香港中文大學
The Chinese University of Hong Kong

Large Language Models for Scientific Discovery in Molecular Property Prediction

Jingyi Liu

Nat. Mach. Intell., 2025, 7, 437

Outline

1. Background
2. Results
3. Discussion
4. Conclusion

Research Background

Research Background

Scientific discovery

Building on existing knowledge to analyze experimental data, recognize data patterns and formulate well-reasoned hypotheses

Prior knowledge understanding

Reasoning abilities

LLM (Large Language Models)

Research Background

- Molecular property prediction

- LLMs possess extensive prior knowledge of molecular property prediction tasks.
  E.g. predict molecular weight and number of aromatic rings as key factors in solubility

- LLMs can understand formal scientific languages, such as the Simplified Molecular Input Line Entry System (SMILES).

Research Background

- Molecular property prediction

  - Can LLMs leverage their prior knowledge and reasoning abilities to facilitate scientific discovery?
  - Can LLMs be effectively used to help predicting the properties of molecules?

- **LLM4SD (LLMs for Scientific Discovery)**, functioning by ：

  1. **synthesizing knowledge from existing literature。**
  2. **inferring knowledge by observing experimental data.**

Research Background

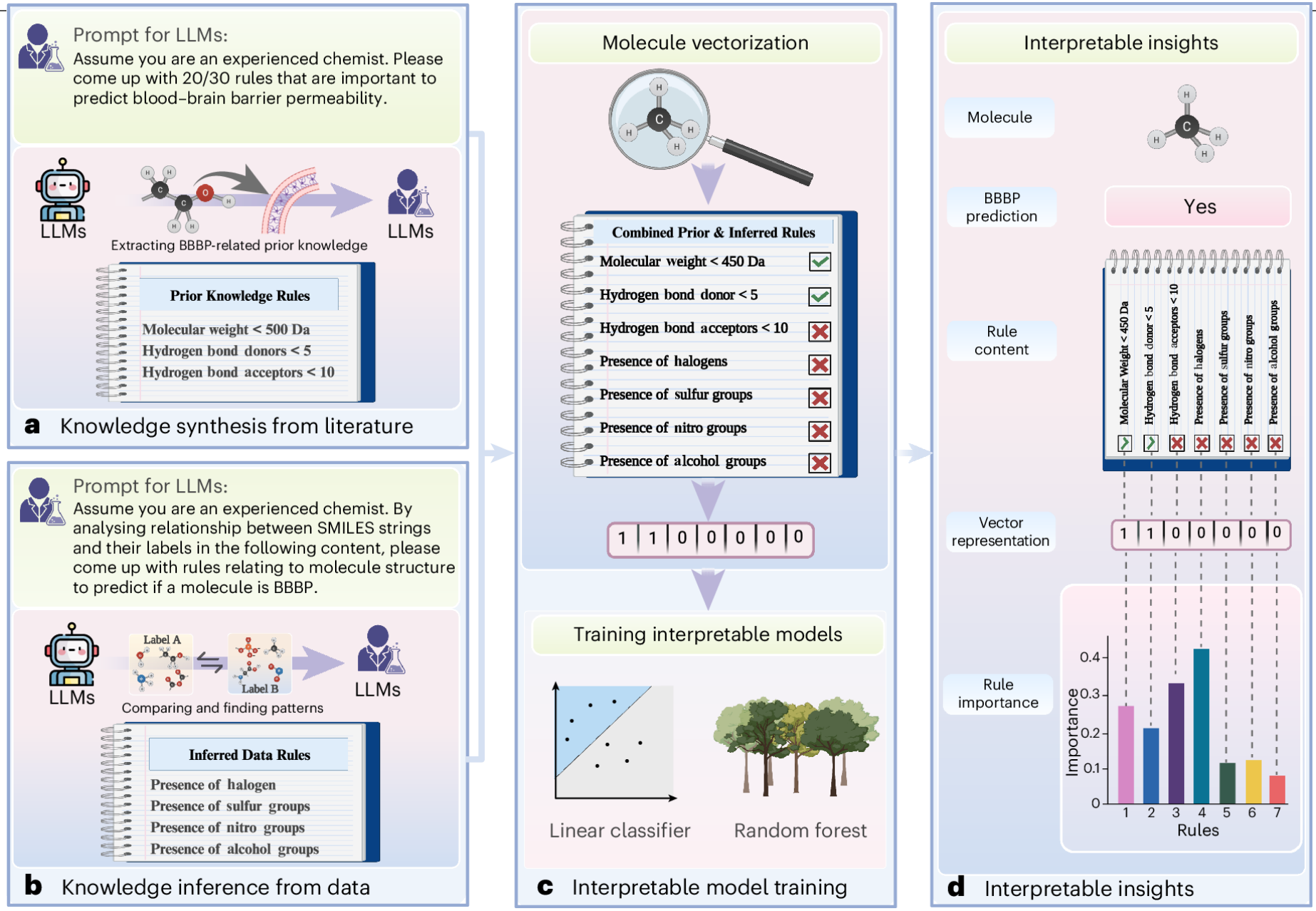
- **LLM4SD (LLMs for Scientific Discovery)**
  - Retrieves known rules to predict molecular properties based on its pretrained literature
    - molecules with molecular weight under 500 Da being more likely to pass the. blood–brain barrier (BBB).
  - Identifies patterns from experimental data using its understanding of SMILES notation and chemistry knowledge
    - molecules containing halogens are more likely to pass the BBB. These rules are then used to create interpretable feature vectors for each molecule.
  - Achieves the current state of the art on molecular property prediction across 58 benchmark tasks from the MoleculeNet, spanning four domains: physiology, biophysics, physical chemistry and quantum mechanics.

