions of new materials discovered  
with deep learning

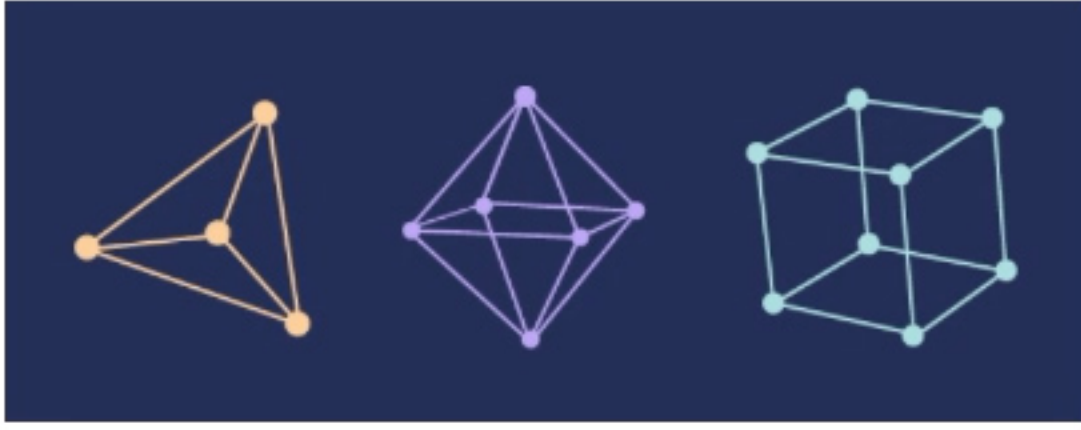
**2.2 M** new  
**380K** Stable  
**0** Synthesizable

**Thermodynamics Stability  
≠ Synthesizability**

*(kinetic barriers, metastability, or synthesis route...)*



**Curse of dimensionality** in Macro Scale



Humans can learn from papers and theories to discover new materials.

**Why not machines?**

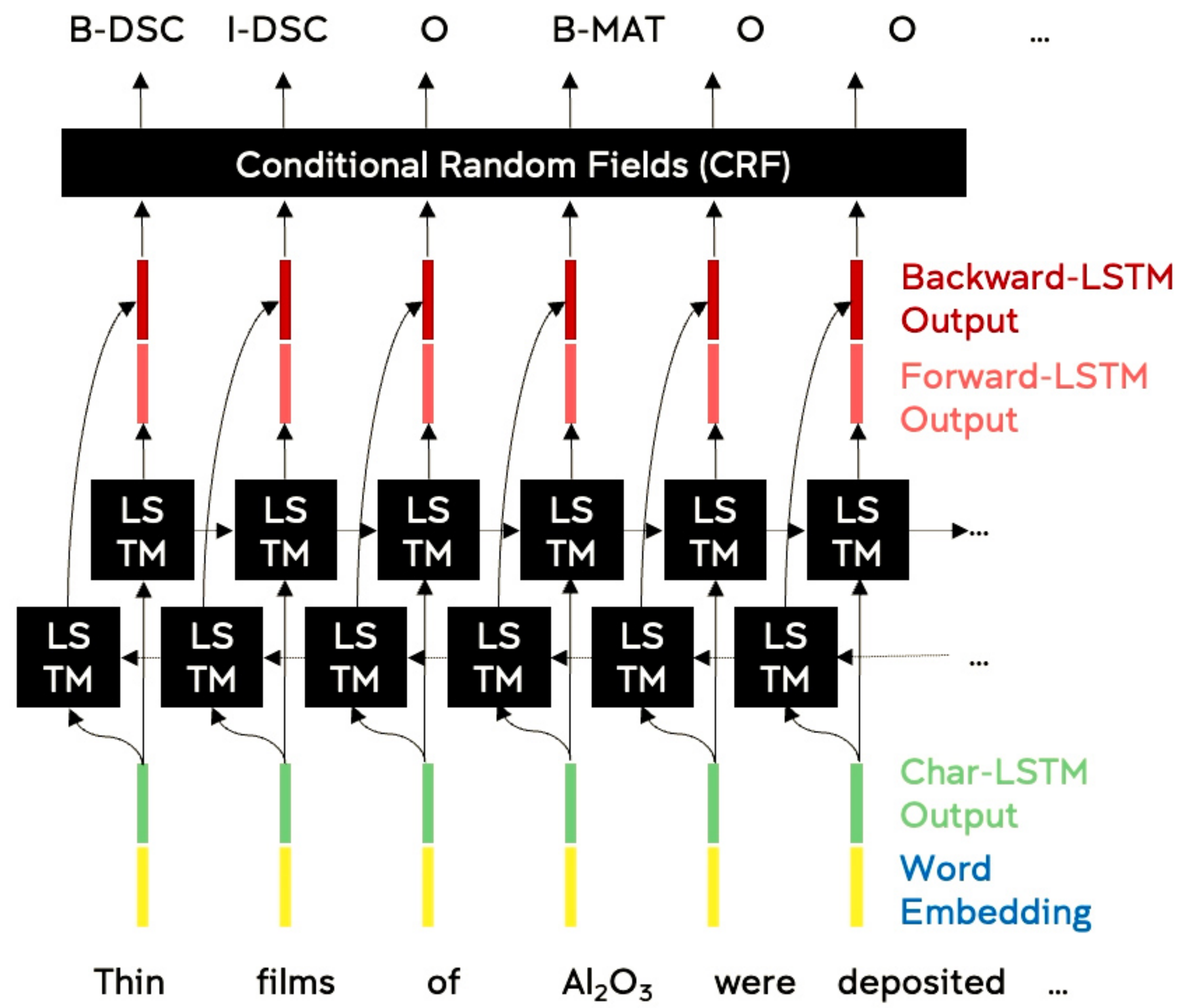
With the right **tokenization** and **architecture**, we can make **material** happen.





# Non-LLM for Material Information

## Model Architecture



## Result & Evaluation

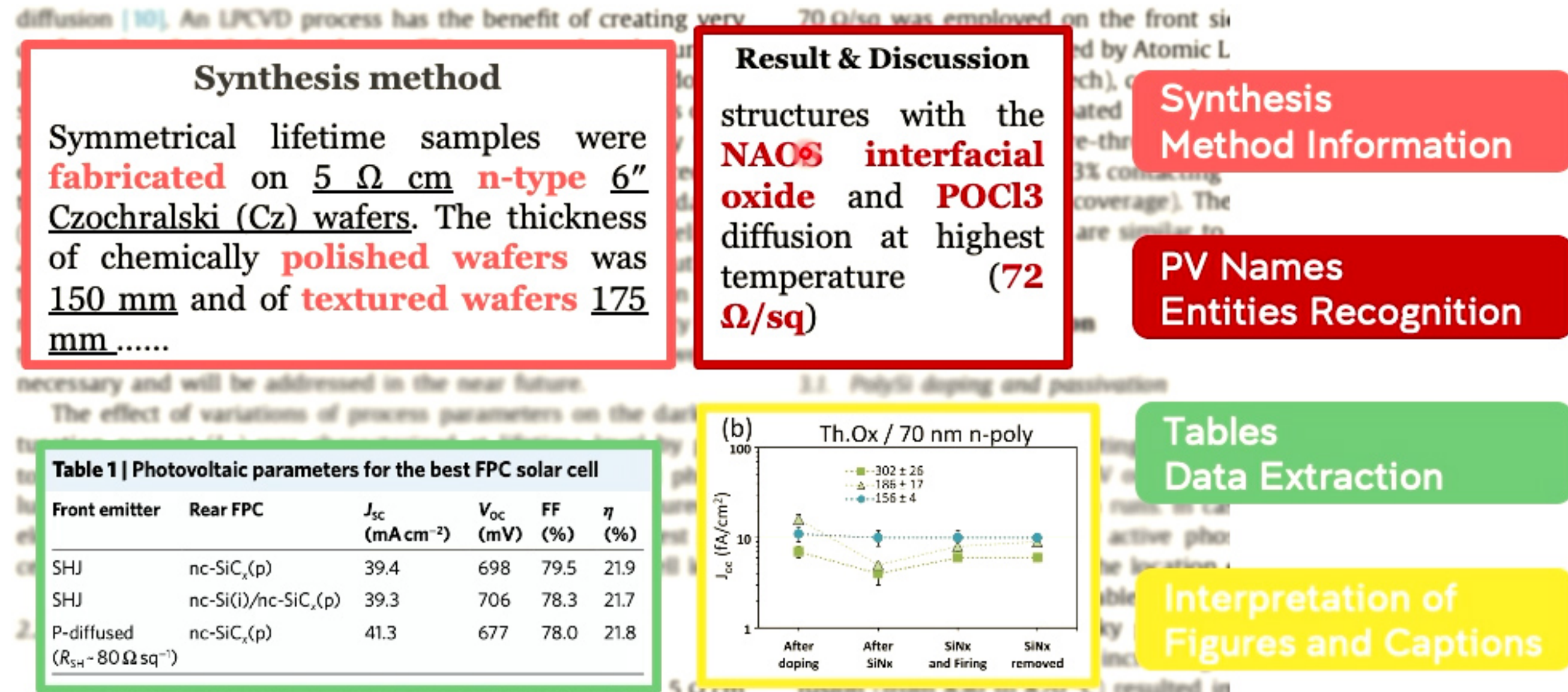


Fig 2. Example of the information could be extracted from a PV paper

| NER | Precision | Recall | F1-score | NER | Precision | Recall | F1-score |
|-----|-----------|--------|----------|-----|-----------|--------|----------|
| MAT | 0.789     | 0.854  | 0.821    | SPL | 0.750     | 0.731  | 0.740    |
| PRO | 0.691     | 0.753  | 0.720    | DSC | 0.806     | 0.825  | 0.815    |
| SMT | 0.621     | 0.639  | 0.630    | APL | 0.755     | 0.689  | 0.721    |
| CMT | 0.736     | 0.771  | 0.740    | PVM | -         | -      | -        |

Fig 1. Neural network architecture for named entity recognition (NER)

Table 1. Detailed results for Name entity Recognition on the test sets\*

\* MAT: Material, PRO: Property, SMT: Synthesis Method, CMT: Characterization Method, SPL: Symmetry Phase Label, DSC: Descriptor, APL: application, PVM: PV Manufacture





# LLM for Material Information

## Previously

It needs multiple sequential steps, including NER, RE, normalisation, and, ultimately, structuring the data.

## Now

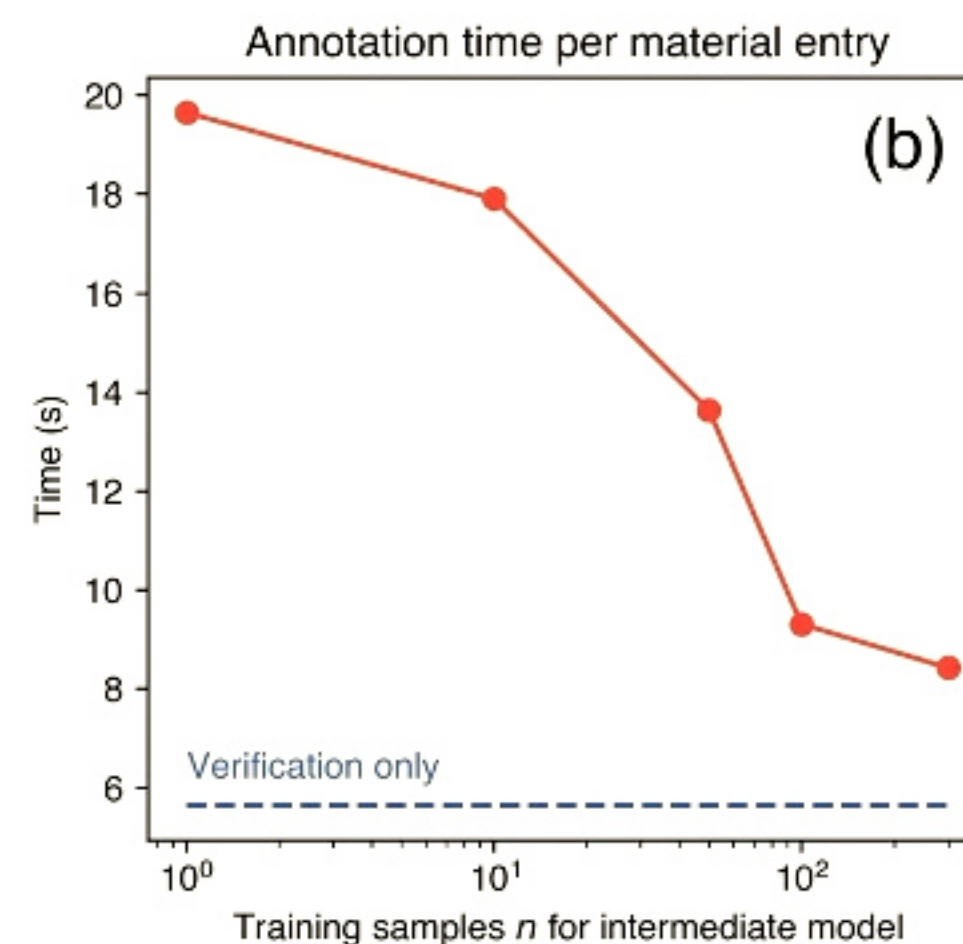
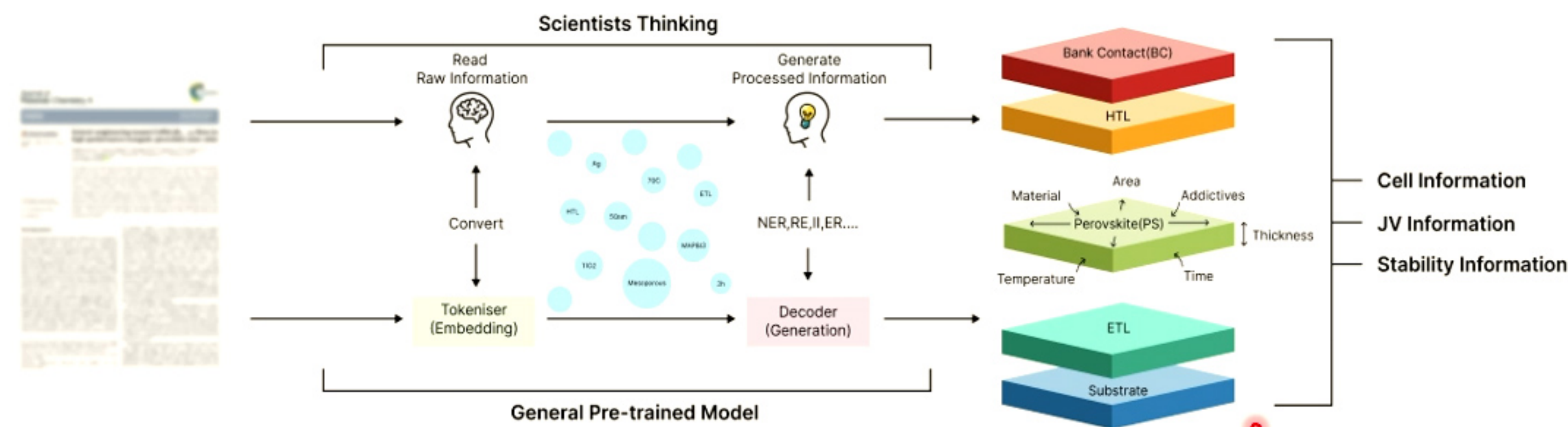
We propose a new NLP task called SII (structured information inference) to perform all sequential steps in one step jointly:

### Input:

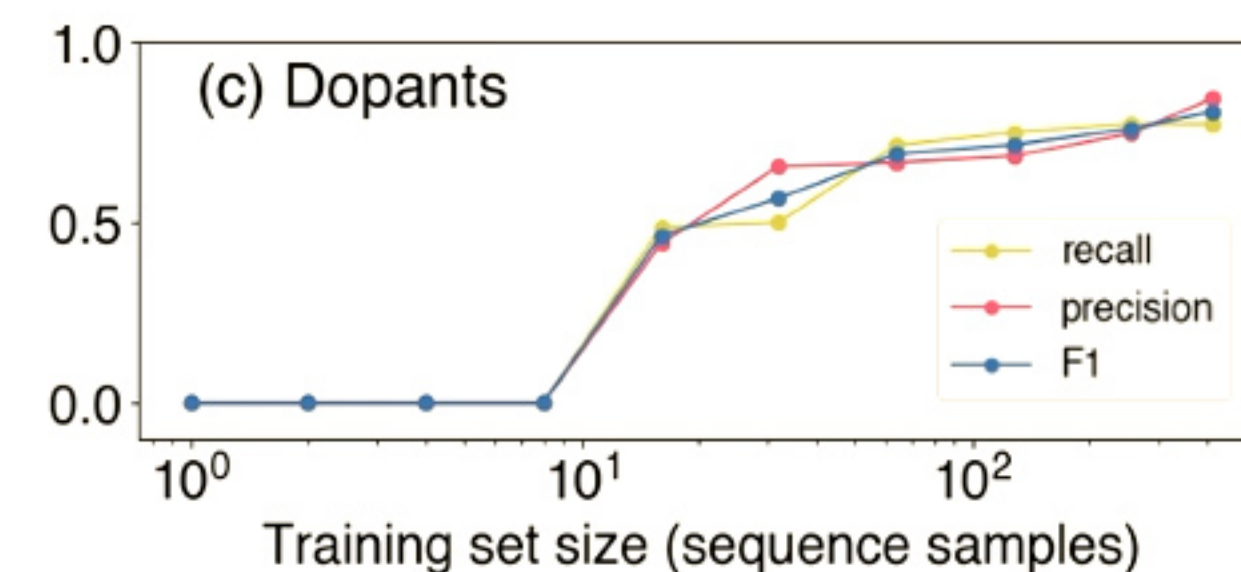
A paragraph / full-text

### Output:

structured data in JSON format (including normalised entity, numeric data, relationship, etc.)



Annotation time continuously reduce **50%**



**~20** minimum number of samples for Supervised Finetuning



[1] Dagdelen, J. et al. . Structured information extraction from scientific text with large language models. Nature Communications, 15(1), 1418.  
[2] Xie, T et al. Creation of a structured solar cell material dataset and performance prediction using large language models. Patterns, 5(5).



# Non-LLM for Science QA

## Extractive question answering

- Takes a question and a document as the input. The document contains a span that can serve as the correct answer to the question.
- The **start** and **end** tokens of the answers with the highest probability will be extracted.

## How

- The training stage will add a span classification layer on the **LSTM/BERT/...** model.
- Predict the **start** and **end** location of answer phrases (for example, multiplication of the token level probabilities).

**Inputs:** A question  $q = \{q_1, \dots, q_Q\}$  of length  $Q$  and a context paragraph  $p = \{p_1, \dots, p_P\}$  of length  $P$

**Output:** An answer span  $\{a_s, a_e\}$  where  $a_s$  is the index of the first answer token in  $p$ ,  $a_e$  is the index of the last answer token in  $p$ ,  $0 \leq a_s, a_e \leq m$ , and  $a_s \leq a_e$

**Context:** The blended slurry was then cast onto a clean current collector (Al foil for the cathode and Cu foil for the anode)

**Question:** What is the anode?

**Answer:** Cu foil

**Context:** Li and co-authors applied PTCDA-derived polymer P10 as the anode for aqueous K-ion batteries. Saturated solutions of K<sub>2</sub>SO<sub>4</sub> (~0.69 M) or KNO<sub>3</sub> (~3.75 M) were tested as the electrolytes.

**Question:** What is the electrolyte?

**Answer:** K<sub>2</sub>SO<sub>4</sub> (~0.69 M) or KNO<sub>3</sub> (~3.75 M)

**Context:** Among the state-of-the-art cathode materials, Ni-rich layered oxides, LiNi<sub>x</sub>CoyAlzO<sub>2</sub> (NCA) and LiNi<sub>x</sub>CoyMnzO<sub>2</sub> (NCM) with  $x + y + z = 1$  and  $x \geq 0.8$ , have an unbeatable high capacity of ~200 mAhg<sup>-1</sup>.

**Question:** What is the cathode?

**Answer:** Ni-rich layered oxides, LiNi<sub>x</sub>CoyAlzO<sub>2</sub> (NCA) and LiNi<sub>x</sub>CoyMnzO<sub>2</sub> (NCM)



[1] Huang, S., & Cole, J. M. (2022). BatteryBERT: A pretrained language model for battery database enhancement. Journal of chemical information and modeling, 62(24), 6365-6377.

[2] Haigh, A., van de Graaf, C., & Rachleff, J. Extractive Question Answering Using Match-LSTM and Answer Pointer.



# LLM for Science QA: SciQAG

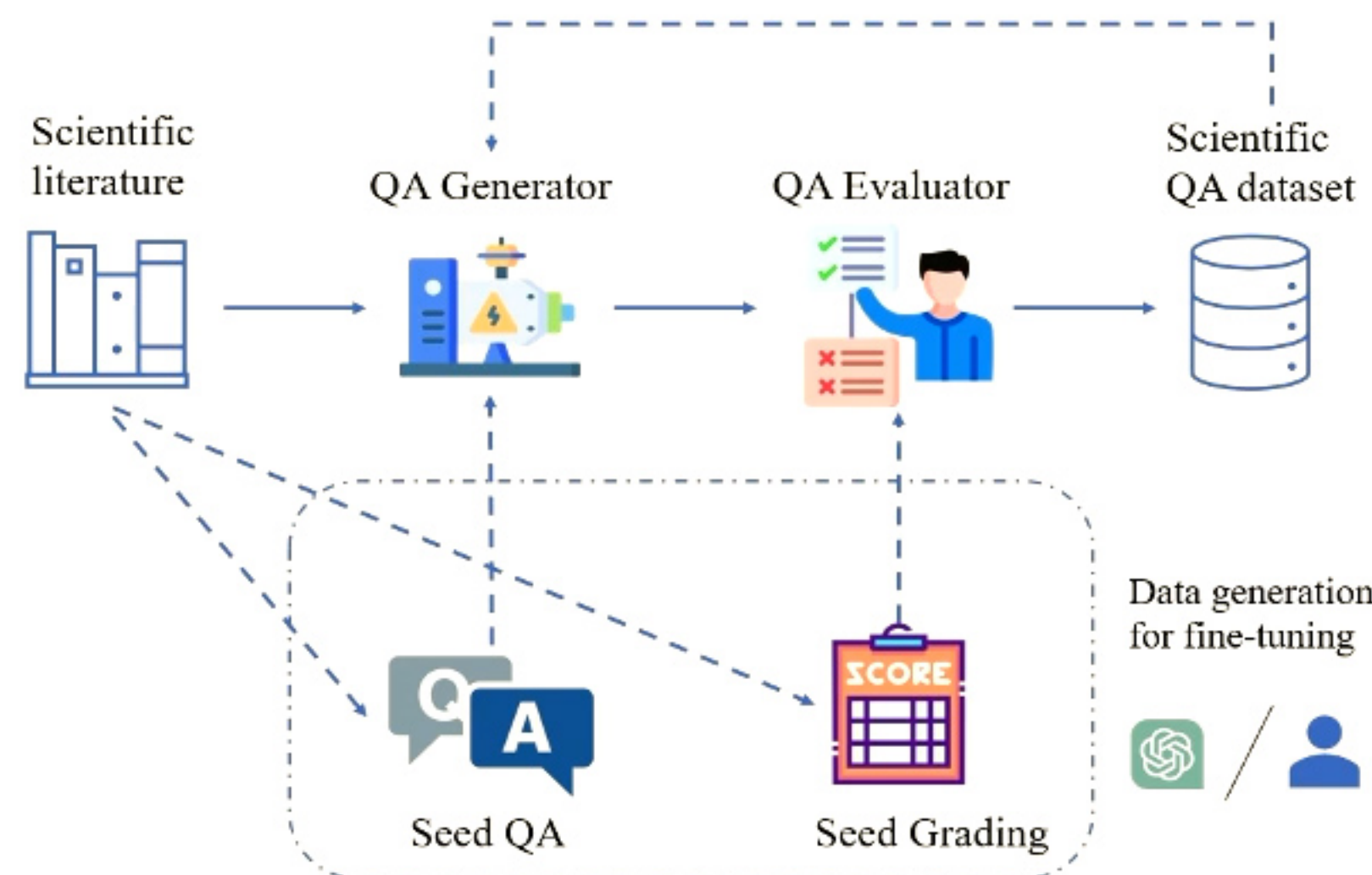
## Previously

- Only **span in the document** can be extracted as answer --> simple QA
- May need an extra filtering process

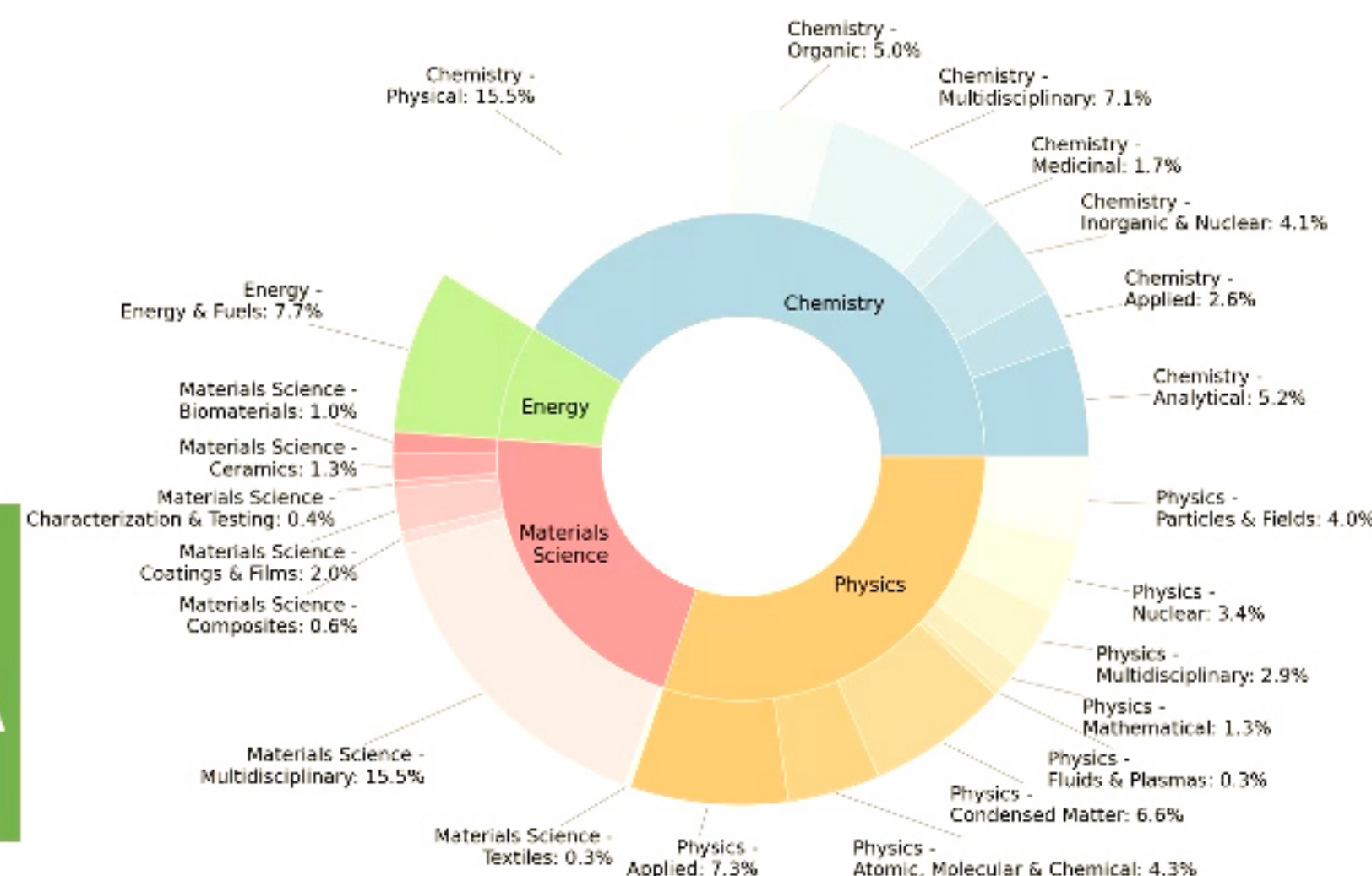
## Now

Cover more kinds of QA tasks (also more **complex** and **longer context**):

- Translation
- Summary
- Multiple-choice
- Information extraction
- Product prediction
- Open-ended QA
- Multi-modal QA
- ...



**LLaMA 2-7B with Flash Attention 2**  
Generate **~188,042 QA**  
from **~22,743 papers**



**Yuwei Wan**  
GD/CityU



**Marc Wilkins**  
UNSW



**Ian Foster**  
Argonne/Uchicago





# SciQAG: LLM for Science QA

## Previously

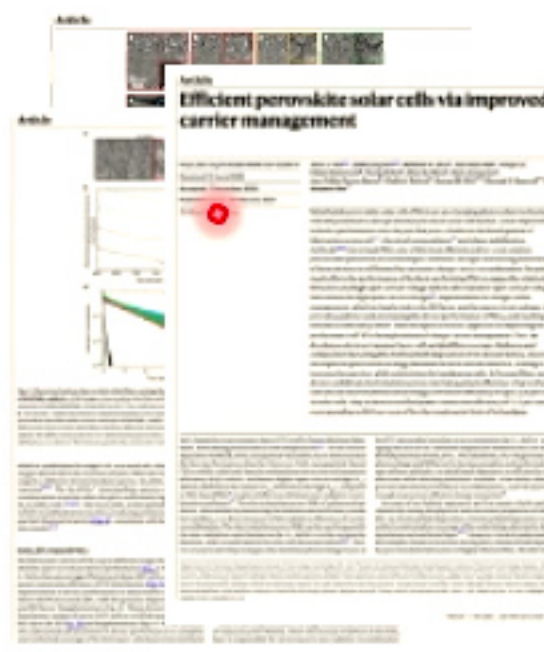
- Only **span in the document** can be extracted as answer --> simple QA
- May need an extra filtering process

## Now

Cover more kinds of QA tasks (also more **complex** and **longer context**):

- Translation
- Summary
- Multiple-choice
- Information extraction
- Product prediction
- Open-ended QA
- Multi-modal QA
- ...

## Content



**Question:** Describe how chemical bath deposition affects the morphology and performance of SnO<sub>2</sub> layers in solar cells.

**Answer:** Chemical bath deposition allows for uniform and complete coverage, which is essential for reducing pinholes and defects that can lead to non-radiative recombination, thereby improving the efficiency and stability of the solar cells.

## Other Examples with LLM

- ScienceQA (Multi-modal, CoT)
- SciQA (knowledge graph vs. LLM)
- ChemBench (chemistry-specific dialogue)
- QASA (rationale-generation)
- MMScl(Multi-modal, Open-ended QA)...





# Our Domain Application Examples

APPLIED  
NANO MATERIALS

www.acsnano.org

## An Artificial-Intelligence-Assisted Investigation on the Potential of Black Silicon Nanotextures for Silicon Solar Cells

Shaozhou Wang,<sup>1</sup> Tong Xie,<sup>2</sup> Ran Liang, Yu Zhang, Fa-Jun Ma, David Payne, Giuseppe Scardera, and Bram Hoex<sup>3</sup>

Cite This: <https://doi.org/10.1021/acsnano.2c02619>

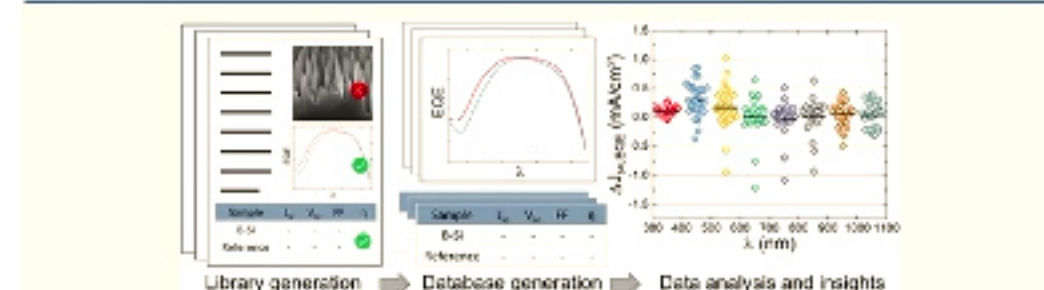
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information



**ABSTRACT:** Black silicon (b-Si) nanotextures are of interest for Si solar cells because of their enhanced light trapping properties. However, the wide range of complex nanotextured b-Si surface morphologies makes a systematic investigation of b-Si solar cells challenging. A comprehensive performance review is necessary to determine the promising b-Si nanotextures for solar cell applications. In this work, we use artificial-intelligence approaches to assist in compiling a systematic and highly refined performance review of b-Si solar cells. We also perform numerical simulations of electrical properties for various nanotextured b-Si morphologies. We find that the weighted average reflectance (WAR) is an effective surface morphology metric for a wide range of surface textures. By correlating solar cell performance parameters to WAR, we show that multicrystalline Si solar cell efficiency can be improved with b-Si nanotexturing, and this is predominately attributed to an increase in short-circuit current density via the blue response improvement. We also show that some b-Si nanotextures can improve the performance of monocrystalline Si solar cells. Device simulations show that the electrical performance of hierarchical (combination of microtexture and nanotexture) and inverted-pyramidal b-Si nanotextures and microtextures can be comparable to or even better than random pyramids. As such, these textures show great potential for monocrystalline Si solar cells.

**KEYWORDS:** nanotexture, black silicon, silicon solar cell, natural language processing, computer vision, machine learning, numerical simulation

Black Silicon  
2022



Nano Thinfilm Synthesis  
2023

scientific data

OPEN  
DATA DESCRIPTOR

## Automatically Generated Datasets: Present and Potential Self-Cleaning Coating Materials

Shaozhou Wang,<sup>1,2,3</sup> Yuwei Wan,<sup>1,2,4</sup> Ning Song,<sup>1,2,5</sup> Yixuan Liu,<sup>2</sup> Tong Xie,<sup>1,2,6</sup> & Bram Hoex<sup>1,2,6</sup>

The rise of urbanization coupled with pollution has highlighted the importance of outdoor self-cleaning coatings. These revolutionary coatings contribute to the longevity of various surfaces and reduce maintenance costs for a wide range of applications. Despite ongoing research to develop efficient and durable self-cleaning coatings, adopting systematic research methodologies could accelerate these advancements. In this work, we use Natural Language Processing (NLP) strategies to generate open- and traceable-sourced datasets about self-cleaning coating materials from 39,031 multi-disciplinary papers. The data are from function-based and property-based corpora for self-cleaning purposes. These datasets are presented in four different formats for diverse uses or combined uses: material frequency statistics, material dictionary, measurement value datasets for self-cleaning-related properties and optical properties, and sentiment statistics of material stability and durability. This provides a literature-based data resource for the development of self-cleaning coatings and also offers potential pathways for material discovery and prediction by machine learning.

Self-Cleaning Coating  
2023

Patterns

CellPress  
OPEN ACCESS

Article

## Creation of a structured solar cell material dataset and performance prediction using large language models

Tong Xie,<sup>1,2</sup> Yuwei Wan,<sup>1,2</sup> Yufei Zhou,<sup>2</sup> Wei Huang,<sup>2</sup> Yixuan Liu,<sup>2</sup> Qingyuan Linghu,<sup>1,2</sup> Shaozhou Wang,<sup>1,2</sup> Chunyu Kit,<sup>1,2</sup> Clara Grazian,<sup>1,2</sup> Wenjie Zhang,<sup>2</sup> and Bram Hoex<sup>1,2,6</sup>

<sup>1</sup>School of Photovoltaic and Renewable Energy Engineering, University of New South Wales, Kensington, NSW, Australia

<sup>2</sup>GreenDynamics Pty. Ltd, Kensington, NSW, Australia

<sup>3</sup>Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China

<sup>4</sup>School of Mathematics and Statistics, University of Sydney, Camperdown, NSW, Australia

<sup>5</sup>DAARE ARC Training Centre in Data Analytics for Resources and Environments, South Eveleigh, NSW, Australia

<sup>6</sup>School of Computer Science and Engineering, University of New South Wales, Kensington, NSW, Australia

<sup>\*</sup>Lead contact

<sup>\*</sup>Correspondence: [ctoxt@cityu.edu.hk](mailto:ctoxt@cityu.edu.hk) (C.X.), [b.hoex@unsw.edu.au](mailto:b.hoex@unsw.edu.au) (B.H.)

<https://doi.org/10.1016/j.patter.2024.100955>

**THE BIGGER PICTURE** Big data's importance in materials science is clear, yet its effective use is challenging due to the sheer volume and complexity of the data. Natural language processing (NLP) offers a solution by transforming unstructured text into structured formats, facilitating tasks such as extraction and summarization. In materials science, this means converting information from scientific papers into structured datasets, a process often slowed by the continuous influx of new data. To circumvent the inefficiencies of multi-step NLP workflows, there is a growing need for streamlined, one-step NLP methods. Employing fine-tuned large language models could be key, allowing for the rapid updating of datasets and providing valuable training data for further model development. This approach not only expedites research but also accelerates material prediction, leading to faster scientific breakthroughs.

Solar Cell  
2024





# Perception & Action

Intelligence has two essential parts, **perception** and **action**, which are jointly learned and optimised.

| Perception                                                                          |                              |
|-------------------------------------------------------------------------------------|------------------------------|
| Traditional AI tasks: classification, 2D/3D detection, depth estimation, etc.       |                              |
|    | Literature understanding     |
|  | Characterization analysis    |
|  | Lab environmental perception |
|  | Safety monitor               |

| Action                                                                                |                      |
|---------------------------------------------------------------------------------------|----------------------|
| Traditional AI tasks: path planning, robot navigation, robot manipulation             |                      |
|    | Hypothesis propose   |
|  | Plan optimization    |
|  | Simulation execution |
|  | Experiment execution |





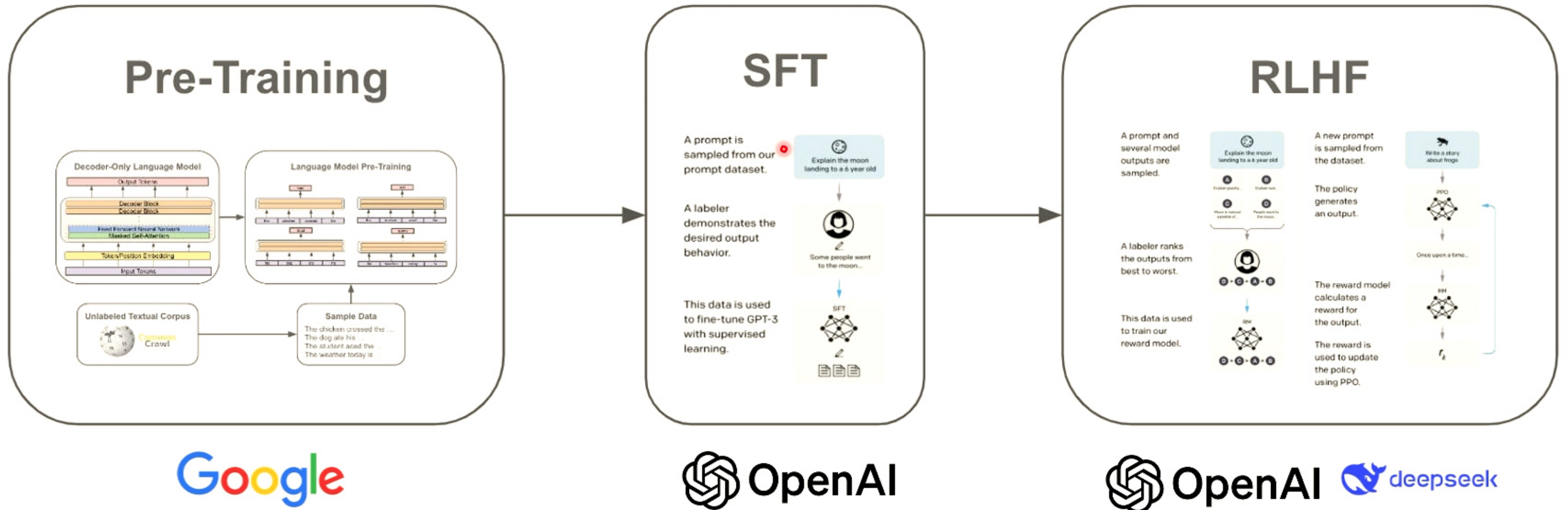
# Unlocking Scientific Action: The Power of Foundational Models (Large Language Models)





# How to train a LLM?

## Alignment







# GPT-3.5 for PV Material Design(2023.03)

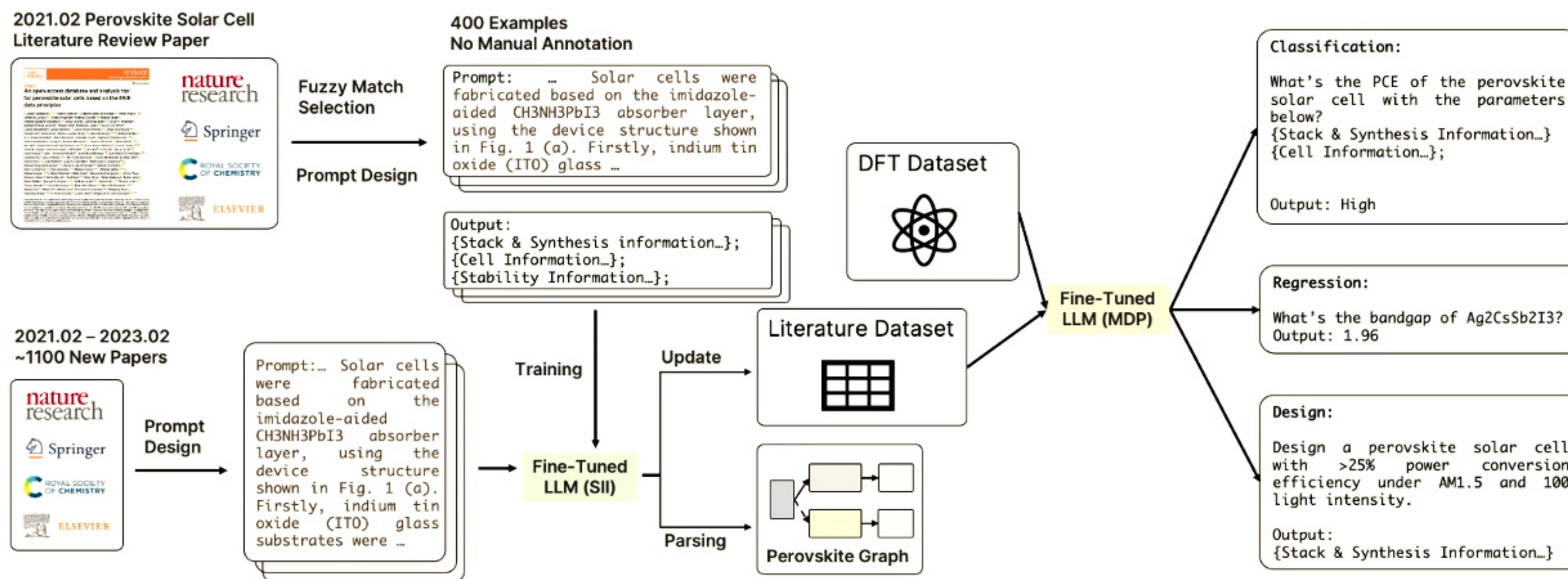
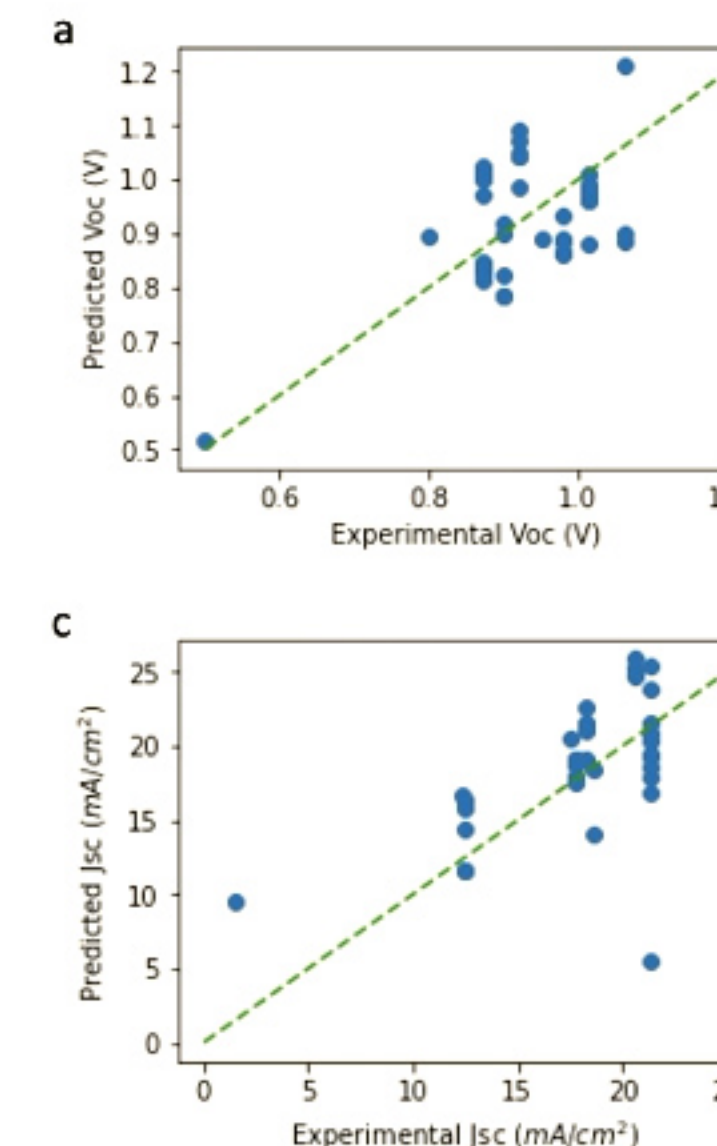


Figure 3: Flowchart of the fine-tuned LLM for structured information inference(SII) tasks and material & device prediction(MDP) tasks