Results

Results

- LLM4SD pipeline
- Experimental results
- Validation of established rules

LLM4SD pipeline



a. Knowledge synthesis from the scientific literature
b. Knowledge inference from data
c. Model training
d. Interpretable insights.

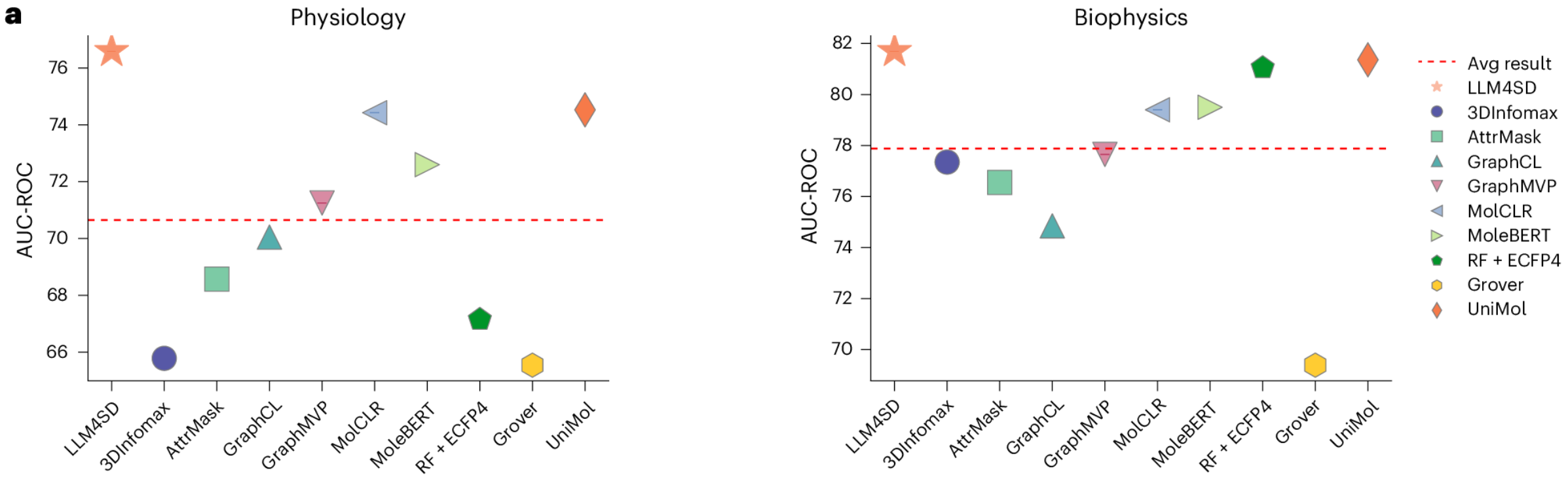
➢ Experimental Results

- 4 domains: physiology, biophysics, quantum mechanics, physical chemistry
  - Overall performance
  - Study of key components

➢ Experimental Results: overall performance

❖ Overall performance:

LLM4SD demonstrated superior efficacy and performance compared with nine other specialized, state-of-the-art supervised GNNs: AttrMask, GraphCL, MolCLR, 3DInfomax, GraphMVP, MoleBERT, Grover and UniMol2