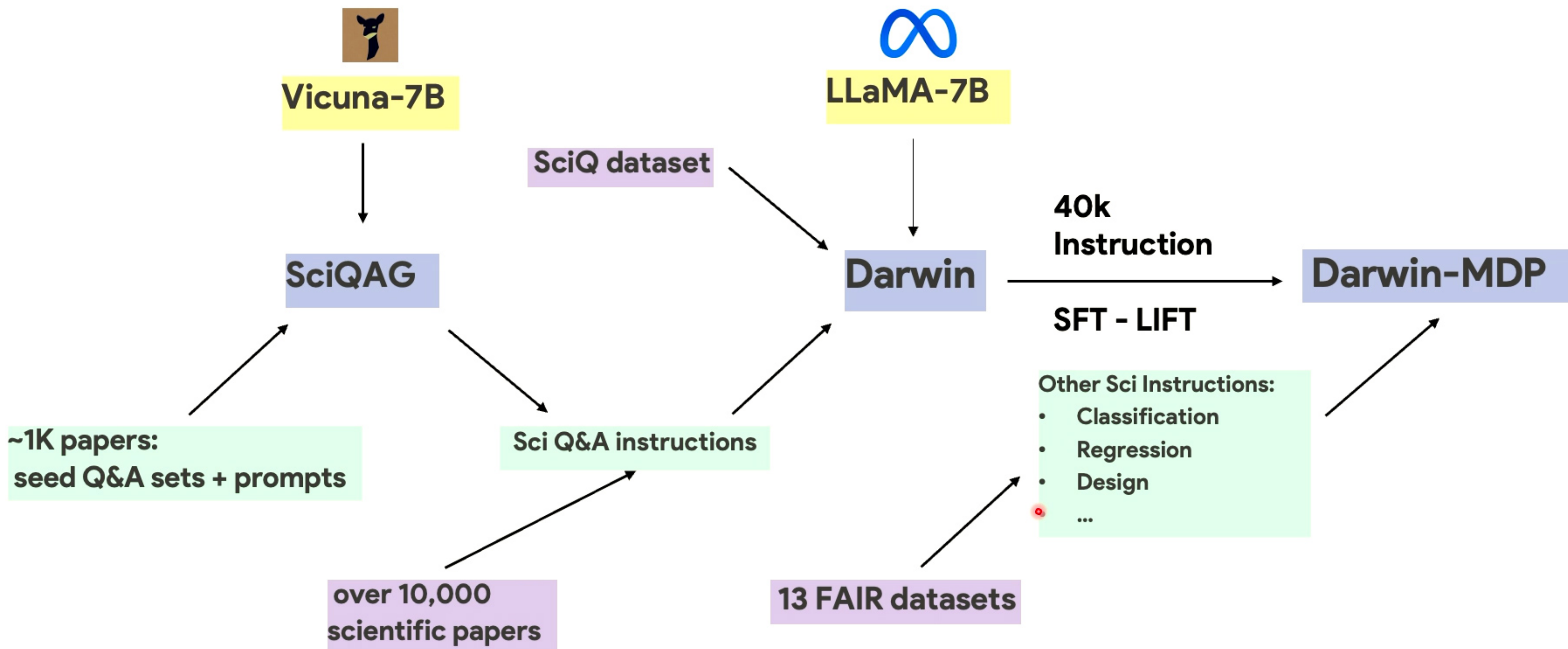
Clara Grazian  
University of Sydney





# DARWIN 7B (2023.08)

World First Material Science LLM





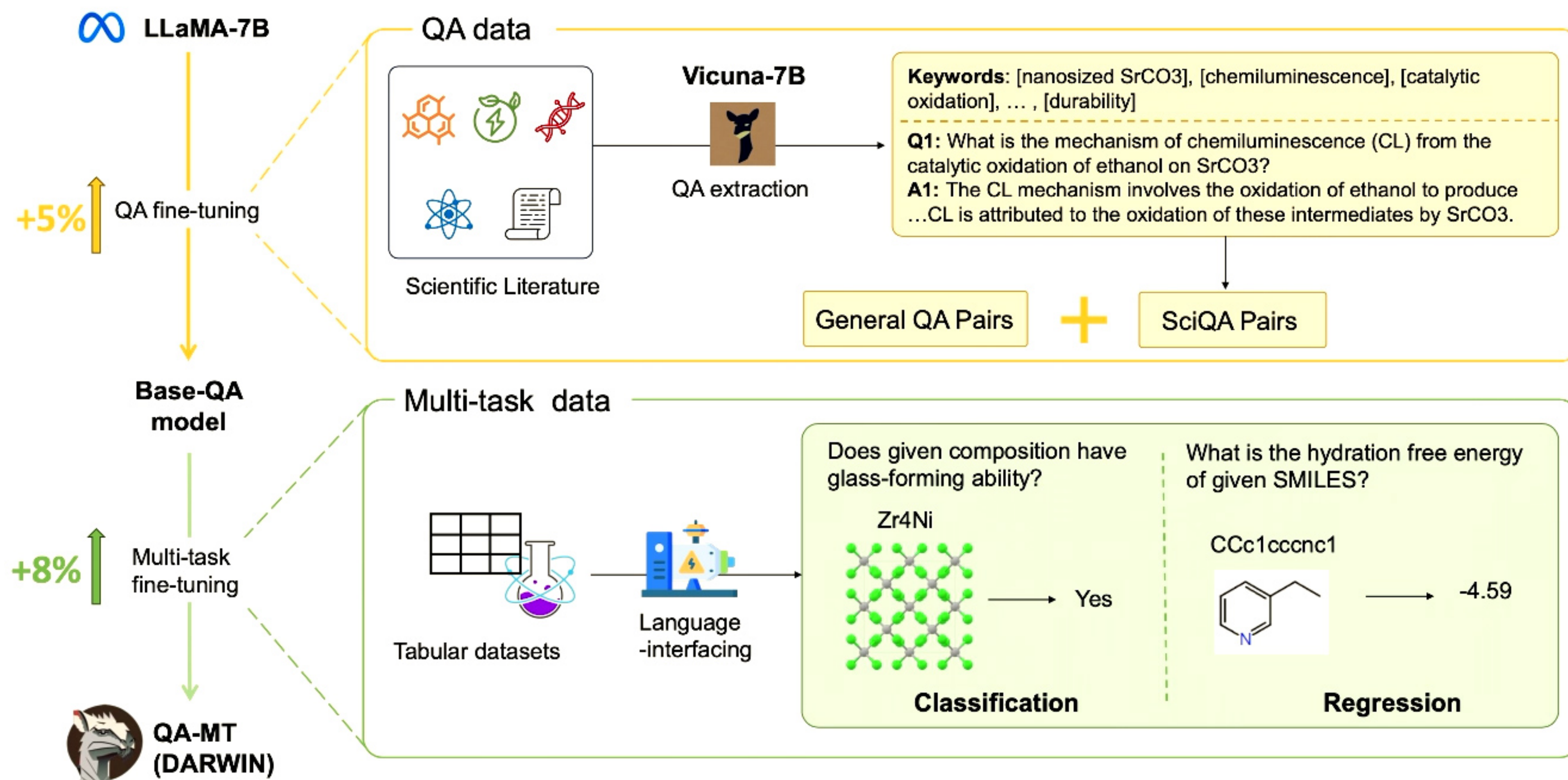


# DARWIN 1.5 7B(2024.06) – focused on MDP

(Material & Device Prediction)



➤ Language Interface FT(LIFT) makes LLM as a universal regressor in material science



**Yuchen Zeng**  
Microsoft



**Yuwei Wan**  
GreenDynamics



# FT-LLM Powerd Material Prediction

## ➤ Performance on material prediction task

**World Record in Experimental Material Design  
Verified by Material Project - MatBench**

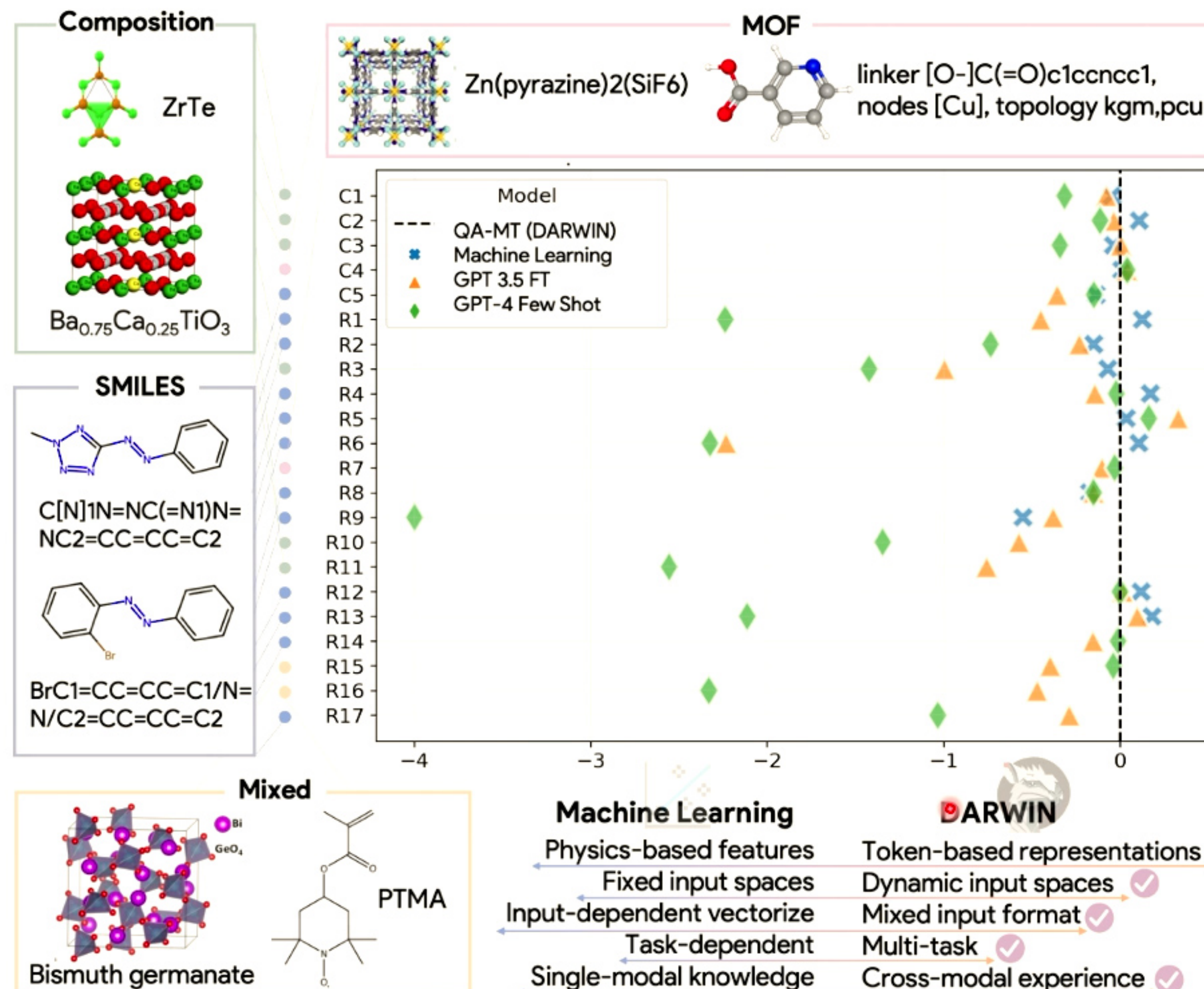
matbench\_v0.1 matbench\_expt\_gap

Individual Task Leaderboard for matbench\_expt\_gap

Leaderboard for an individual task. Algorithms shown here may include both general purpose and specialized algorithms (i.e., algorithms which are only valid for a subset of tasks in the benchmark).

Leaderboard

| algorithm                | mean mae | std mae | mean rmse | max max_error |
|--------------------------|----------|---------|-----------|---------------|
| Darwin                   | 0.2865   | 0.0083  | 0.7185    | 7.6000        |
| Au/SAASBO CrabNet v1.2.7 | 0.3310   | 0.0071  | 0.8123    | 11.1001       |
| MODNet (v0.1.12)         | 0.3327   | 0.0239  | 0.7685    | 9.8955        |
| CrabNet                  | 0.3463   | 0.0088  | 0.8504    | 9.8002        |
| MODNet (v0.1.10)         | 0.3470   | 0.0222  | 0.7437    | 9.8567        |
| Ax+CrabNet v1.2.1        | 0.3566   | 0.0248  | 0.8673    | 11.0998       |
| Ax(10/90)+CrabNet v1.2.7 | 0.3632   | 0.0196  | 0.8679    | 11.1003       |
| CrabNet v1.2.1           | 0.3757   | 0.0207  | 0.8805    | 10.2572       |
| AMMExpress v2020         | 0.4161   | 0.0194  | 0.9918    | 12.7533       |
| RF-SCM Magpie            | 0.4461   | 0.0177  | 0.8243    | 9.5428        |
| gptchem                  | 0.4544   | 0.0123  | 1.0737    | 11.7000       |
| Dummy                    | 1.1435   | 0.0310  | 1.4438    | 10.7354       |





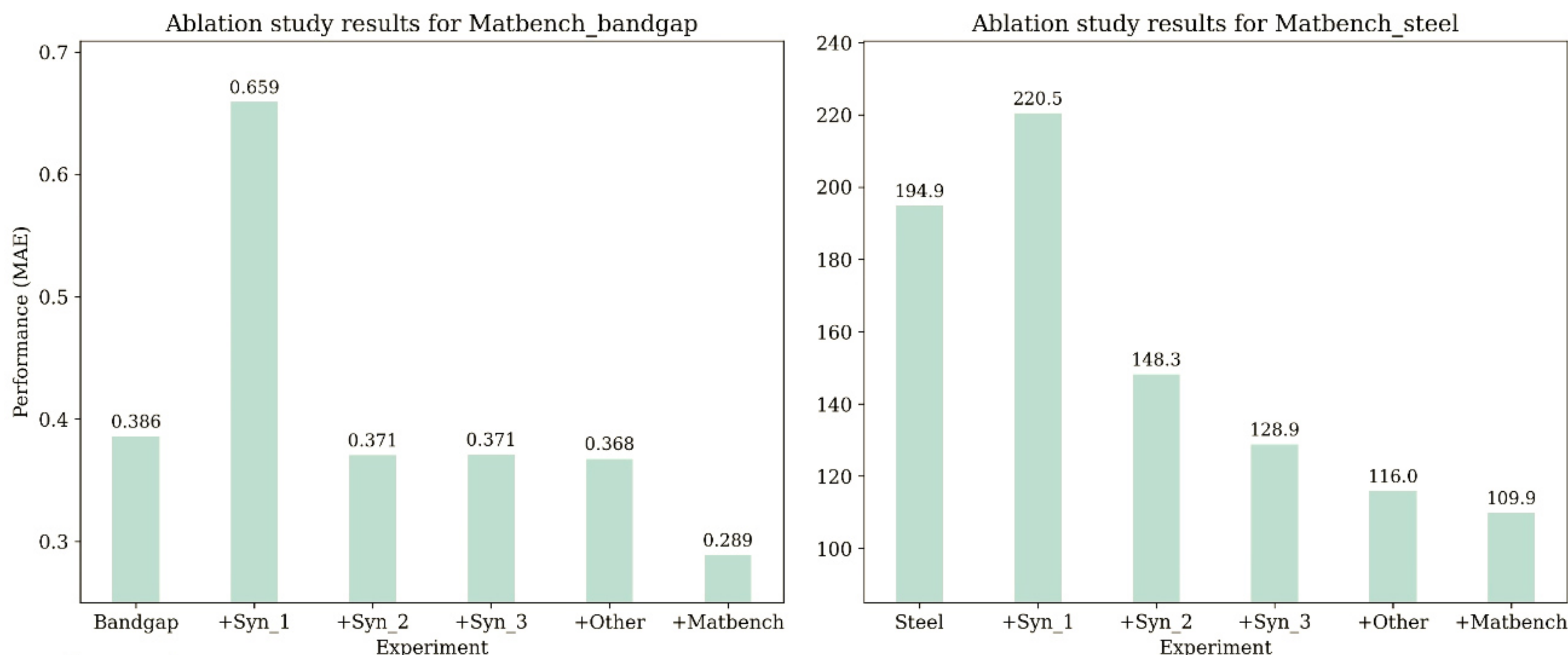
# FT-LLM Powerd Material Prediction

## ➤ Performance on bandgap prediction task

| Composition                        | Eg,exp | Eg,PBE      | Eg,HSE       | Eg,GWE     | Eg,AFLOWE   | Eg,SVR      | Eg,DARWIN 1.5 |
|------------------------------------|--------|-------------|--------------|------------|-------------|-------------|---------------|
| GaN                                | 3.2    | 1.62 (-49%) | 3.14 (-2%)   | 3.32 (4%)  | 1.85 (-42%) | 4.45 (39%)  | 2 (-38%)      |
| CdTe                               | 1.6    | 0.62 (-61%) | 1.52 (-3%)   | 1.76 (12%) | 0.67 (-57%) | 1.43 (-9%)  | 1.08 (-33%)   |
| LiF                                | 14.2   | 9.2 (-35%)  | 11.47 (-19%) | 15.1 (6%)  | 8.27 (-42%) | 9.87 (-30%) | 11.7(-18%)    |
| TiO <sub>2</sub>                   | 3.42   | 2.13 (-37%) | 3.67 (7%)    | 3.73 (9%)  | 2.09 (-38%) | 3.99 (16%)  | 3.3 (-4%)     |
| CuSbS <sub>2</sub>                 | 1.38   | 0.9 (-35%)  | 1.69 (22%)   | 1.1 (-20%) | 0.79 (-42%) | 1.39 (1%)   | 1.38 (0%)     |
| ZnS                                | 3.91   | 2.07 (-47%) | 3.49 (-11%)  | 4.15 (6%)  | 2.6 (-33%)  | 3.12 (-20%) | 2.9 (-26%)    |
| Cu <sub>2</sub> ZnSnS <sub>4</sub> | 1.6    | 0.28 (-83%) | 0.09 (-94%)  | 1.64 (3%)  | N/A         | 1.75 (9%)   | 1.46 (-8.8%)  |
| PbTe                               | 0.19   | 0 (-100%)   | 0.19 (0%)    | 0.26 (36%) | 0 (-100%)   | 0.2 (5%)    | 0.3 (58%)     |
| GaAs                               | 1.52   | 0.19 (-86%) | 1.12 (-26%)  | 1.52 (0%)  | 0.24 (-84%) | 1.28 (-15%) | 1.18 (-22%)   |
| ZnO                                | 3.44   | 0.67 (-81%) | 2.49 (-28%)  | 3.2 (-7%)  | 1.87 (-46%) | 3.41 (-1%)  | 3.2 (-7.0%)   |
| <b>MAD</b>                         |        | 1.65        | 0.67         | 0.22       | 1.6         | 0.75        | 0.72          |
| <b>RMSE</b>                        |        | 2.1         | 1            | 0.33       | 2.25        | 1.46        | 0.686         |



# Effectiveness of Merge Tra



## Ablation Experiments

- (1) Random Parameter LLM vs. Pretrained LLM
- (2) Professional Descriptor (SMILES, MOF) vs. Natural Language (Formula Name, Composition)
- (3) Multi-Task vs. Synthetic Data ( Real/Fake X & Real/Fake Y)

**Pre-training improves format learning; QA-FT improves instruction following;  
MTL improves transferability**





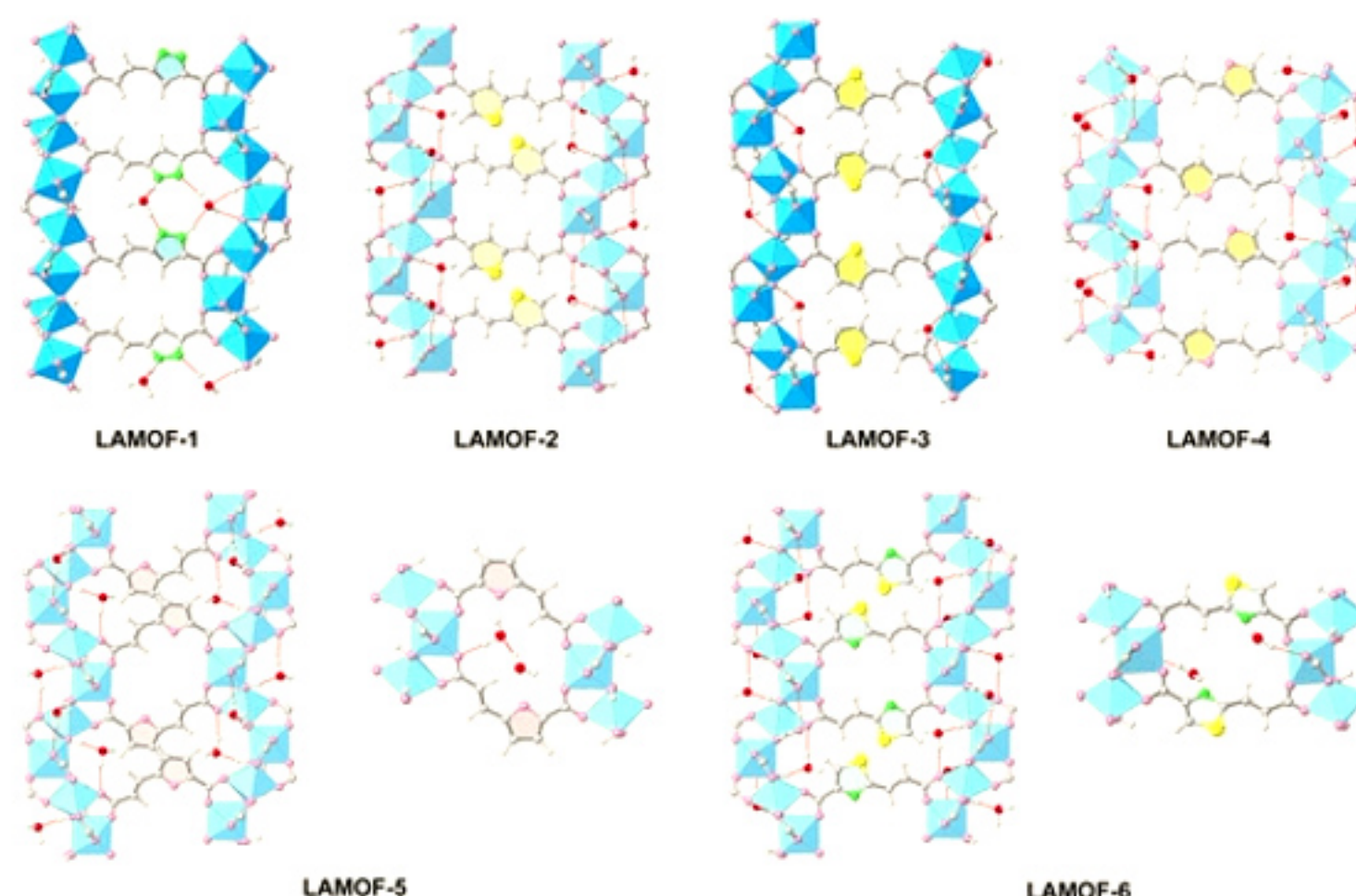
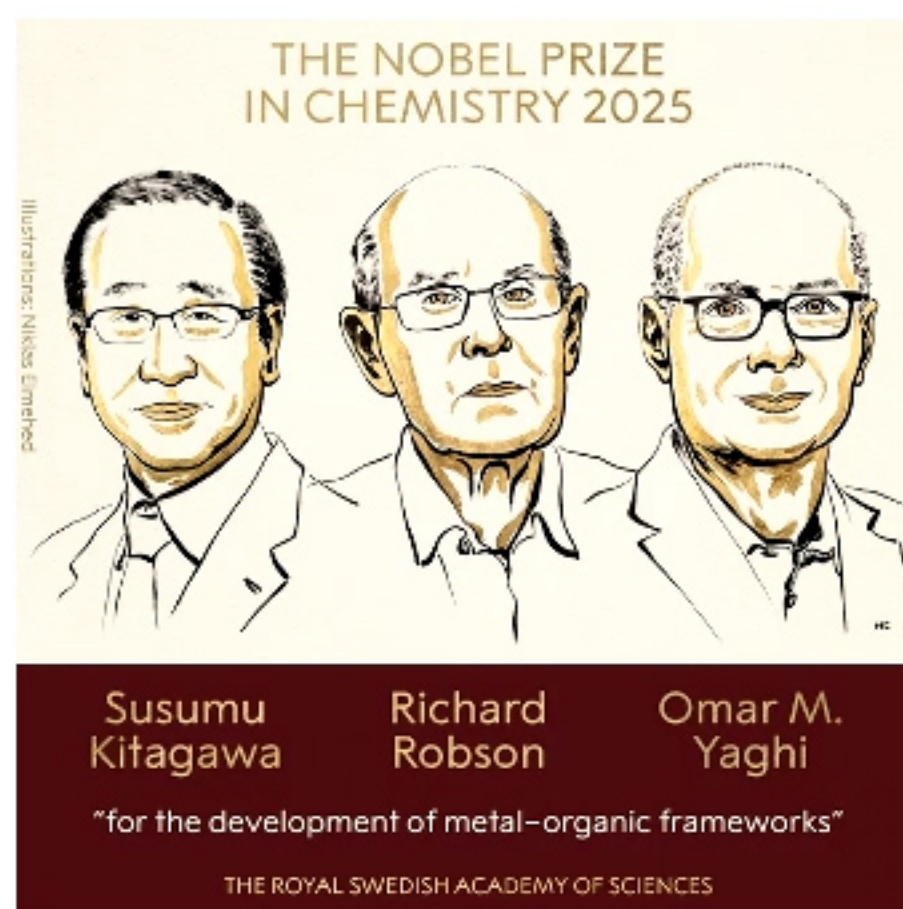
# We Unlocked a New Era in ML for chemistry

ARTICLE | December 13, 2023

## Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models

Zhiling Zheng, Ali H. Alawadhi, Saumil Chheda, S. Ephraim Neumann, Nakul Rampal, Shengchao Liu, Ha L. Nguyen, Yen-hsu Lin, Zichao Rong, J. Ilja Siepmann, Laura Gagliardi, Anima Anandkumar, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi\*

engineering strategies, (54,56–59) as the MOF linker mutation methods and actions are diverse, and the principles behind molecular editing encompass not only chemical rules like bonding and valency but also the construction of syntactically correct text-based chemical representations (Figures S3–S5). Fine-tuning through API, in comparison to in-context learning in prompt engineering, is not constrained by token limits, allowing the training of models on a substantial volume of examples and case studies (Table S2 and Figure S6). This approach is instrumental in adapting generalized pretraining models to more specific tasks in fields like chemistry. (60–62)



December 27, 2023  
Volume 145  
Number 51  
pubs.acs.org/JACS

# J | A | C | S

JOURNAL OF THE AMERICAN CHEMICAL SOCIETY



ACS Publications  
Most Trusted. Most Cited. Most Read.