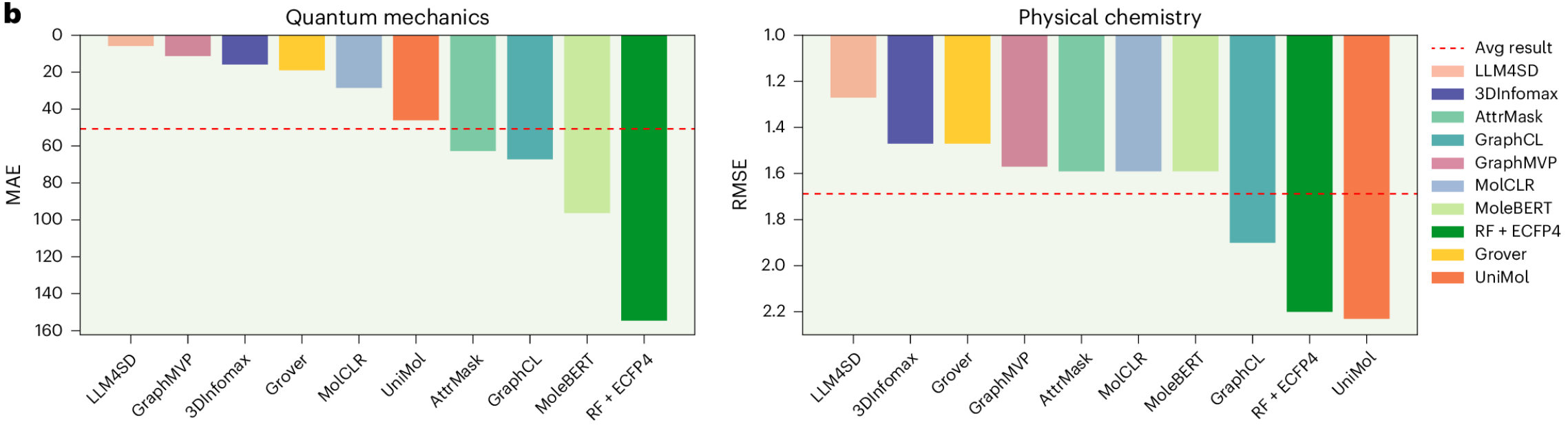


*AUC-ROC: Area under the Receiver Operating Characteristic Curve. It measures a model's ability to distinguish between classes, regardless of threshold. It's standard for classification problems

➢ **Experimental Results:** overall performance

❖ Overall performance:

LLM4SD demonstrated superior efficacy and performance compared with nine other specialized, state-of-the-art supervised GNNs: AttrMask, GraphCL, MolCLR, 3DInfomax, GraphMVP, MoleBERT, Grover and UniMol2



$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

➢ **Experimental Results:** overall performance

❖ Compared to GNNs, LLM4SD:

1. Leverages prior knowledge accumulated from previous scientific literature.
2. Provides interpretability for generating clear scientific hypotheses.

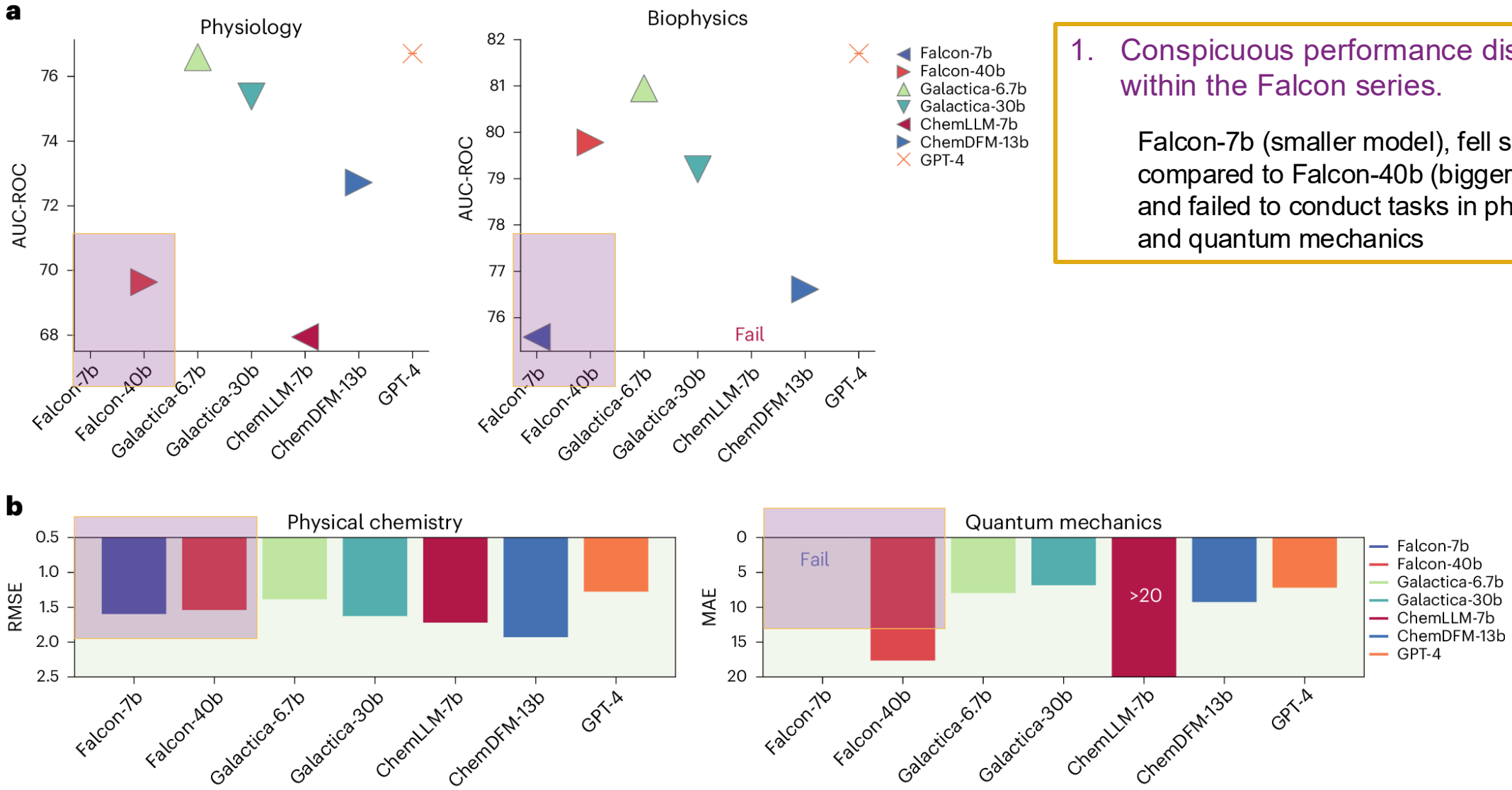
	LLM4SD	GNNs
Knowledge Integration	Inherently embeds scientific knowledge, avoiding additional intervention beyond natural language interaction	Requires explicit careful curation and integration of domain-specific features or hard-coded knowledges
Interpretability	More interpretable through natural language reasoning and explanation capabilities	Encoding molecules into embeddings leads to a lack of interpretability.

➢ **Experimental Results:** study of key components

- The influence of scale and pretraining datasets on its performance are studied.
- Contributions of knowledge synthesis and inference are accessed.
- Foundational LLM backbones are evaluated:
  - General LLMs: GPT-4, Falcon-7b, Falcon-40b
  - Domain-specific LLMs: Galactica-6.7b, Galactica-30b, ChemLLM-7b, and ChemDFM-13b