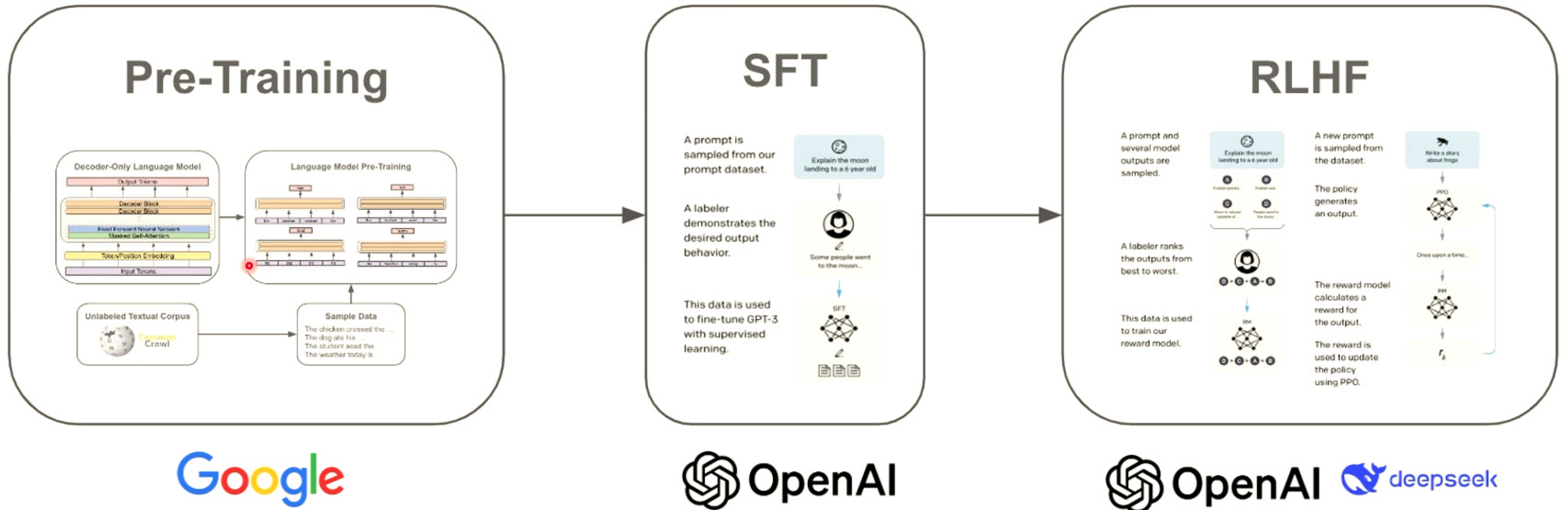
www.acs.org





# How to train a LLM?

## Alignment





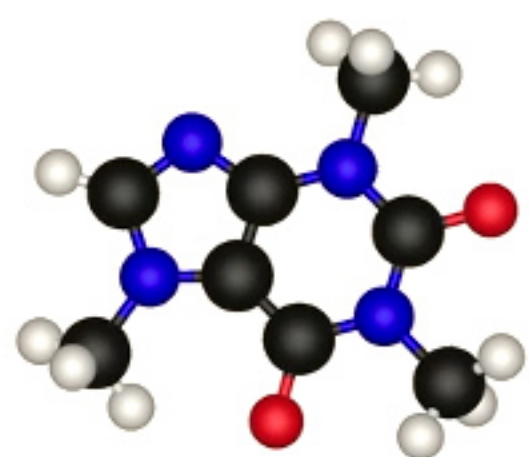
# CheMiST(2025.08) – Chemical Reasoning Model

First Reasoning Model for Scientific Field – “Chemical Deepseek”

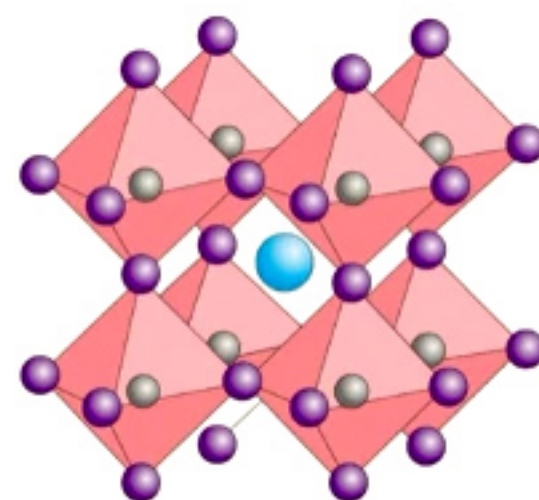


ALPS Cluster by Swiss AI Centre - >10K H200

## Mid Stage Scientific Training (MiST) (~3.9B)

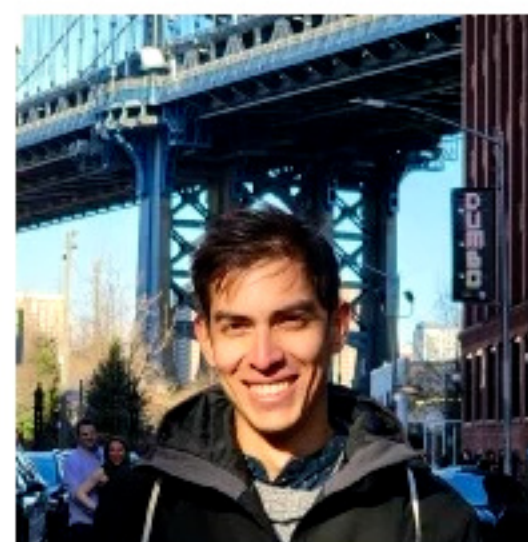


**750K** experimental  
Synthesised Molecular



**150K** experimental  
Synthesised Material

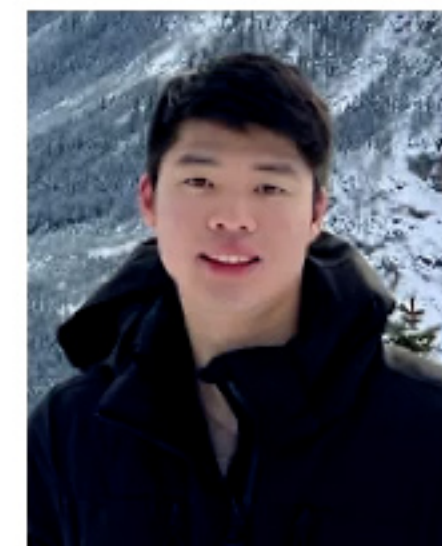
Largest Pre-training in the **Physical Science**



**Andres m  
Bran**  
EPFL



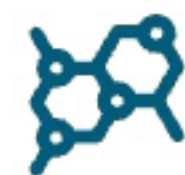
**Philippe  
Schwaller**  
EPFL



**Jeffery  
Meng**  
UNSW

## Reinforcement Learning from Scientific Feedback (RLSF)

Symbolic  
Competence



Chemical  
Knowledge



*Example Reward Function Design for Inorganic Material Structure Generation Tasks*

$$R = V \cdot \exp\left(-\frac{\sigma E}{\sigma_0}\right) \cdot \left(0.5 \cdot H(\epsilon - |\Delta E|) + H(|\Delta E| - \epsilon) \cdot \text{clip}\left(\frac{E_{\text{in}} - E_{\text{out}}}{S}, 0, 1\right)\right)$$

*Where the validity gate*

$$V = C \cdot H(d_{\min} - d) \cdot H\left(v - \left|\Delta \frac{V}{V}\right|\right) \cdot H(f - F_{\max})$$

Interatomic  
distance

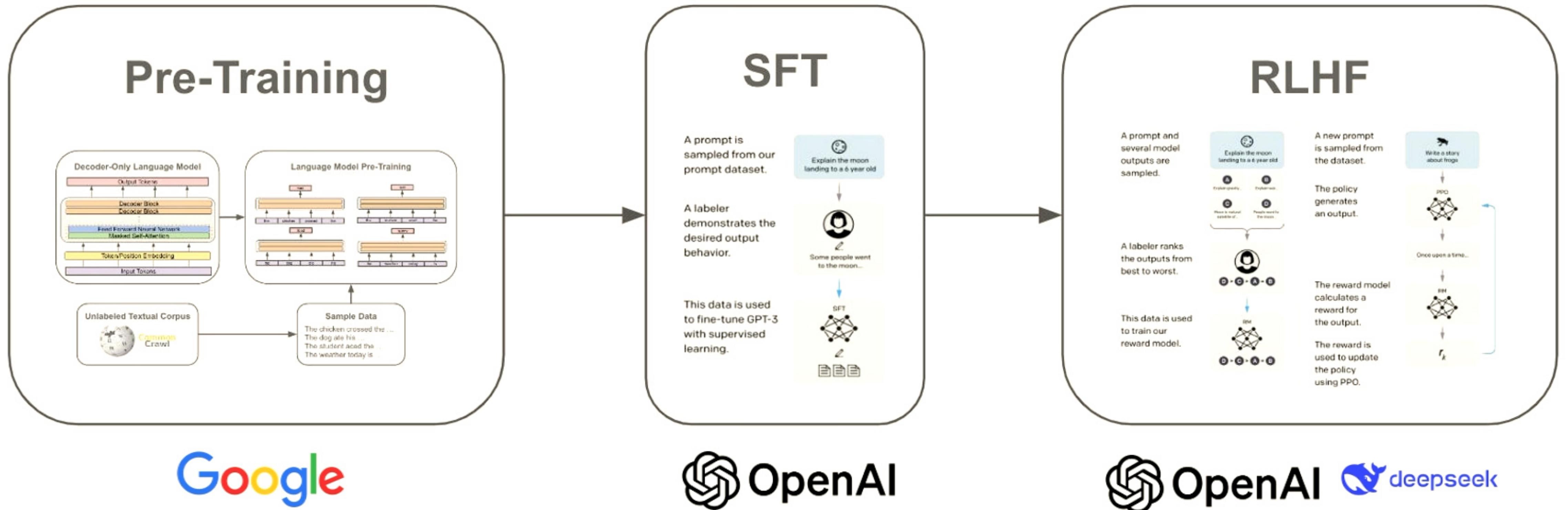
Volume  
Change

Atomic  
Force

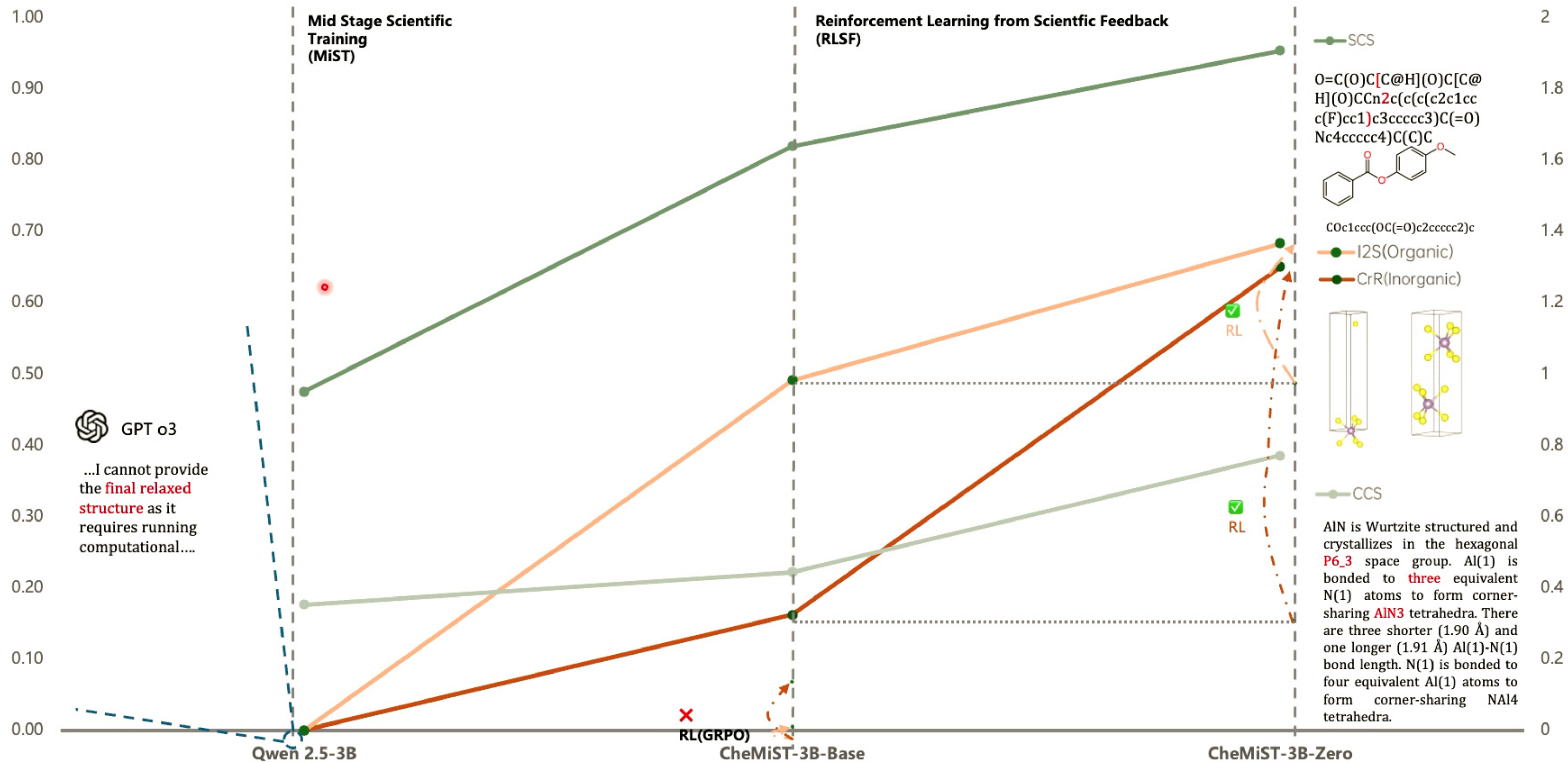


# How to train a LLM?

## Alignment









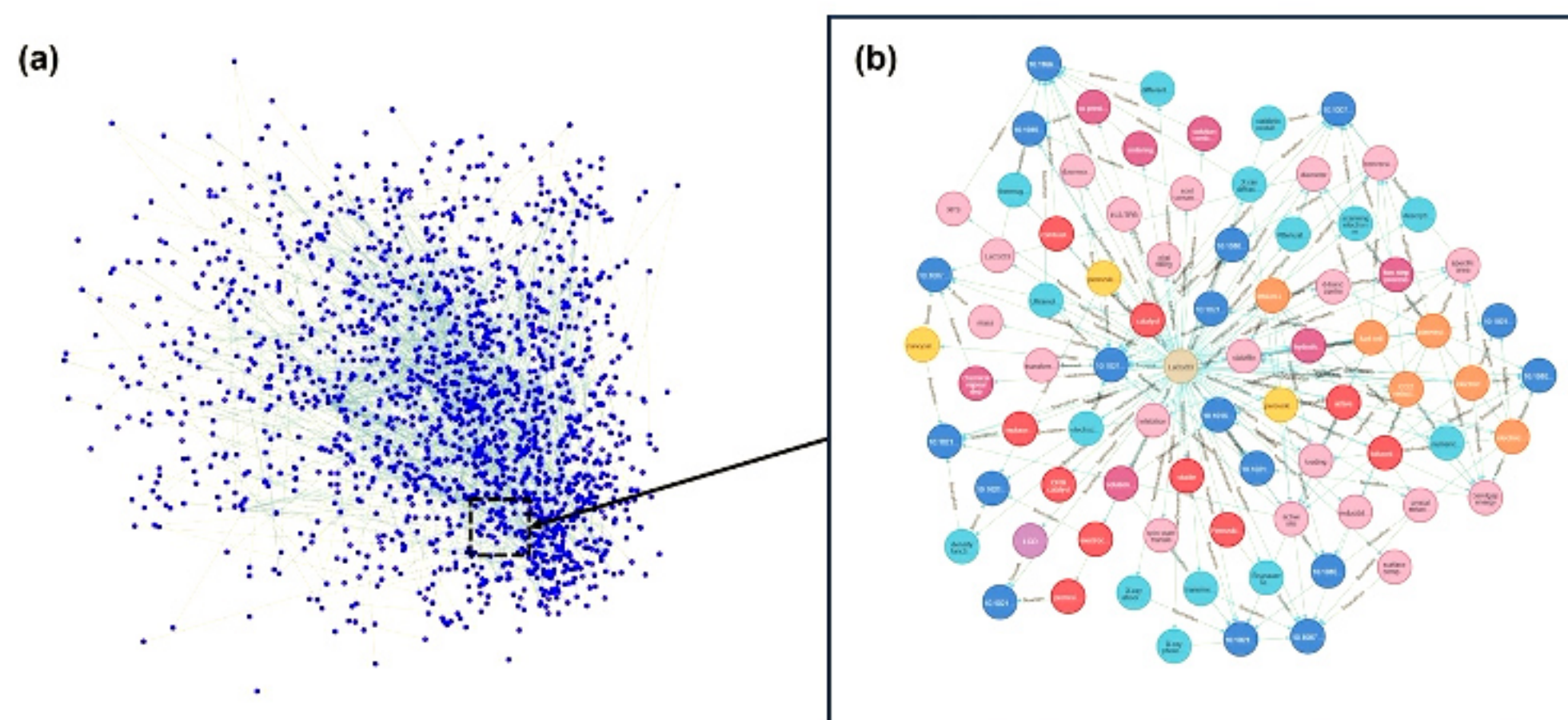
# The Road to “AI Scientists”

Scale up **foundational elements** like open datasets, specialised packages, and knowledge graphs



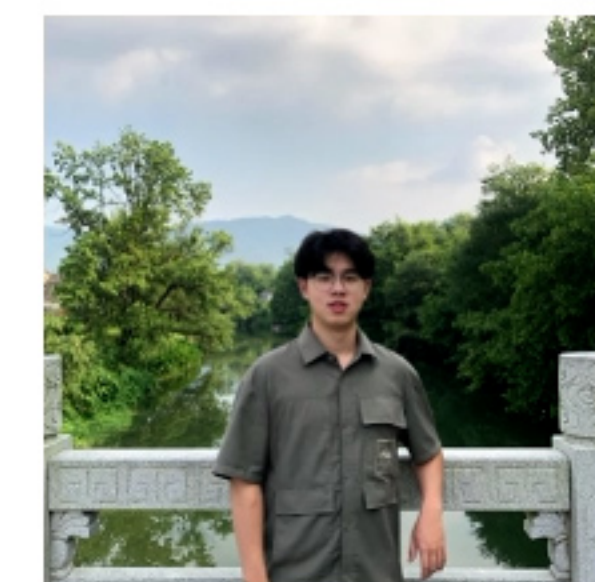


# MKG: LLM driven Material Science Knowledge Graph



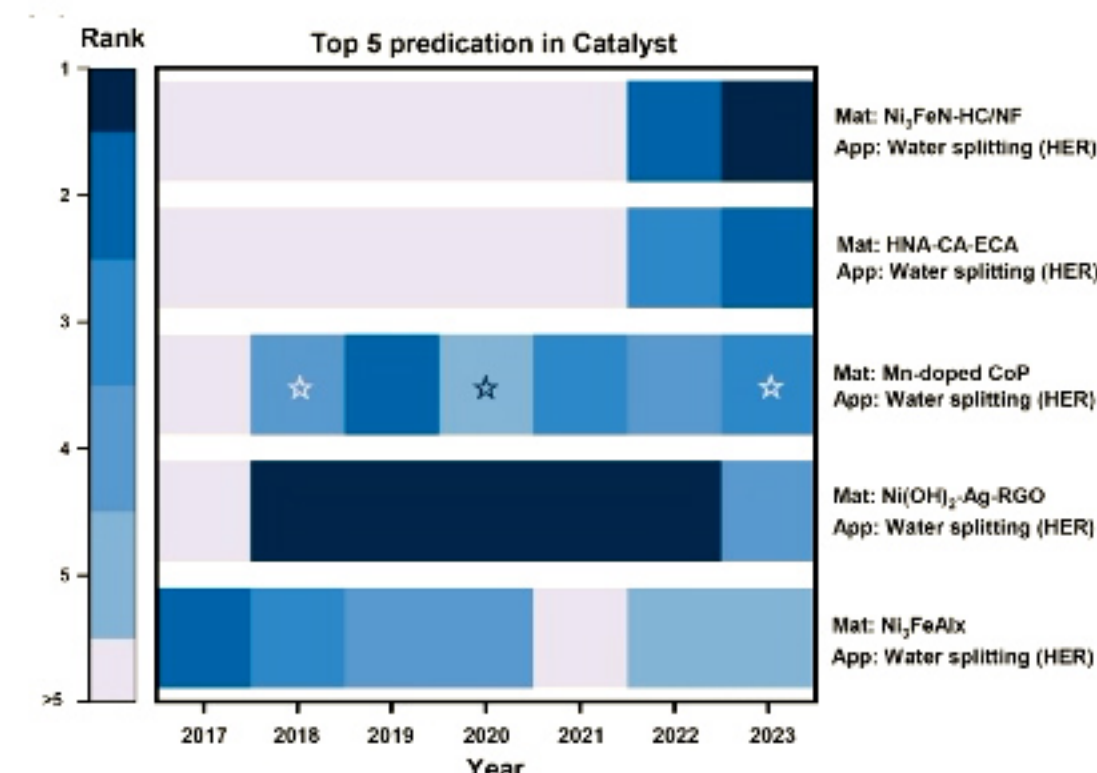
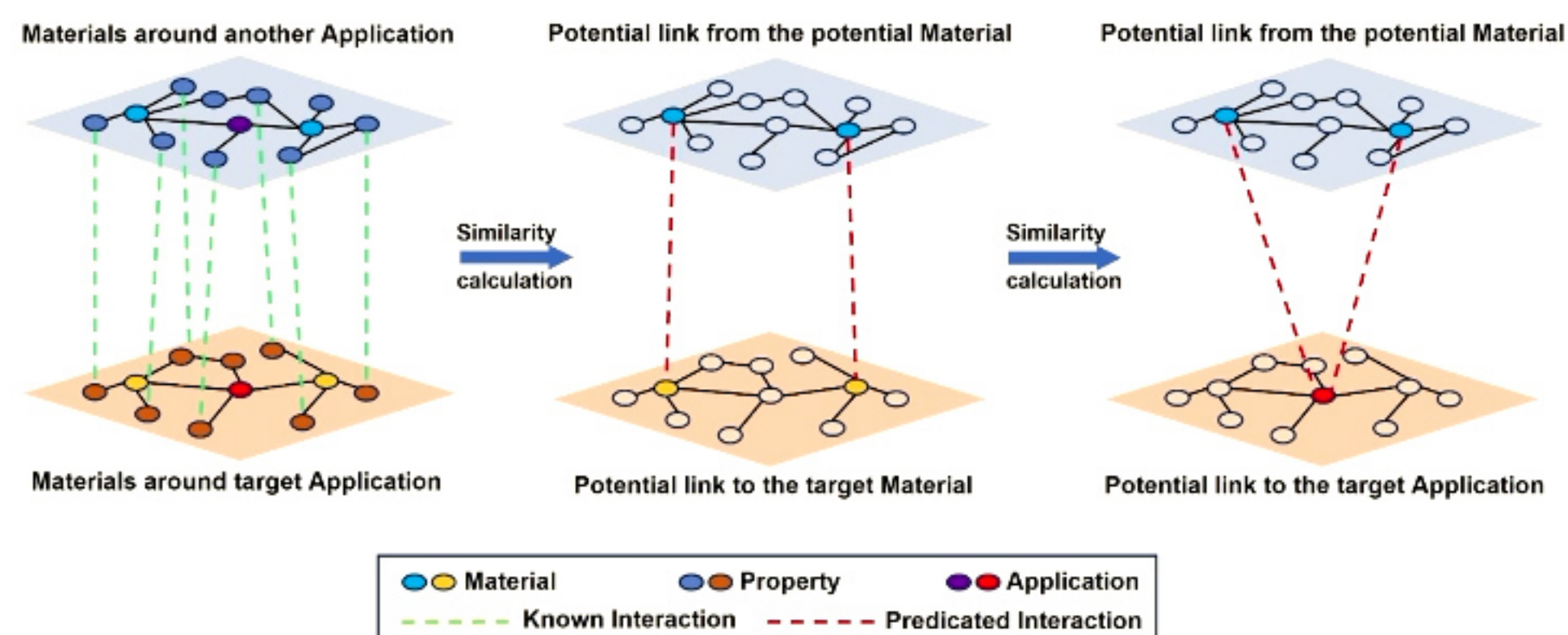
**162,605 Nodes**  
**731,772 Edges**  
**>10M Unique Triples**