➢ **Experimental Results:** study of key components

❖ Effect of scale and pretraining datasets

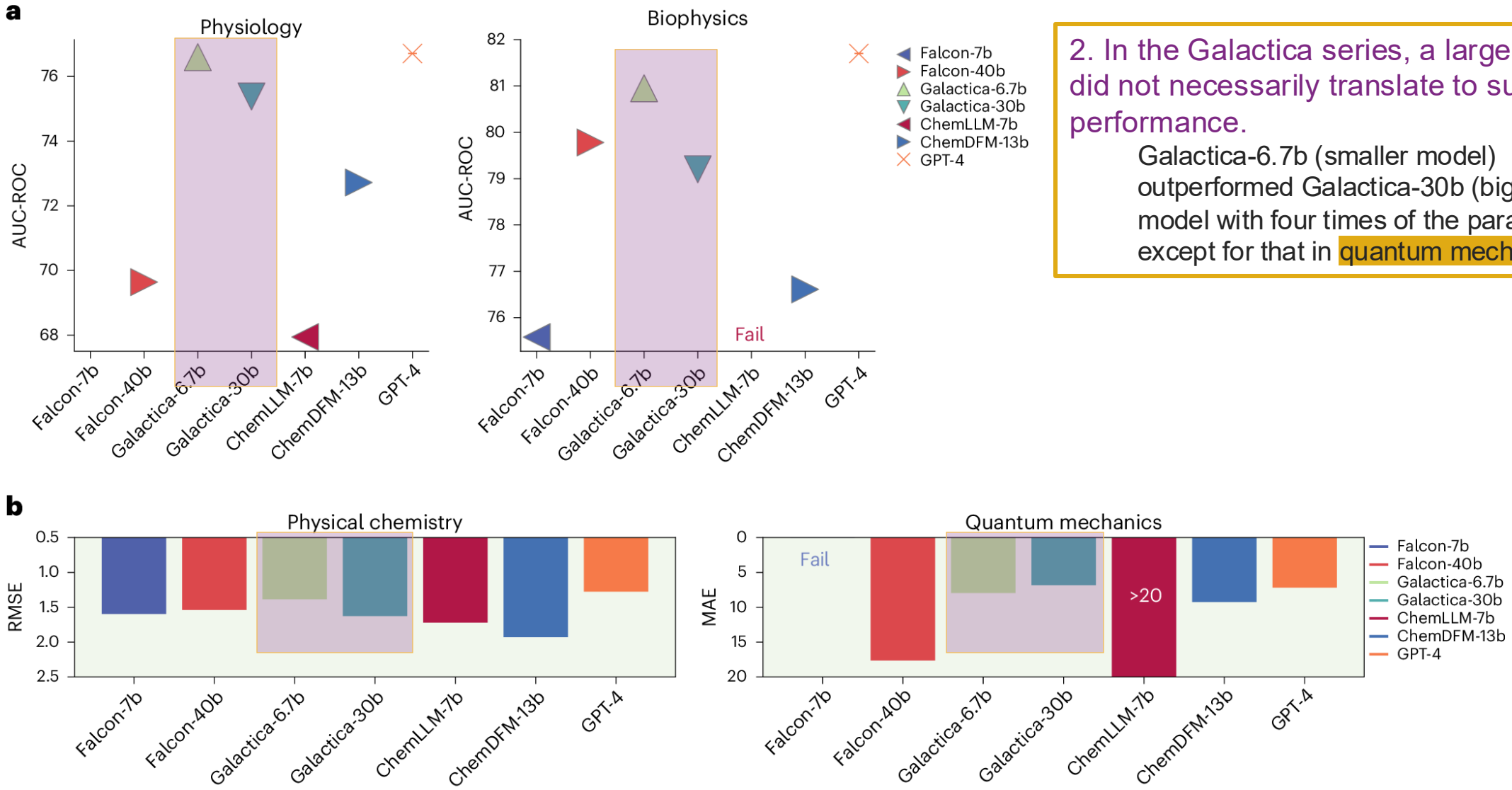


1. Conspicuous performance disparities within the Falcon series.

   Falcon-7b (smaller model), fell short compared to Falcon-40b (bigger model) and failed to conduct tasks in physiology and quantum mechanics