**Largest KG in material science**



**Yanpeng Ye**  
 UNSW, KAUST

## Network-based Link Prediction

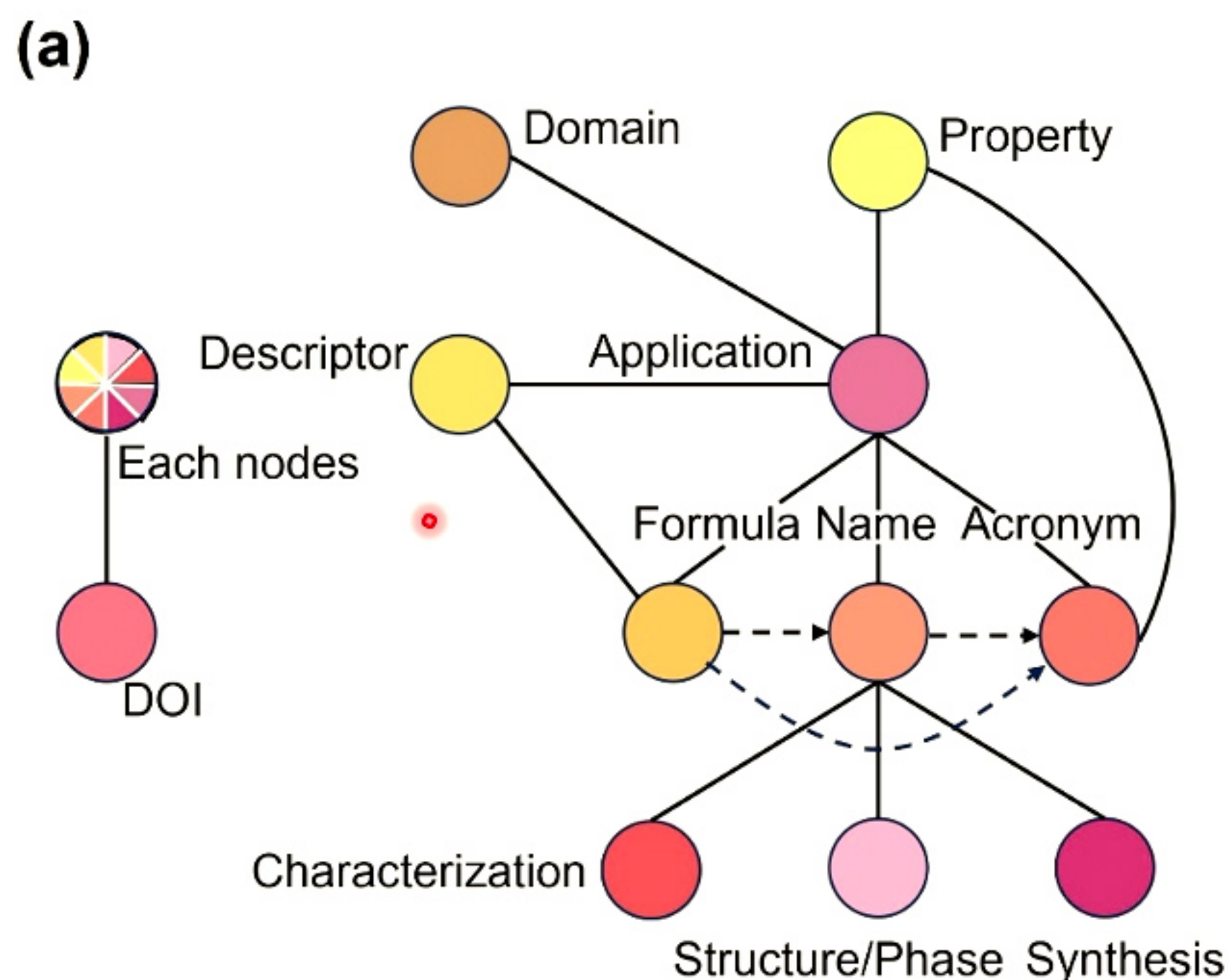


[1] Ye et al. Construction and Application of Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model, NeurIPS 2024

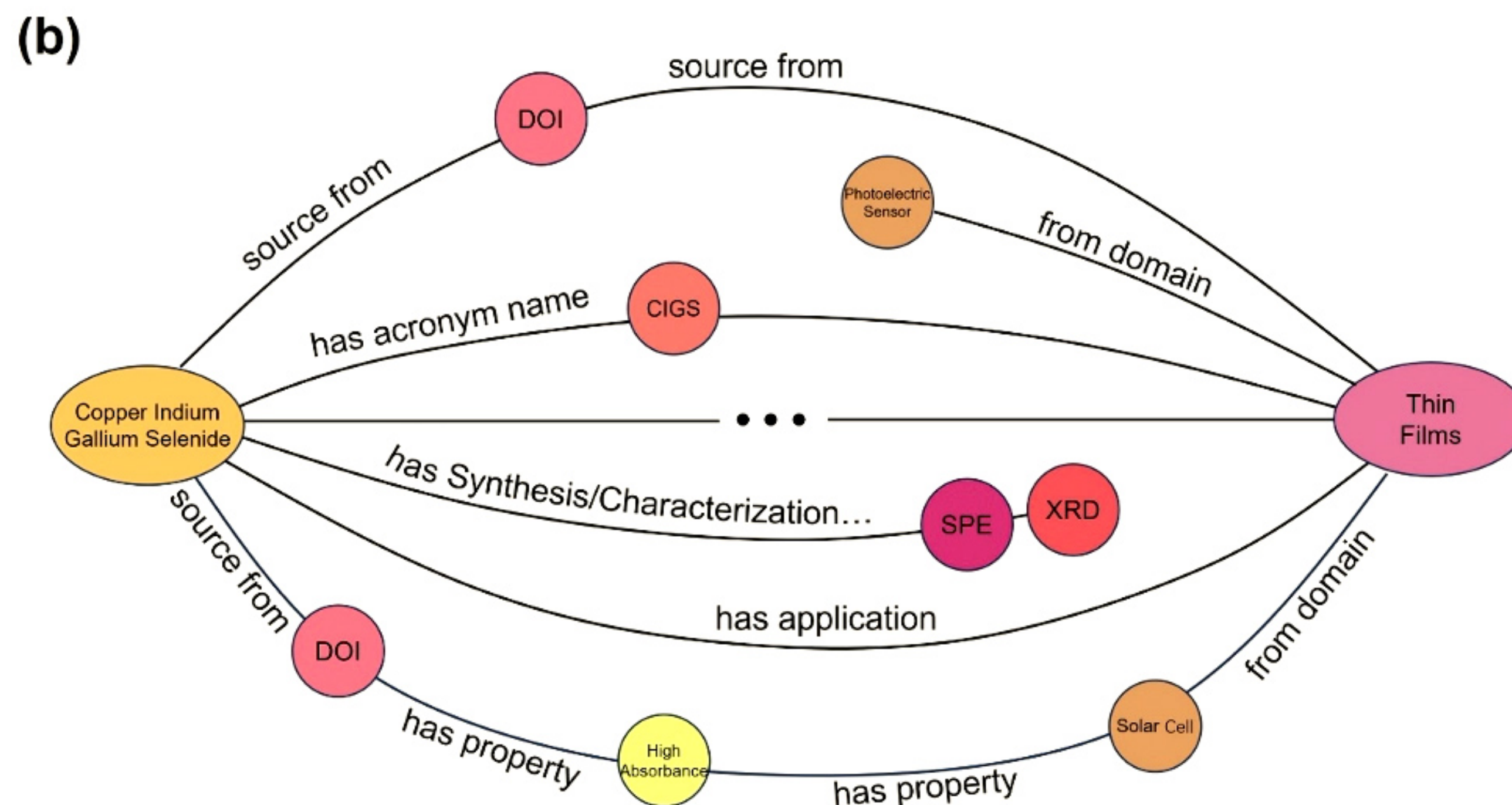


# MKG: LLM driven Material Science Knowledge Graph

- MKG Schema



- **Example Path**





# MKG: LLM driven Material Science Knowledge Graph

## Paper:

...  
DOI: xxxx1  
TiO2 nanosheets with exposed (001) facets were fabricated by a simple reduction deposition ...  
...  
DOI: xxxx2  
TiO2 are grown by atomic layer deposition and used in perovskite solar cells...  
...

## Named entity recognition:

...  
DOI: xxxx1  
Entity : TiO2, nanosheets, (001) facets, reduction deposition ...  
...  
DOI: xxxx2  
Entity : TiO2, atomic layer deposition, perovskite solar cells ...  
...

## Entity resolution :

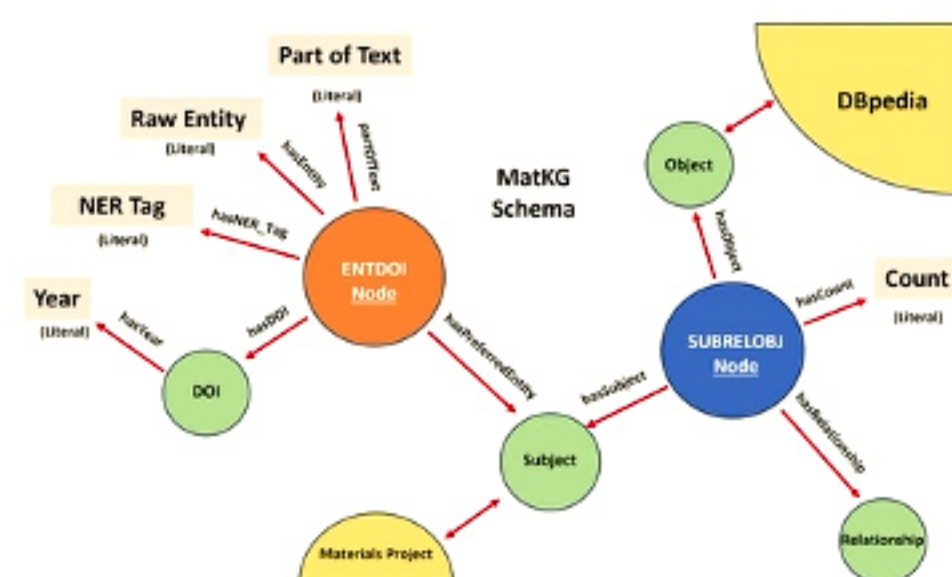
...  
Entity A: TiO2  
Entity B: nanosheets  
Entity C: (001) facets  
Entity D: reduction deposition  
Entity E: atomic layer deposition  
Entity F: perovskite solar cells ...  
...

## Triples:

...  
(A, has\_structure, B)  
(A, has\_structure, C)  
(A, has\_synthesis, D)  
(A, has\_synthesis, E)  
(A, has\_application, F)  
...

## Relation extraction:

...  
(A, B): has\_structure  
(A, C): has\_structure  
(A, D): has\_synthesis  
(A, E): has\_synthesis  
(A, F): has\_application  
...

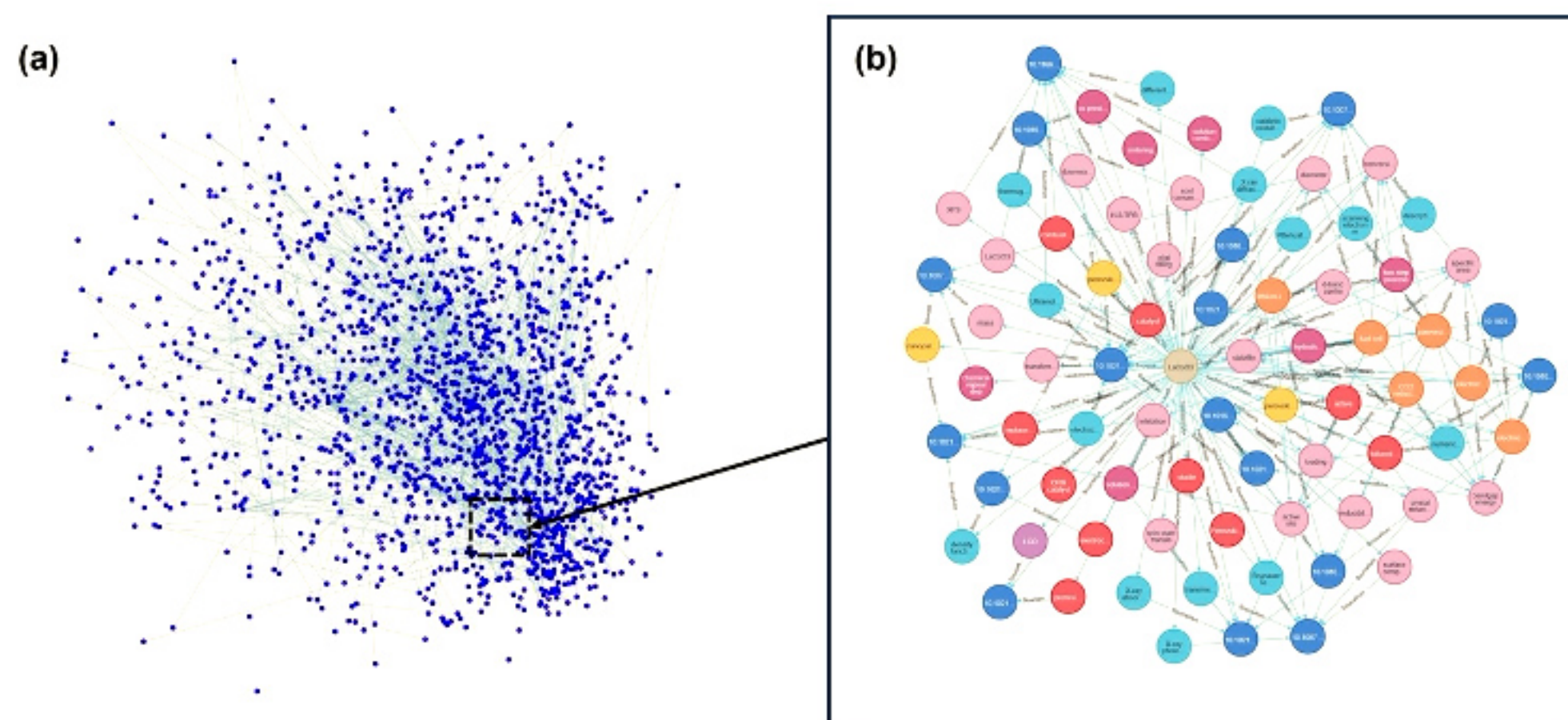


MatKG



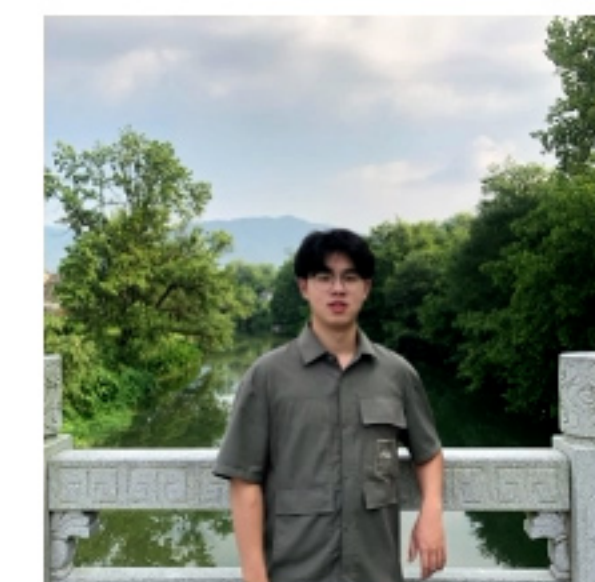


# MKG: LLM driven Material Science Knowledge Graph



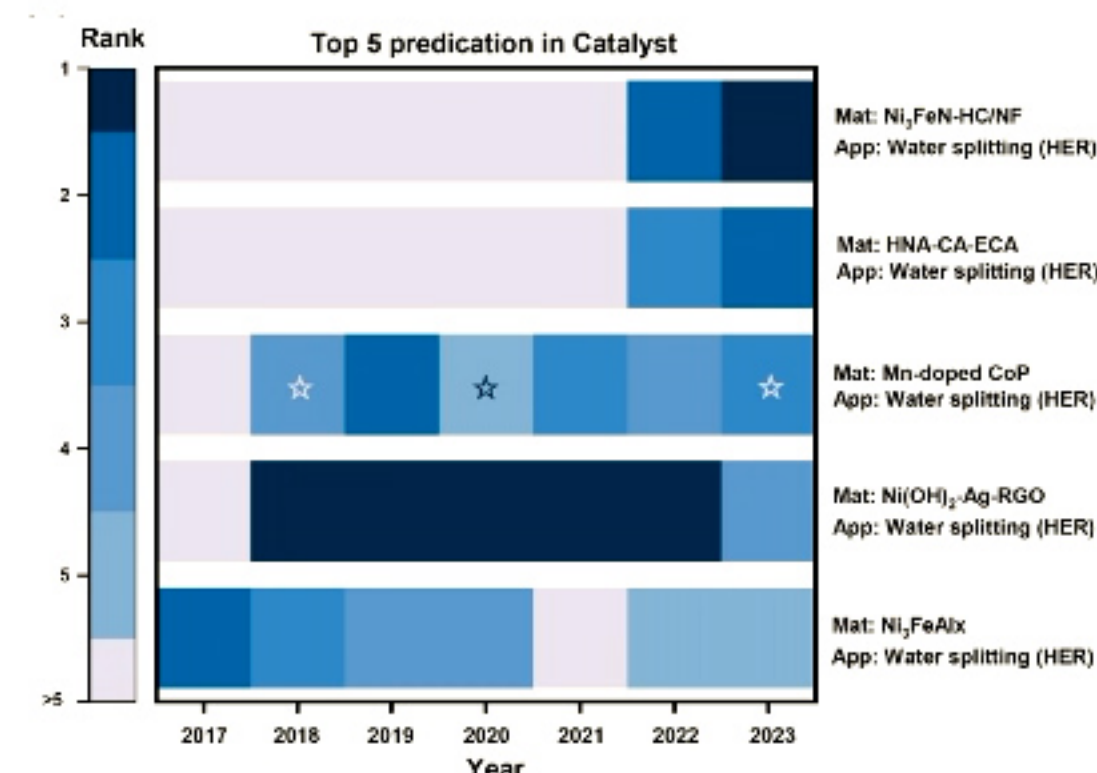
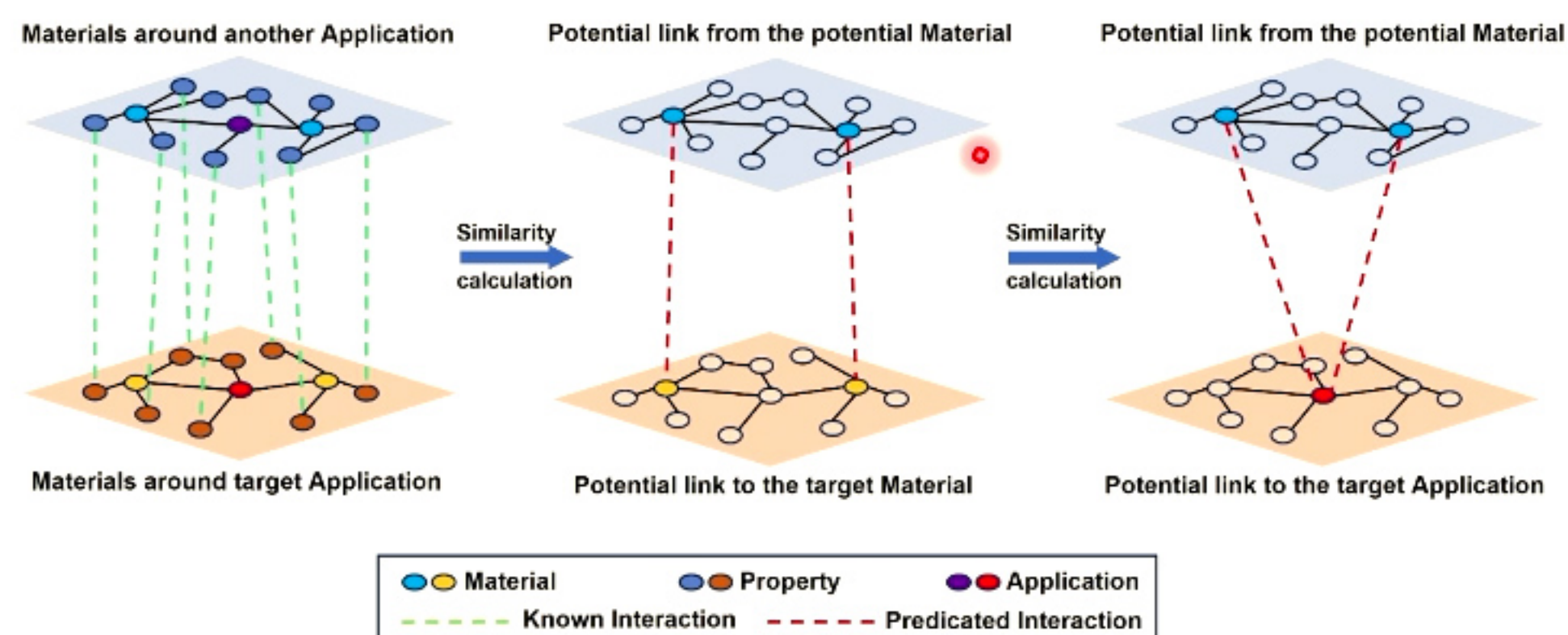
162,605 Nodes  
731,772 Edges  
>10M Unique Triples

Largest KG in material science



Yanpeng Ye  
UNSW, KAUST

## Network-based Link Predication



[1] Ye et al. Construction and Application of Materials Knowledge Graph in Multidisciplinary Materials Science via Large Language Model, NeurIPS 2024



# MKG: LLM driven Material Science Knowledge Graph

## Paper:

...  
DOI: xxxx1  
TiO2 nanosheets with exposed (001) facets were fabricated by a simple reduction deposition ...  
...  
DOI: xxxx2  
TiO2 are grown by atomic layer deposition and used in perovskite solar cells...  
...

## Named entity recognition:

...  
DOI: xxxx1  
Entity : TiO2, nanosheets, (001) facets, reduction deposition  
...  
DOI: xxxx2  
Entity : TiO2, atomic layer deposition, perovskite solar cells  
...

## Entity resolution :

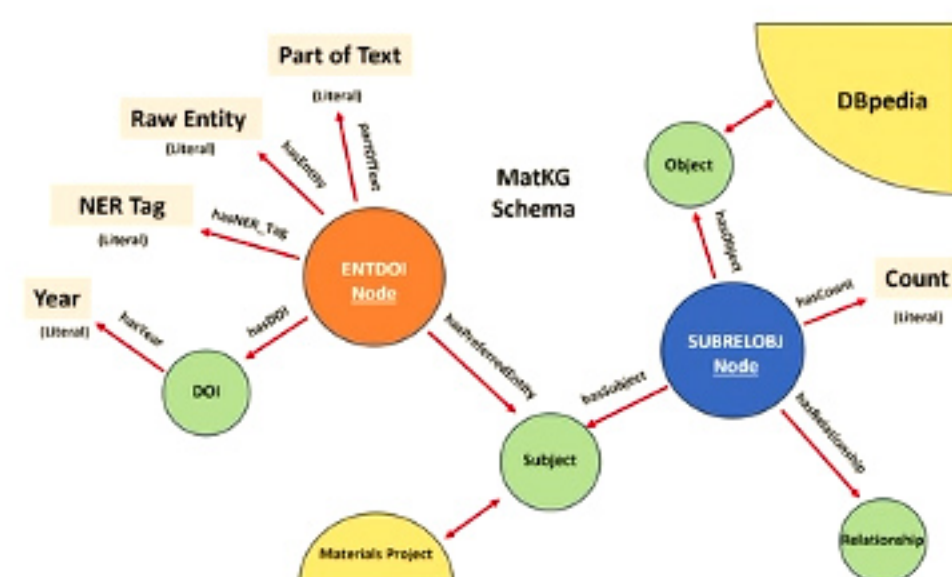
...  
Entity A: TiO2  
Entity B: nanosheets  
Entity C: (001) facets  
Entity D: reduction deposition  
Entity E: atomic layer deposition  
Entity F: perovskite solar cells  
...

## Triples:

...  
(A, has\_structure, B)  
(A, has\_structure, C)  
(A, has\_synthesis, D)  
(A, has\_synthesis, E)  
(A, has\_application, F)  
...

## Relation extraction:

...  
(A, B): has\_structure  
(A, C): has\_structure  
(A, D): has\_synthesis  
(A, E): has\_synthesis  
(A, F): has\_application  
...



MatKG





# The Road to “AI Scientists”

AI needs to really understand the **latent knowledge, like structure** behind texts.





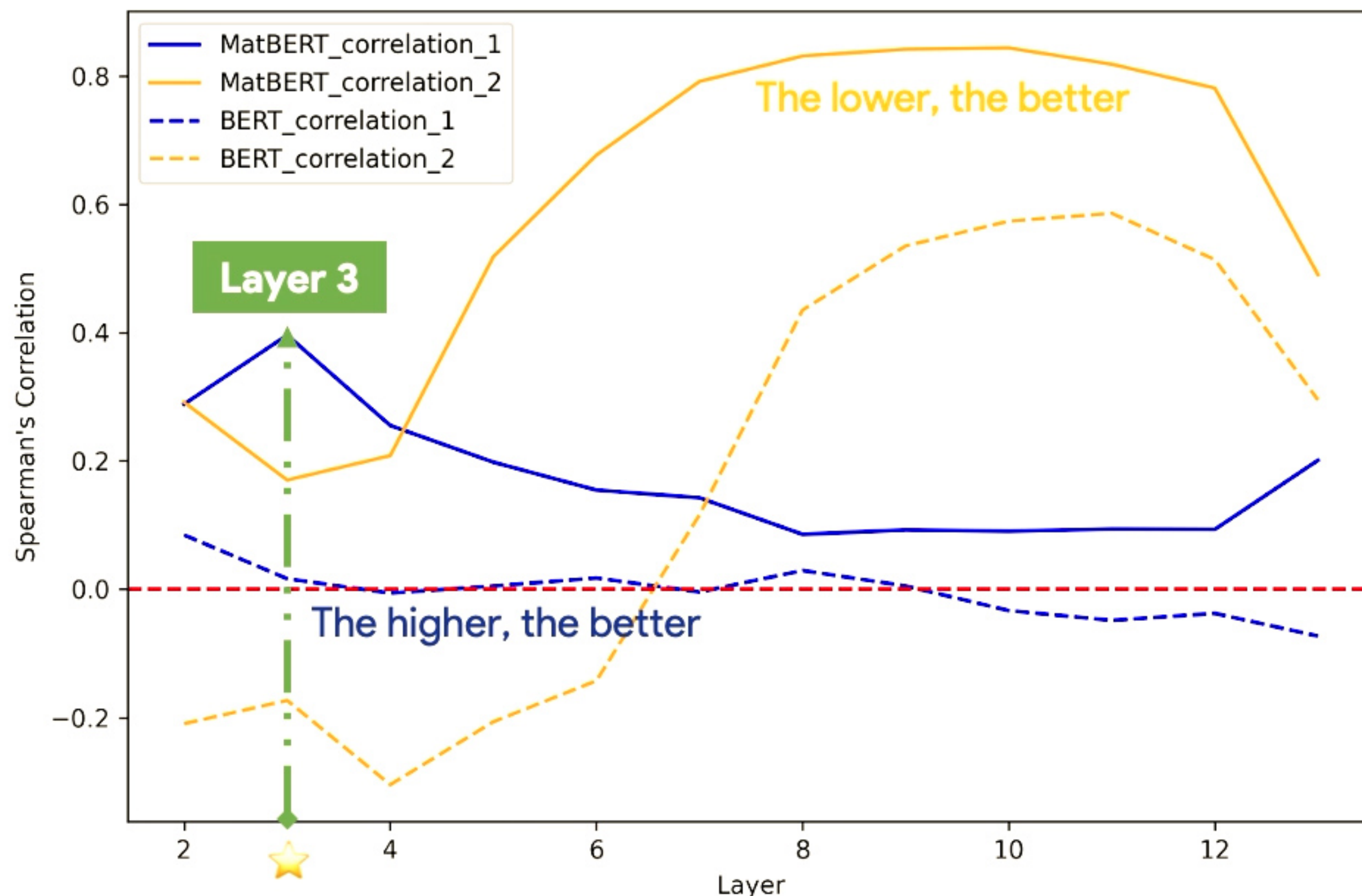
# From Token to Discovery

## Previously

- **Word2vec** embedding is used for material prediction (better than DFT on zT prediction)
- Out-of-vocabulary (OOV) problem and limited paper resources (cannot find candidates from literature)

## Tokeniser effect

- Material names composed of fewer tokens tend to be ranked higher (which is not aligned with reality).
- The **low tokeniser effect** is a necessary but not sufficient condition for good prediction performance.
- BERT embedding needs contrastive learning to improve representation quality



The Spearman's correlation between the predicted rank of 84 materials and their actual rank as 'Correlation\_1', and the Spearman's correlation between the predicted rank of 84 materials and their tokenized length as 'Correlation\_2'.



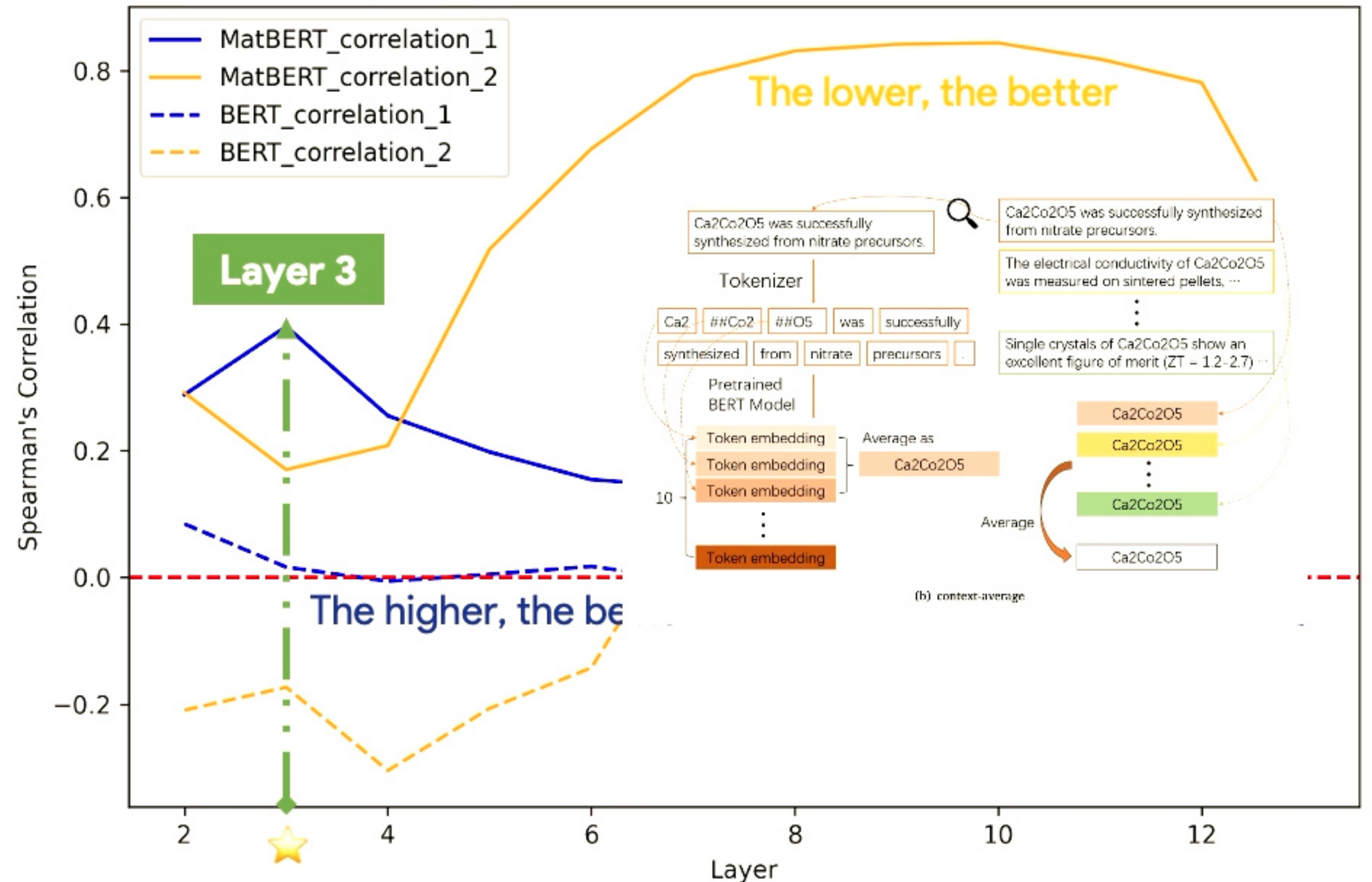
# From Token to Discovery

## Previously

- **Word2vec** embedding is used for material prediction (better than DFT on zT prediction)
- Out-of-vocabulary (OOV) problem and limited paper resources (cannot find candidates from literature)

## Tokeniser effect

- Material names composed of fewer tokens tend to be ranked higher (which is not aligned with reality).
- The **low tokeniser effect** is a necessary but not sufficient condition for good prediction performance.
- BERT embedding needs contrastive learning to improve representation quality



The Spearman's correlation between the predicted rank of 84 materials and their actual rank as 'Correlation\_1', and the Spearman's correlation between the predicted rank of 84 materials and their tokenized length as 'Correlation\_2'.



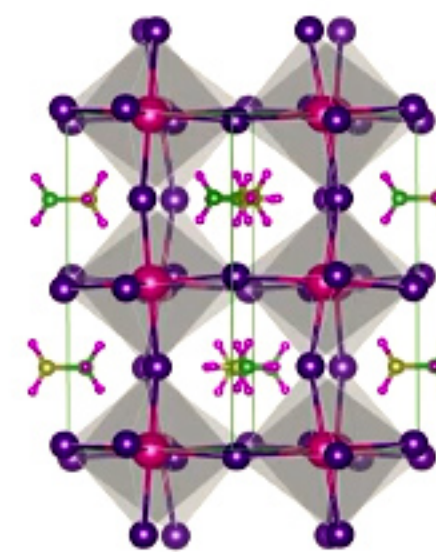
# MatFusion: Universal Embedding for Material

- We're already creating AI-generated content; can we do the same for materials?

A dog in Beach  
**GPT o3**



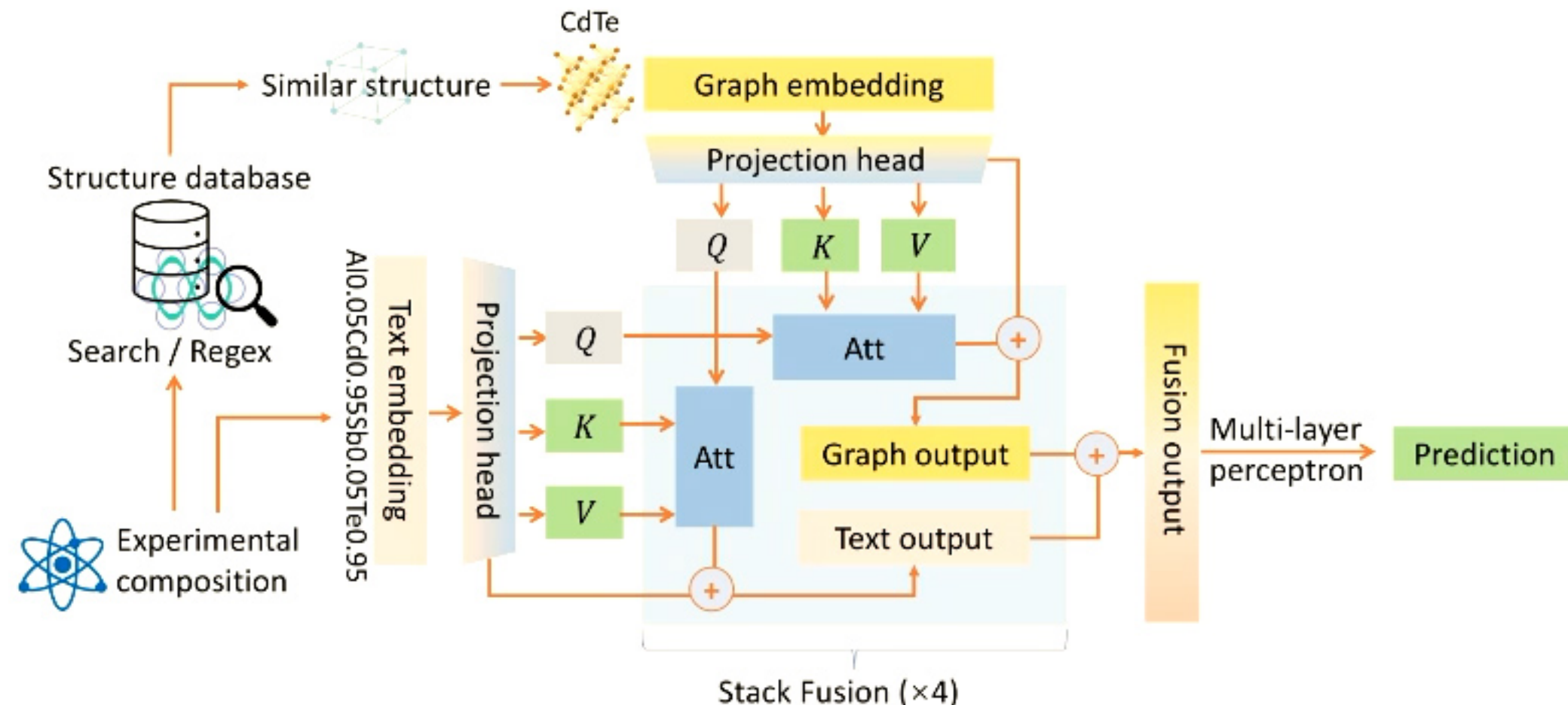
A high efficiency  
optoelectrical  
materials  
???



Yingheng  
Cornell

We need a model, learn basic physics information, and have experimental science knowledge together!

- **MatFusion**: Multimodal framework bridging 3D physical structure and experimental science



Text guided crystal generation

**1000x efficient than high-throughput DFT!**

- In the thermoelectrics property benchmark, **53.4%** accuracy improvement.

| Model  | Embedding Type | MAE <sub>exp</sub> | STD <sub>exp</sub> | MAE <sub>comp</sub> | STD <sub>comp</sub> |
|--------|----------------|--------------------|--------------------|---------------------|---------------------|
| BERT   | Text@Formula   | 0.6201             | 0.0266             | 0.6978              | 0.0288              |
|        | Fusion@Formula | 0.3720             | 0.0251             | 0.2751              | 0.0037              |
| T5     | Text@Formula   | 0.5585             | 0.0663             | 0.6353              | 0.0276              |
|        | Fusion@Formula | 0.3803             | 0.0278             | <b>0.2698</b>       | 0.0018              |
| LLaMA2 | Text@Formula   | 0.4844             | 0.0360             | 0.5514              | 0.0062              |
|        | Fusion@Formula | 0.3782             | 0.0195             | 0.2781              | 0.0036              |
| DARWIN | Text@Formula   | 0.4714             | 0.0298             | 0.5409              | 0.0131              |
|        | Fusion@Formula | <b>0.3596</b>      | 0.0170             | 0.2857              | 0.0062              |



[1] Xie, T, et al. MATFUSION: A multi-modal framework bridging llms and structural embeddings for experimental materials property prediction, ICLR 2025 AI4Mat Workshop





# AtomWorld: Benchmarking CIF modification aptitude

Prior work such as **MatterGen**, **CrystalLM** for generating materials are limited to simple crystals (training data is the limitation).

- We argue the task of CIF modification - although requires strong reasoning capability from the LLM - is a more tractable problem because it isn't limited by training data. **Learns *how* rather than *what* to generate.**



**Alex Chen**  
UNSW

Table 1: Actions and the corresponding Action Prompt for AtomWorld.

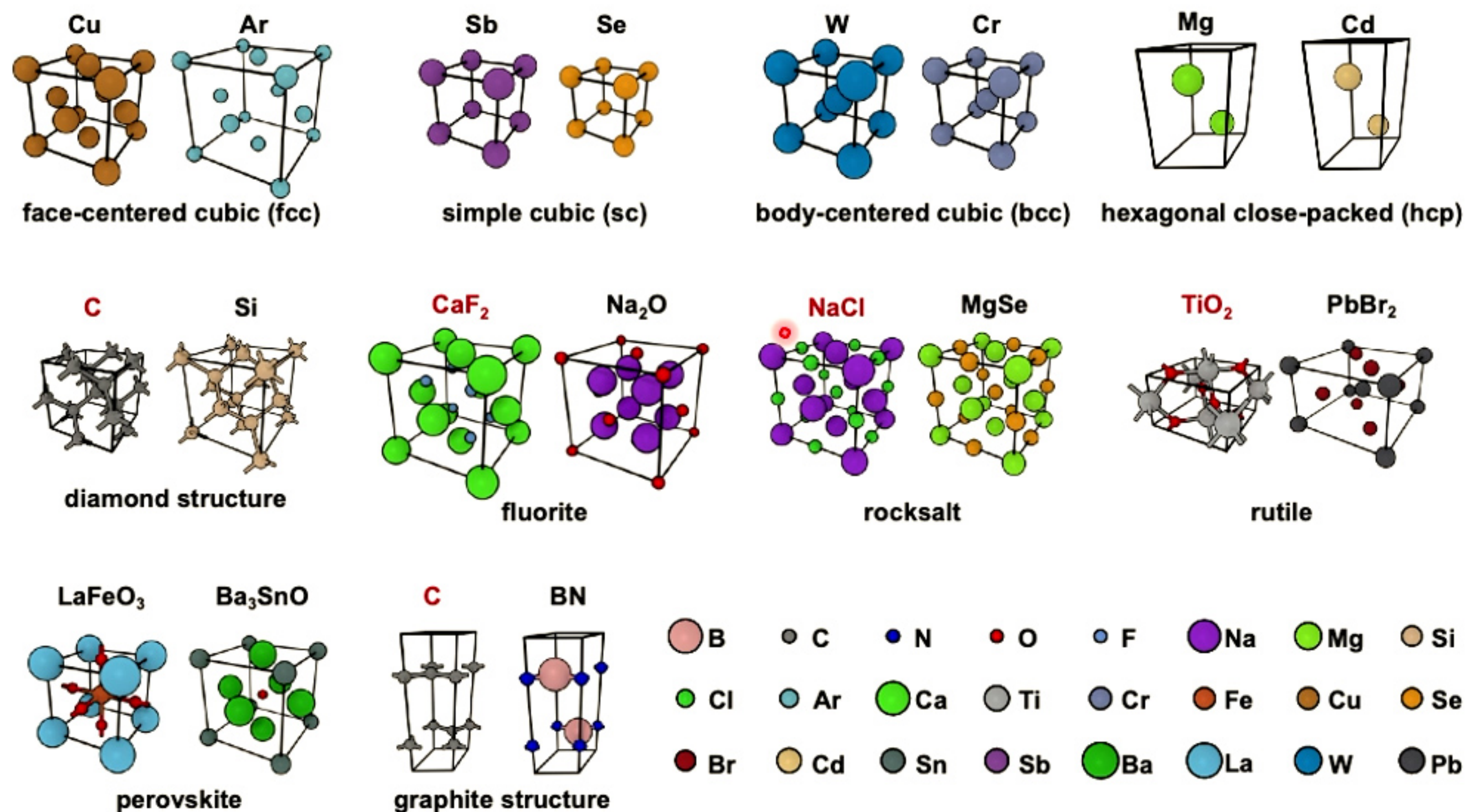
| Action name    | Action prompt                                                                                                                                                                                          |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| change         | Change the atom at index {index} into {new_symbol} in the cif file. The indices of atoms are started from 0.                                                                                           |
| remove         | Remove the atom at index {index} from the cif file. The indices of atoms are started from 0.                                                                                                           |
| add            | Add one {symbol} atom at the Cartesian coordinate {position} to the cif file.                                                                                                                          |
| move           | Move the atom at index {index} by {d_pos} angstrom in the cif file.                                                                                                                                    |
| move_towards   | Move the atom at index {index1} towards the atom at index {index2} by {distance} angstrom in the cif file.                                                                                             |
| insert_between | Insert a {symbol} atom in the line between atoms at indices {index1} and {index2}, and the inserted atom must be {distance:.2f} angstrom from atom at {index1} in the cif file.                        |
| swap           | Swap atoms at indices {index1} and {index2} in the cif file. The indices of atoms are started from 0.                                                                                                  |
| delete_below   | Delete all atoms whose z coordinate is lower than the atom at index {index} in the cif file. Excluding itself and atoms with the same z coordinate.                                                    |
| rotate_around  | Rotate all surrounding atoms within {radius} angstrom of the center atom at index {index} by {angle} degree around the axis {axis} in the cif file. The rotation should following the right-hand rule. |
| super_cell     | Create a supercell with the size {dim_0}x{dim_1}x{dim_2}.                                                                                                                                              |