➢ **Experimental Results:** study of key components

❖ Effect of scale and pretraining datasets

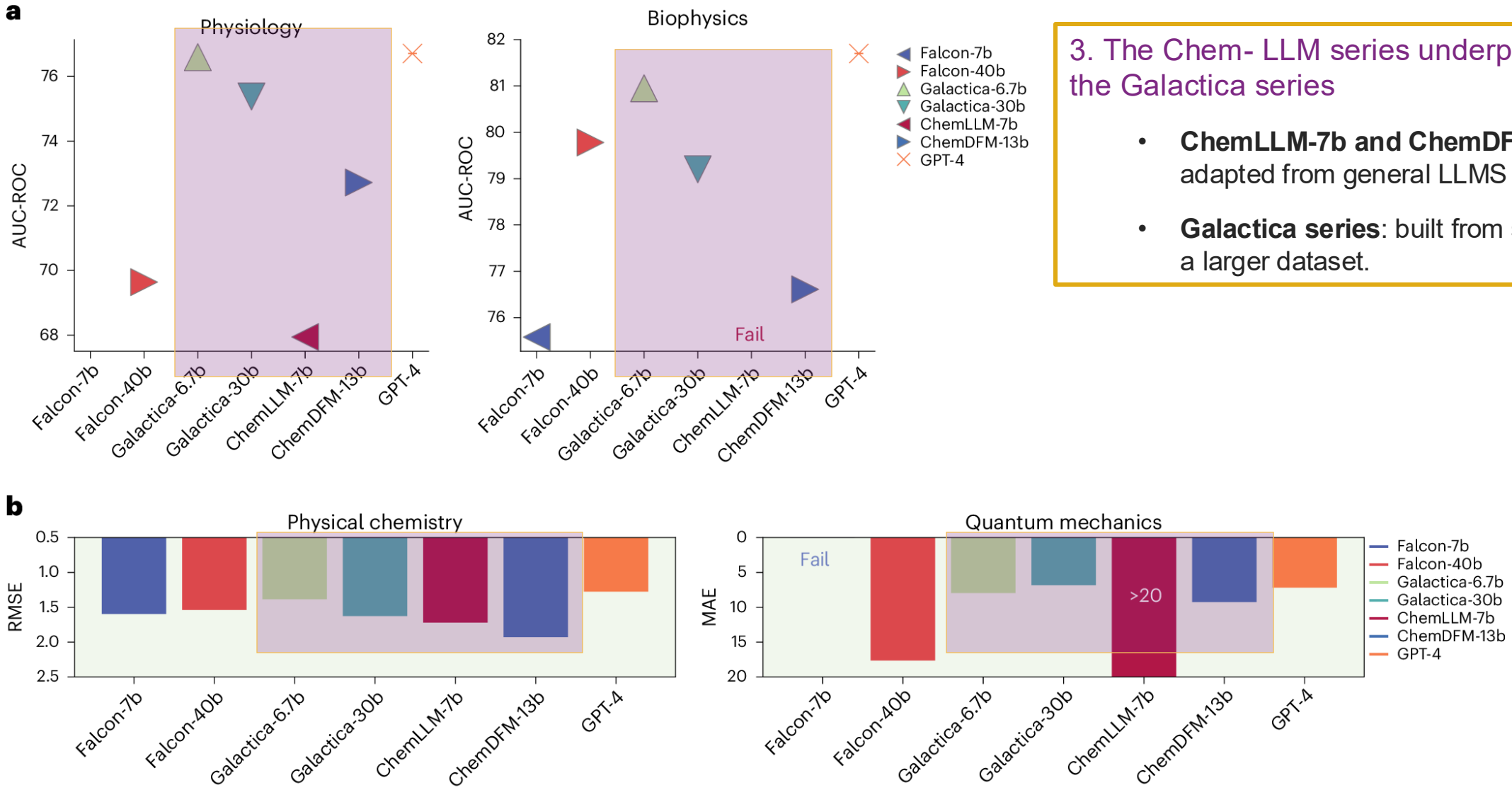


2. In the Galactica series, a larger model did not necessarily translate to superior performance.

Galactica-6.7b (smaller model) outperformed Galactica-30b (bigger model with four times of the parameters) except for that in quantum mechanics.

➢ **Experimental Results:** study of key components

❖ Effect of scale and pretraining datasets



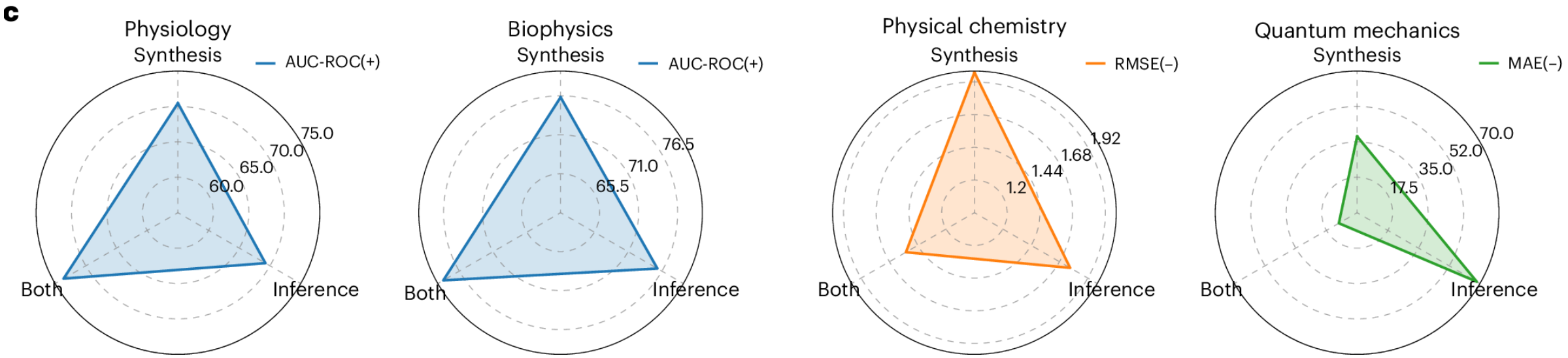
3. The Chem- LLM series underperformed the Galactica series

- **ChemLLM-7b and ChemDFM-13b**: adapted from general LLMS
- **Galactica series**: built from scratch with a larger dataset.