# AtomWorld: Benchmarking CIF modification aptitude

| Structure type | gemini 2.5 pro | gpt o3 | gpt o4-mini | deepseek chat | llama3-70b | qwen3-32b | qwen3-14b | qwen3-8b | qwen3-4b |
|----------------|----------------|--------|-------------|---------------|------------|-----------|-----------|----------|----------|
| graphite       | 1              | 2      | 1           | 1             | 0          | 0         | 1         | 0        | 0        |
| perovskite     | 2              | 2      | 2           | 0             | 0          | 0         | 1         | 0        | 0        |
| rutile         | 2              | 2      | 2           | 1             | 0          | 2         | 0         | 0        | 0        |
| rocksalt       | 2              | 2      | 1           | 1             | 0          | 1         | 2         | 0        | 0        |
| fluorite       | 2              | 2      | 2           | 1             | 0          | 2         | 1         | 0        | 0        |
| diamond        | 2              | 2      | 2           | 2             | 0          | 2         | 0         | 0        | 0        |
| hcp            | 2              | 2      | 1           | 2             | 0          | 1         | 1         | 1        | 0        |
| bcc            | 2              | 2      | 2           | 2             | 0          | 1         | 1         | 1        | 0        |
| sc             | 2              | 2      | 2           | 2             | 0          | 2         | 1         | 0        | 0        |
| fcc            | 2              | 2      | 2           | 2             | 0          | 2         | 2         | 1        | 0        |

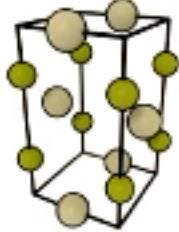
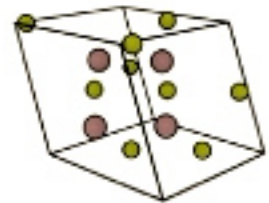
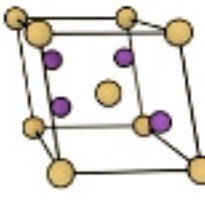


The number of correctly generated CIFs for each structure type in the CIF-Gen task. The squares marked in red indicate cases where the single correct generation is the standard prototype. The right side shows the specific 3D crystal structures for each type, where the chemical compositions in red represent the standard prototypes





# AtomWorld: Benchmarking CIF modification aptitude

| Structure & Goal                                                                                                               | Core thinking process                                                                                                                                                                                                                                                                                                                                                                             |
|--------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>PtS (mp-288)</p>                           | <p>... a potent strategy is to raise the energy of the valence band maximum. This can be achieved by substituting the sulfur atoms with a heavier, less electronegative element from the same group, such as Selenium (Se).<br/><i>The Se 4p orbitals are higher in energy than the S 3p orbitals.</i></p>                                                                                        |
| <p>Band gap ↓</p>                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                   |
| <p>Ga<sub>2</sub>S<sub>3</sub> (mp-539)</p>  | <p>... its band gap is largely determined by the electronegativity difference and bond strength between the Gallium (Ga) cation and the Sulfur (S) anion. <i>To increase the band gap, a modification that strengthens the chemical bonds and increases the material's ionicity is required.</i></p>                                                                                              |
| <p>Band gap ↑</p>                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                   |
| <p>CdAs<sub>2</sub> (mp-471)</p>            | <p>... Cd is a relatively large and electropositive element, leading to bonds with significant ionic character that are less stiff than purely covalent bonds.<br/>... Zn is directly above Cadmium in Group 12 of the periodic table. <i>It has a smaller atomic radius and is more electronegative, which leads to the formation of shorter, stronger, and more covalent bonds with As.</i></p> |
| <p>Bulk modulus ↑</p>                                                                                                          |                                                                                                                                                                                                                                                                                                                                                                                                   |

*Excerpts from Gemini 2.5 Pro reasoning trace during successful completions of StructProp tasks. **Correct** and **incorrect** identifications are highlighted.*

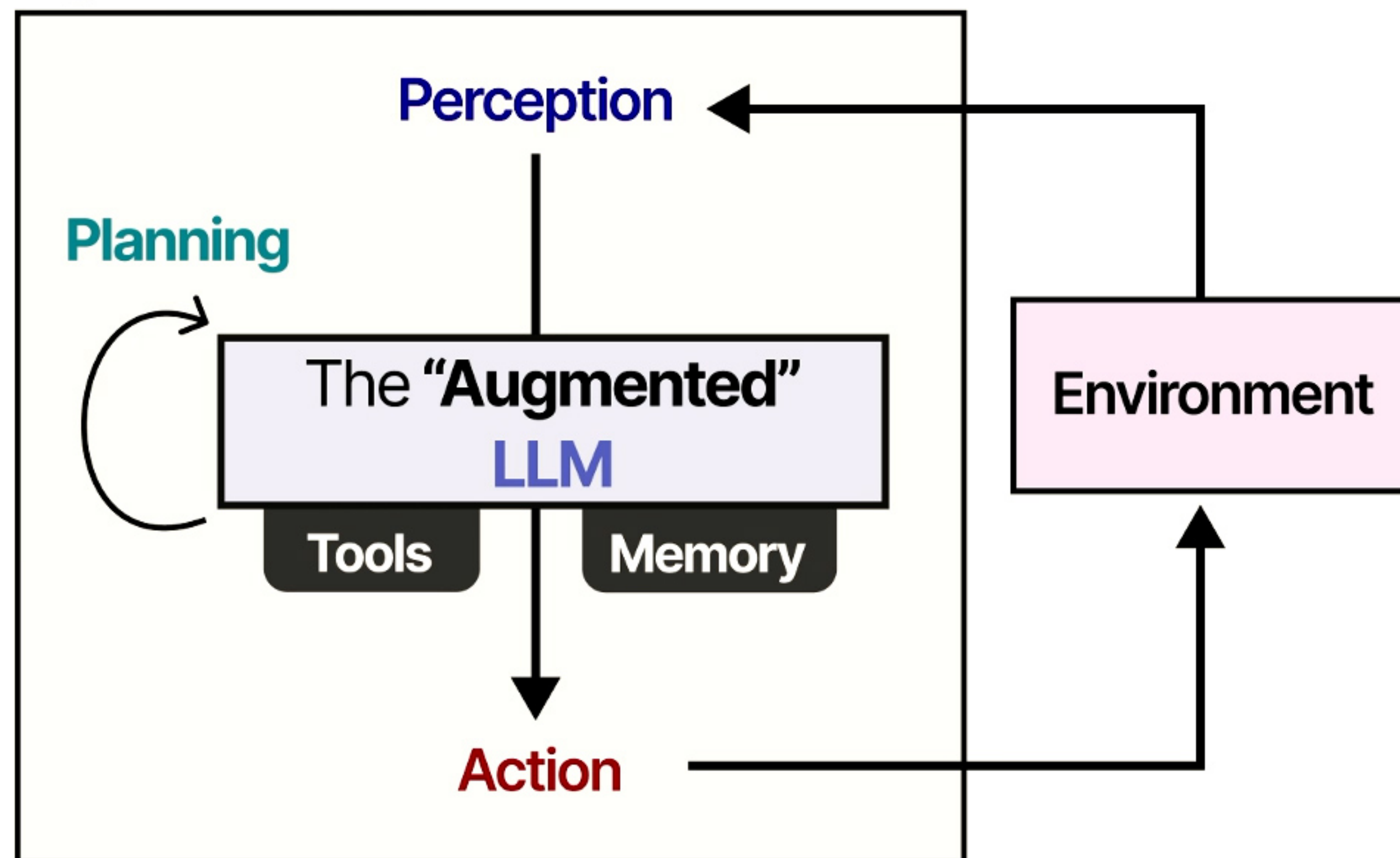




# The Road to “AI Scientists”

**Agent** is coming,  
**the arms** of the  
foundational Model

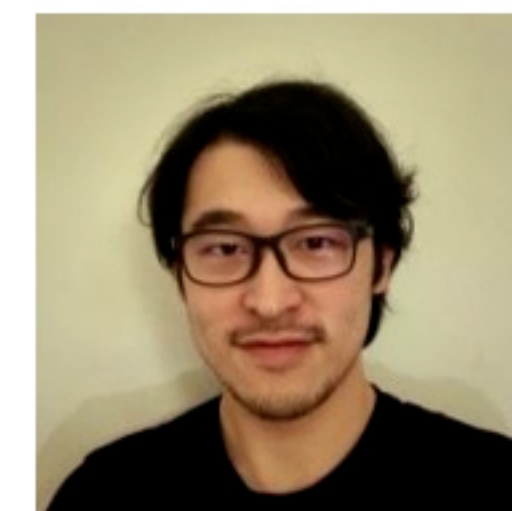
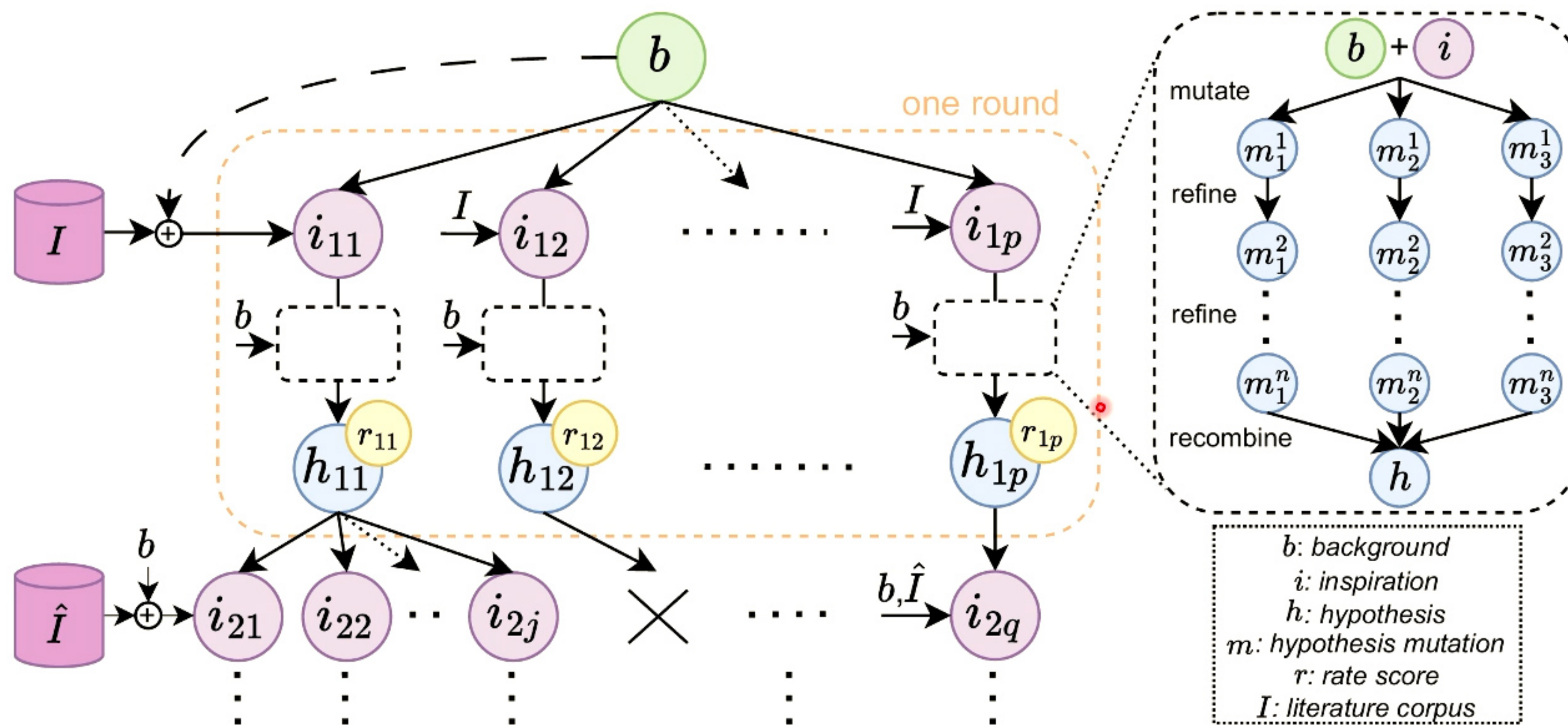
Agent



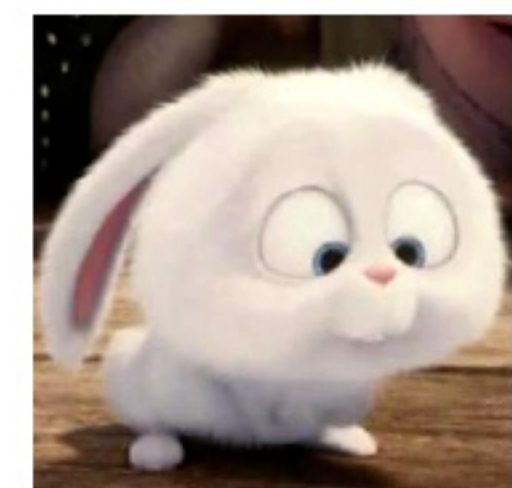


# MOOSE-Chem: LLMs for Hypothesis Generation

Yes, it can! By decoupling research background from inspiration and utilizing diverse operations, AI can automatically generate feasible and innovative scientific ideas.



**Zonglin Yang**  
NTU



**Dongzhan Zhou**  
SAIL





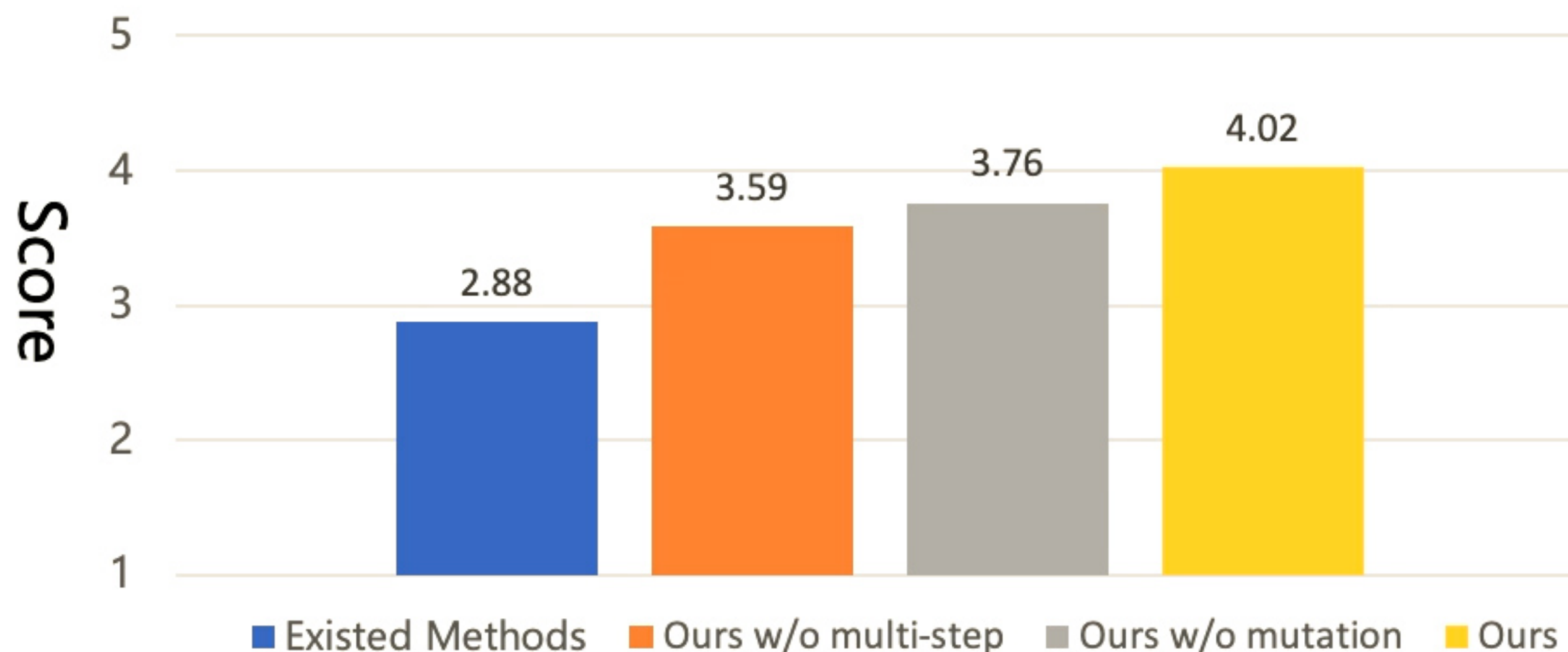
# MOOSE-Chem: LLMs for Hypothesis Generation

**Benchmark:** Including Nature, Science, and sub-journal-level benchmarks to evaluate the model's ability to rediscover knowledge.

## Experiment Results

- The model can retrieve inspiration from the corpus with a top-5% retrieval success rate exceeding 80%.
- Based on the inspiration, it can generate highly matched ideas.

| Category             | Count |
|----------------------|-------|
| Polymer Chemistry    | 21    |
| Organic Chemistry    | 22    |
| Inorganic Chemistry  | 3     |
| Analytical Chemistry | 5     |
| Total                | 51    |





# MOOSE-Chem: LLMs for Hypothesis Generation

Generated *h*: A pioneering integrated electrocatalytic system leveraging **ruthenium** nanoparticles embedded in **nitrogen-doped** graphene, combined with a dual palladium-coated ion-exchange membrane reactor, will catalyze efficient, scalable, and site-selective reductive deuteration of aromatic hydrocarbons and heteroarenes. Utilizing deuterium sources from both  $D_2$  gas and  $D_2O$ , this system will optimize parameters through real-time machine learning-driven dynamic adjustments. Specific configurations include ruthenium nanoparticle sizes (2-4 nm), nitrogen doping levels (12-14%), precisely engineered palladium membranes (5 micrometers, ensuring 98% deuterium-selective permeability), and advanced cyclic voltammetry protocols (1-5 Hz, -0.5V to -1.5V).

Ground truth *h*: The main hypothesis is that a **nitrogen-doped ruthenium (Ru)** electrode can effectively catalyze the reductive deuteration of (hetero)arenes in the presence of  $D_2O$ , leading to high deuterium incorporation into the resulting saturated cyclic compounds. The findings validate this hypothesis by demonstrating that this electrocatalytic method is highly efficient, scalable, and versatile, suitable for a wide range of substrates.

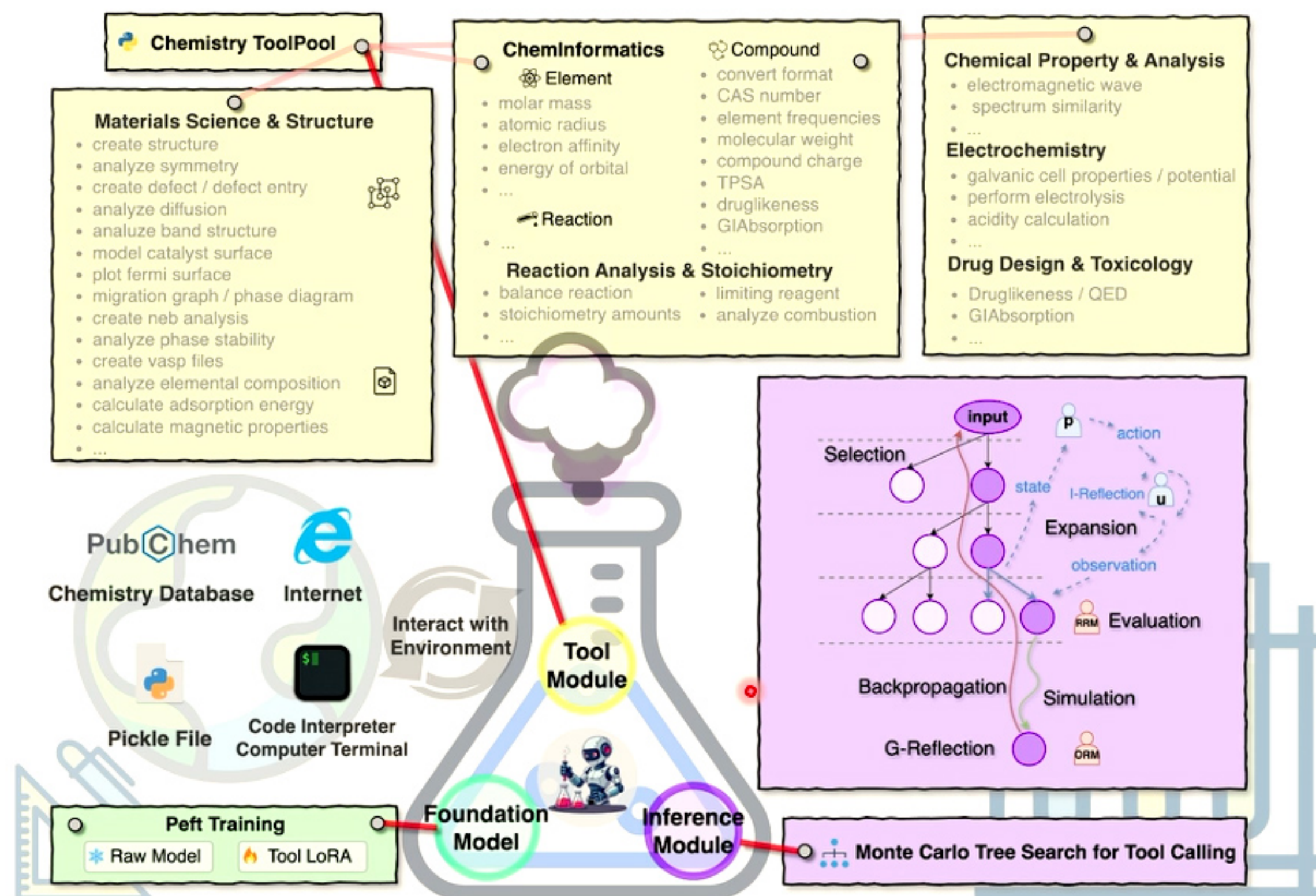
Expert's analysis: The proposed hypothesis effectively covers two key points from the ground truth hypothesis: **the incorporation of ruthenium (Ru) and the use of  $D_2O$  as a deuterium source** within the electrocatalytic system. However, the current content does not detail the mechanism by which Ru-D is produced, which is essential for explaining the process of reductive deuteration. Nevertheless, the results are still insightful. The specific level of nitrogen doping, for example, is highly suggestive and warrants further investigation. Overall, the match remains strong in its alignment with the original hypothesis while also presenting opportunities for deeper exploration.





# ChemMatAgent: Tree-Search Based Tool Learning

➤ **ChemMatAgent** - Independently planning, reasoning, and acting to achieve scientific goals  
By integrating large language models with laboratory automation, these AI agents will autonomously plan experiments, analyse results, and suggest new directions for exploration.