➢ **Experimental Results:** study of key components

❖ Contributions of knowledge synthesis and interference

From literature **From experimental data**



The combination of synthesis and inference features consistently outperformed individual methods.

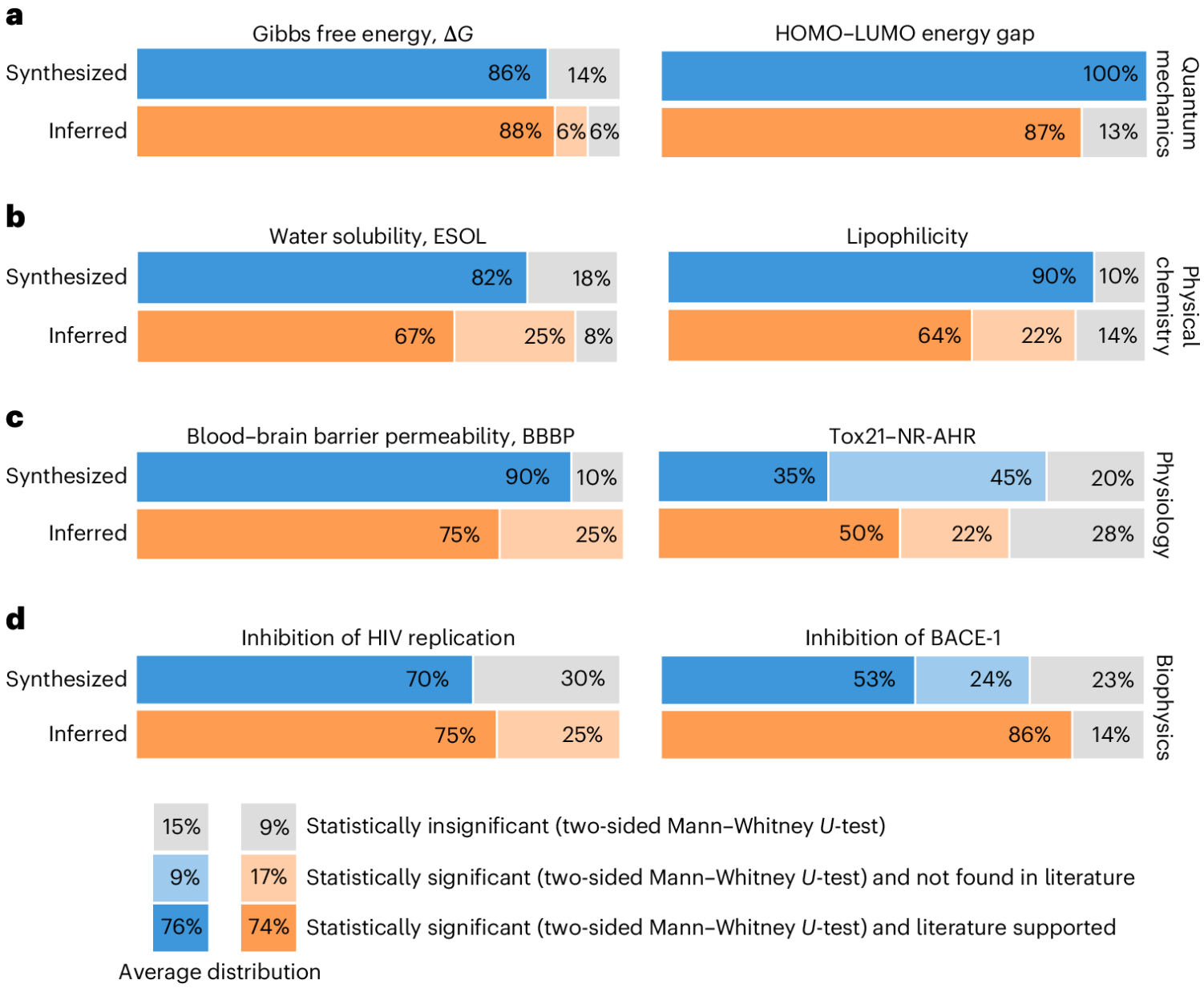
➢ **Experimental Results:** validation of established rules