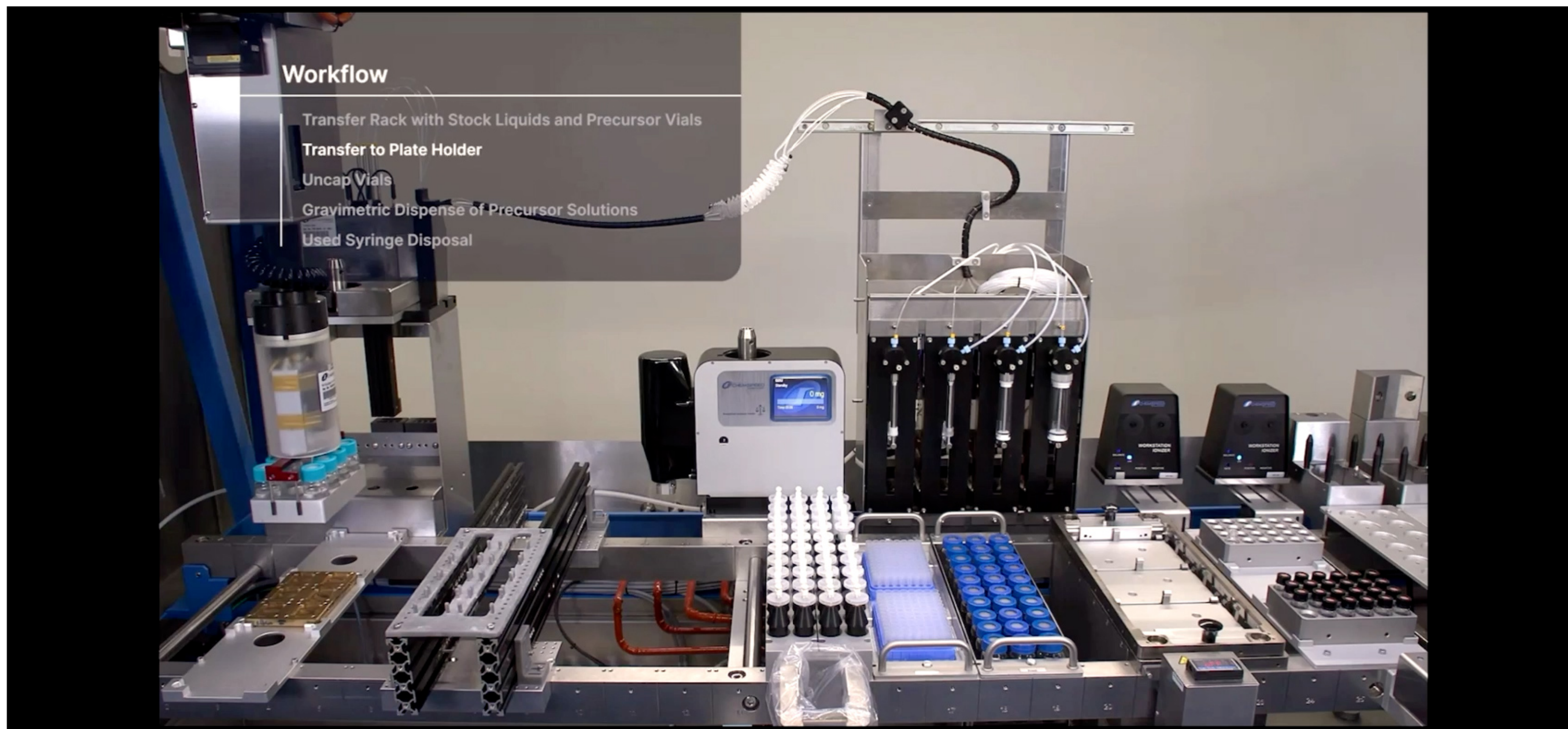


[1] M Wu, T Xie, D Zhou et.al ChemAgent: Enhancing LLMs for Chemistry and Materials Science through Tree-Search Based Tool Learning, ACL 2026 Under Review



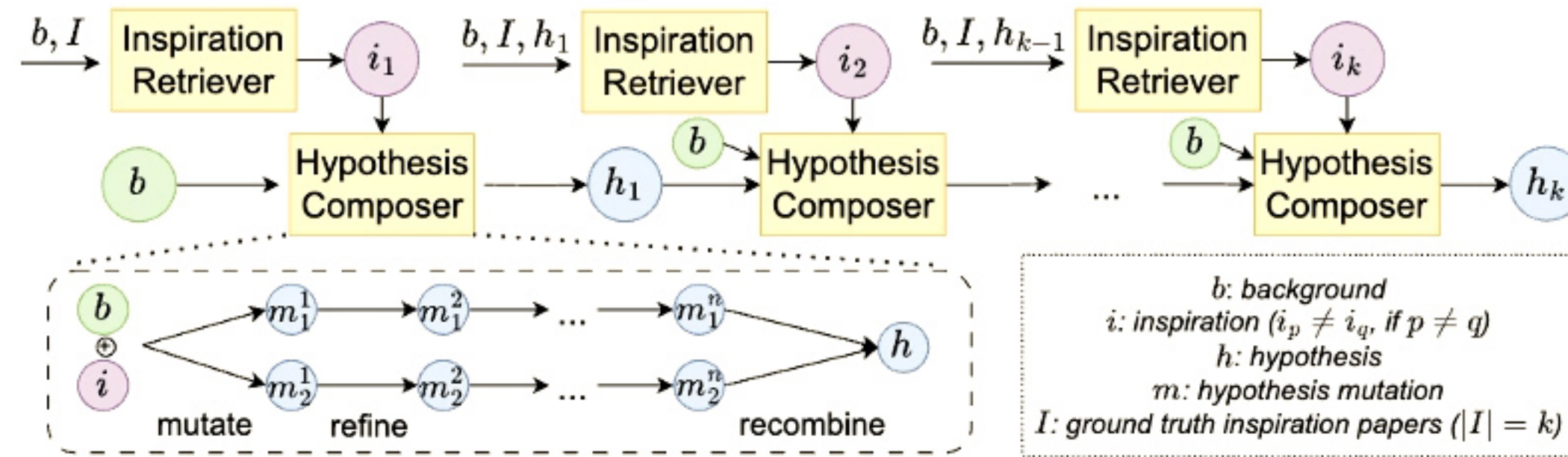


# Nano Thinfilm Self-Driving Labs





# ResearchBench: Tree-Search Based Tool Learning



- Hypothesis composition task**

| Model             | Cell         | C            | ETS          | MS           | P            | EGS          | EVS          | BL           | BS           | Law          | Math         | A            | Overall      |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Llama-3.1-8B      | 76.81        | 85.83        | 77.12        | 80.10        | 84.73        | 79.37        | 77.37        | 77.93        | 74.71        | 73.64        | 83.64        | 81.46        | 78.79        |
| Claude 3.5 Haiku  | 82.14        | 87.15        | 79.32        | 86.57        | 88.71        | 81.13        | 81.13        | 78.44        | 73.14        | 80.00        | 88.73        | 76.59        | 81.87        |
| Gemini 2.0 FH     | 80.37        | 86.67        | 84.85        | 88.36        | 88.92        | 84.93        | 84.93        | 80.12        | 78.15        | 81.86        | 82.05        | 83.68        | 83.27        |
| Claude 3.5 Sonnet | 81.25        | 91.81        | 81.15        | 87.46        | 89.68        | 83.74        | 83.74        | 80.19        | 82.57        | 79.39        | 87.64        | 83.90        | 83.73        |
| GPT-4o Mini       | 83.93        | 91.67        | 79.73        | 87.79        | 88.82        | 86.94        | 86.94        | 82.50        | 82.00        | 82.57        | 86.67        | 87.32        | 84.56        |
| Llama-3.1-70B     | 83.10        | 89.38        | 84.27        | 86.77        | 89.57        | 84.68        | 84.68        | 85.56        | 84.57        | 84.57        | 87.27        | 85.85        | 85.19        |
| Qwen Turbo        | 85.18        | 90.28        | 83.56        | 86.00        | 90.16        | 82.88        | 82.88        | 82.50        | 83.56        | 84.57        | 87.64        | 85.85        | 85.20        |
| GPT-4o            | 88.39        | 93.06        | 87.82        | 90.45        | 89.14        | 89.46        | 89.46        | 86.32        | 85.00        | 87.07        | 87.33        | 86.50        | 87.68        |
| Qwen Plus         | 88.39        | 92.36        | 88.53        | 93.03        | 90.65        | 86.58        | 86.58        | 87.30        | 84.22        | 85.56        | 88.73        | 91.22        | 88.24        |
| Gemini 2.0 Flash  | 88.34        | 91.67        | 88.20        | 91.44        | 90.43        | 87.48        | 87.48        | 89.77        | 85.80        | <b>89.43</b> | 90.91        | 89.35        | 88.38        |
| DeepSeek-V3       | <b>90.03</b> | <b>93.89</b> | <b>90.25</b> | <b>93.53</b> | <b>92.97</b> | <b>90.18</b> | <b>90.18</b> | <b>89.78</b> | <b>88.14</b> | 88.79        | <b>92.73</b> | <b>96.10</b> | <b>90.58</b> |





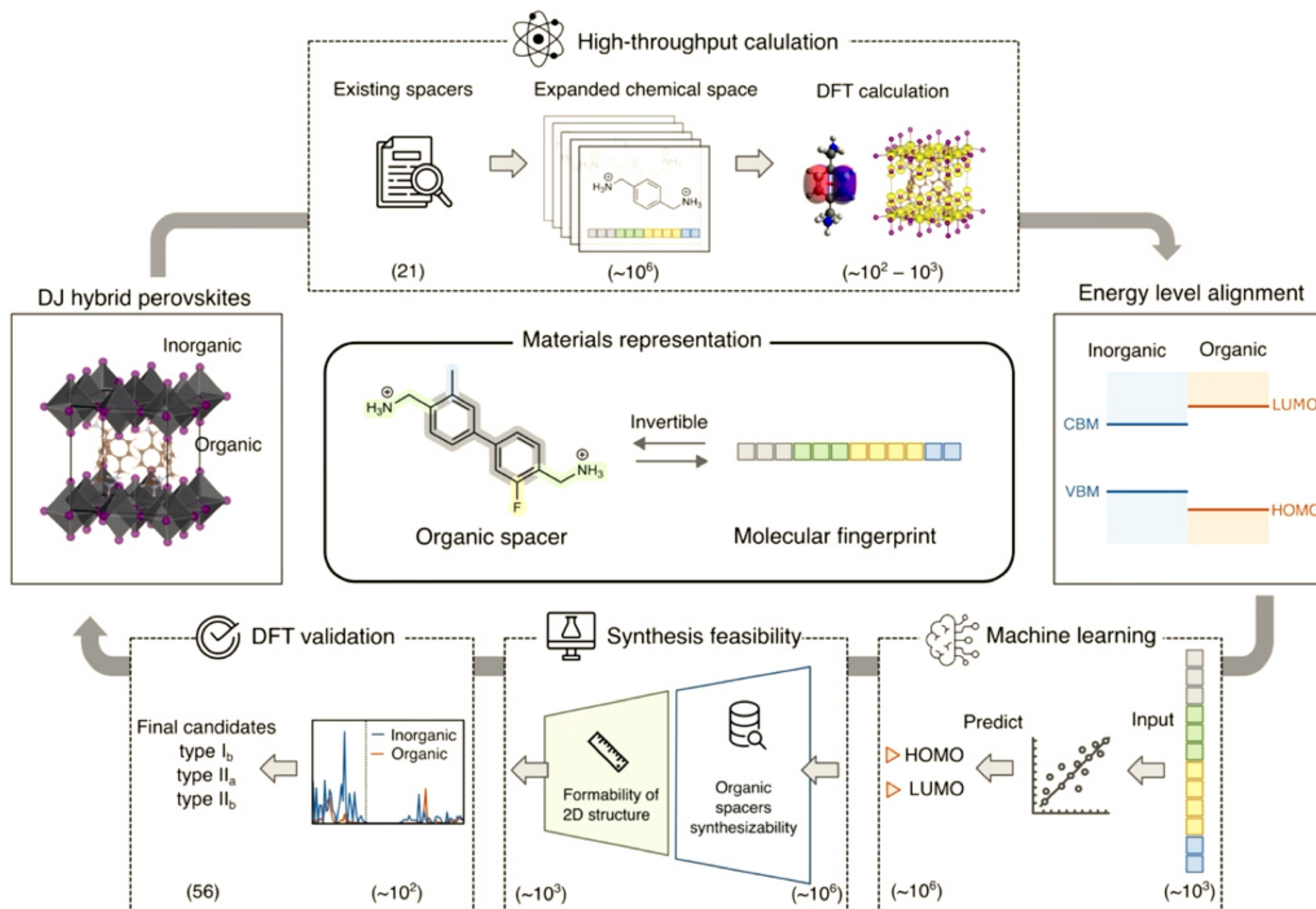
# The Road to “AI Scientists”

**Machine Learning** Accelerate **Material Inverse Design**

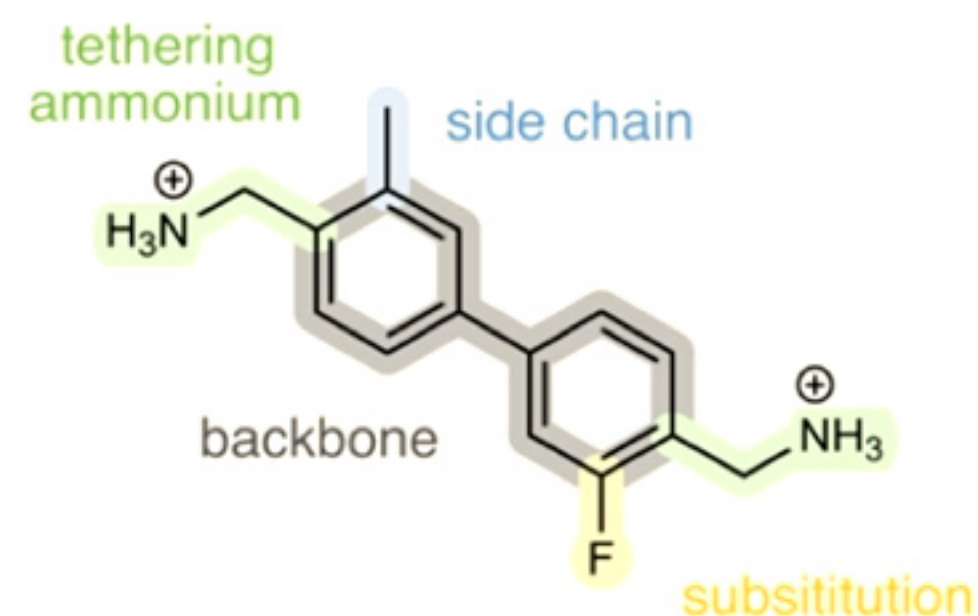




# Inverse Design of 2D Hybrid Perovskites



## (1) Fragmentation



## (2) Encoding

|  |                                                                                                                    |                  |
|--|--------------------------------------------------------------------------------------------------------------------|------------------|
|  | 2 rings,<br>100% linked,<br>100% 6-membered ring                                                                   | 2<br>1<br>1      |
|  | 2 primary ammoniums<br>total linker length: 2<br>position on backbone: 1.0 (para)                                  | 2<br>2<br>1      |
|  | 0 pyridine-type N substitution<br>1 F substitution<br>0 furan-type O substitution<br>0 pyrrole-type N substitution | 0<br>1<br>0<br>0 |
|  | 0 side chain on linker<br>1 side chain on backbone                                                                 | 0<br>1           |



**Yongxin Lyu**  
PolyU

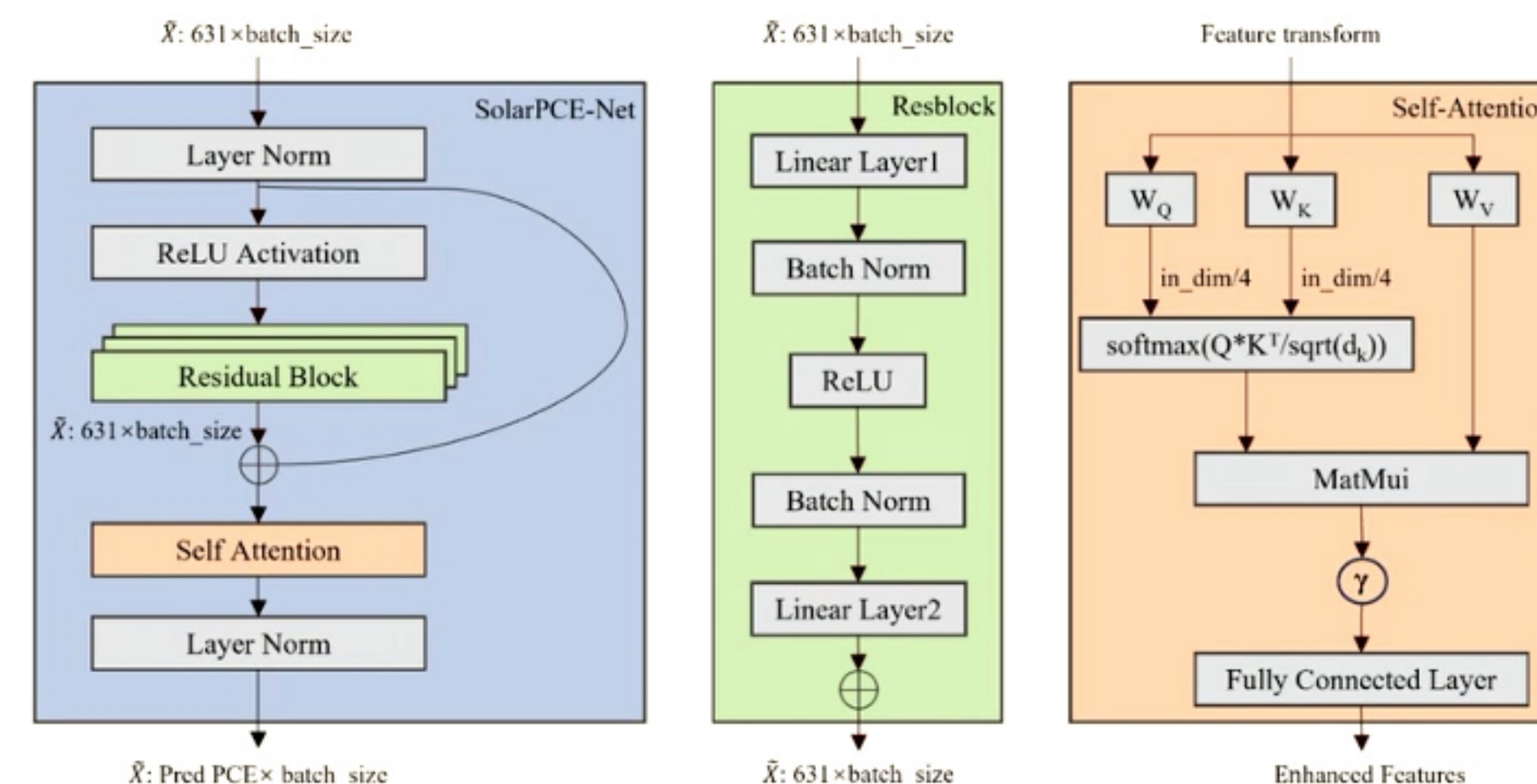
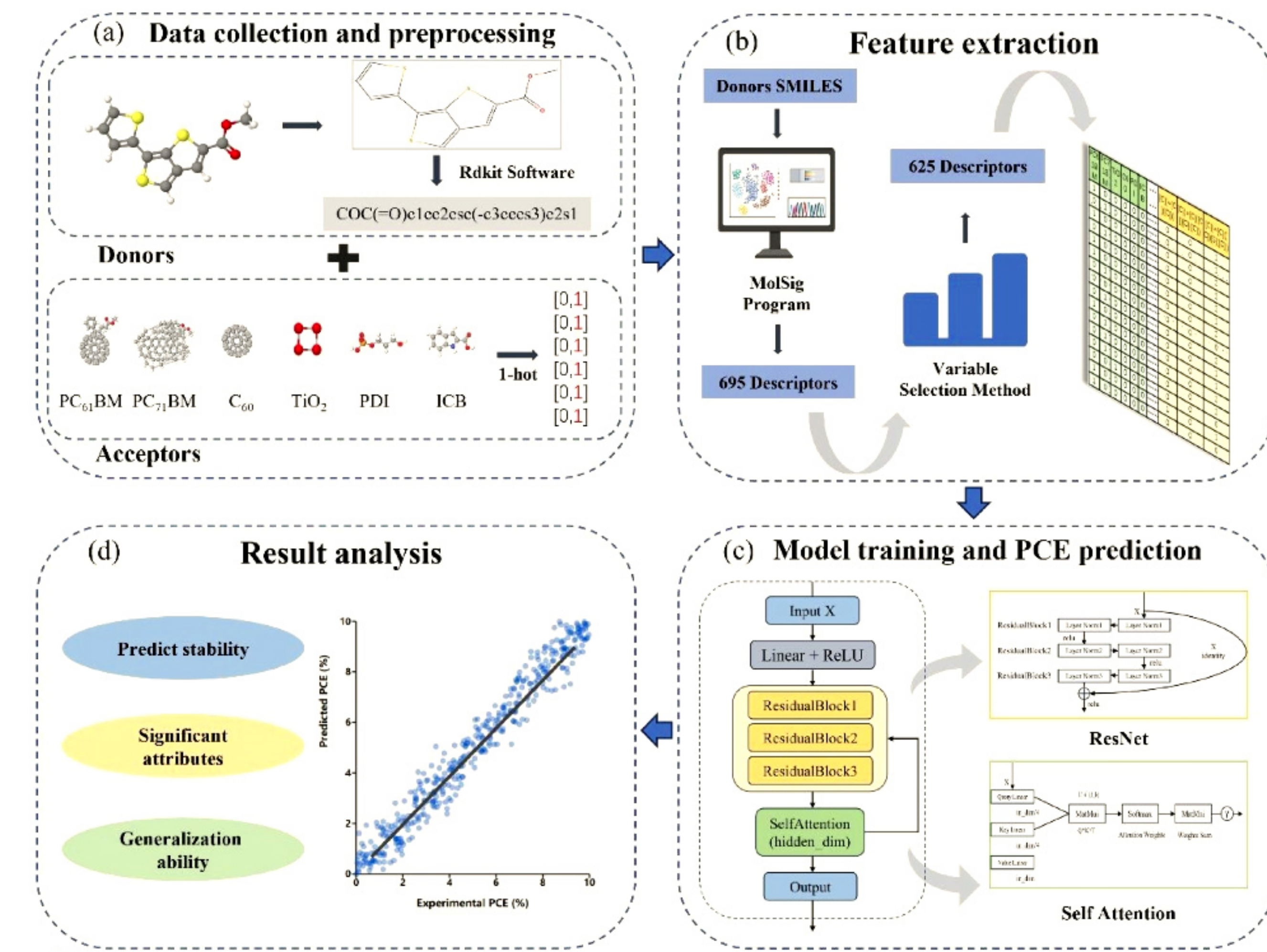


**Tom Wu**  
PolyU





# Inverse Design of Donor-Acceptor Pairs



[1] X. Liu, T. Xie, et.al Accelerating discovery of high-efficiency donor-acceptor pairs in organic photovoltaics via SolarPCE-Net guided screening, Journal of Materials Chemistry A.





# Questions?

tong@greendynamics.com.au

tong.xie@unsw.edu.au

Intelligent **MATERIALS**, Invented **TODAY**

Homepage: <https://unswai4s.notion.site/>

Thanks  
Support



UNSW  
SYDNEY