The rules generated by Galactica-6.7b were validated by:

1. statistical tests to confirm their association with the target molecular attribute:
    - Classification tasks: the Mann–Whitney U-test
      Evaluates the statistical relevance of the rule's ability to distinguish classes
    - Regression tasks: the linear correlation t-test
      Reflects whether the rule contributes to regression prediction

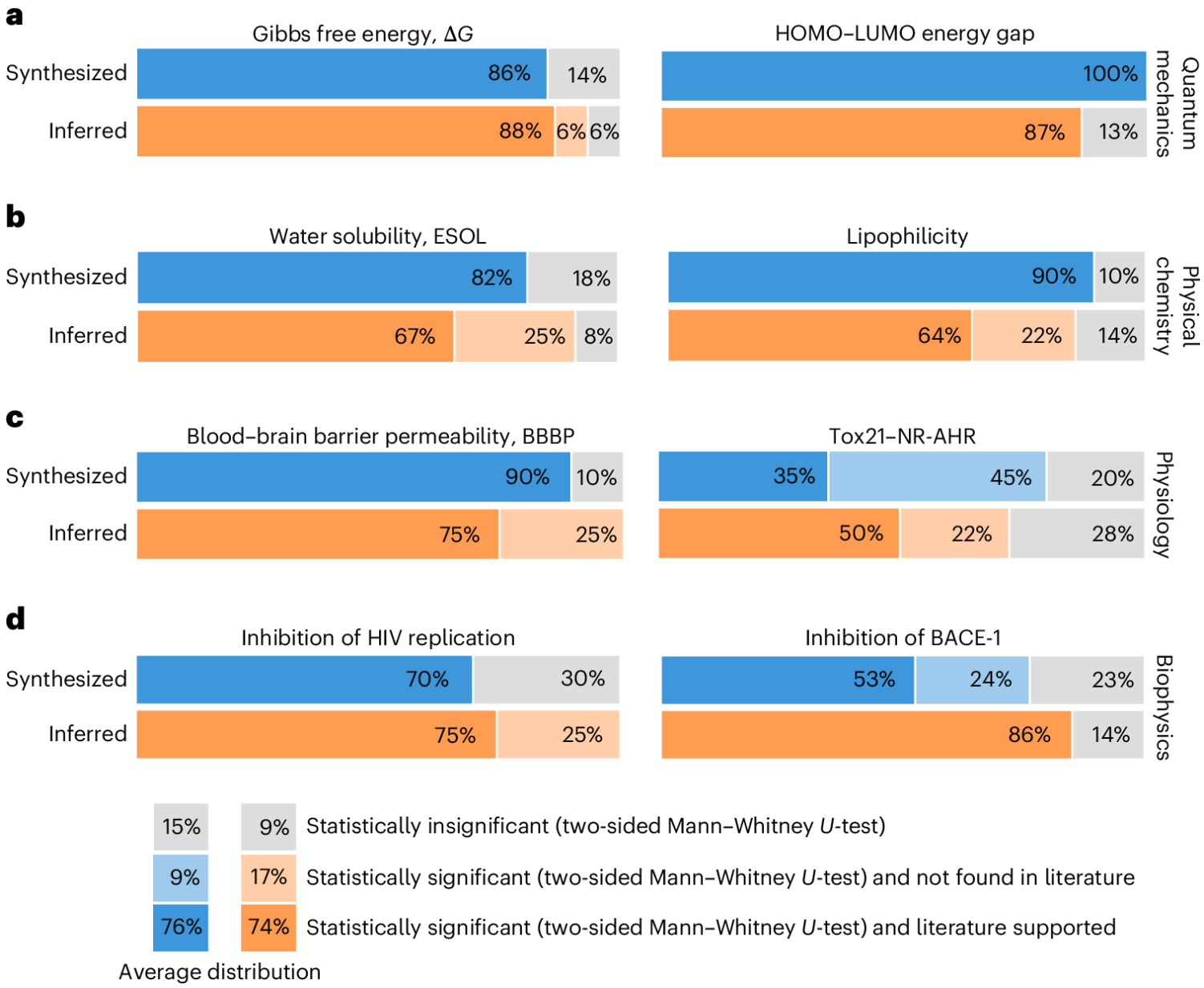
2. a literature review to validate its existence in previous research.

➢ **Experimental Results:** validation of established rules



1. Most of the synthesized rules are readily available in existing scholarly works.
2. Without analysing the data, LLMs tend to aggregate and summarize existing knowledge.

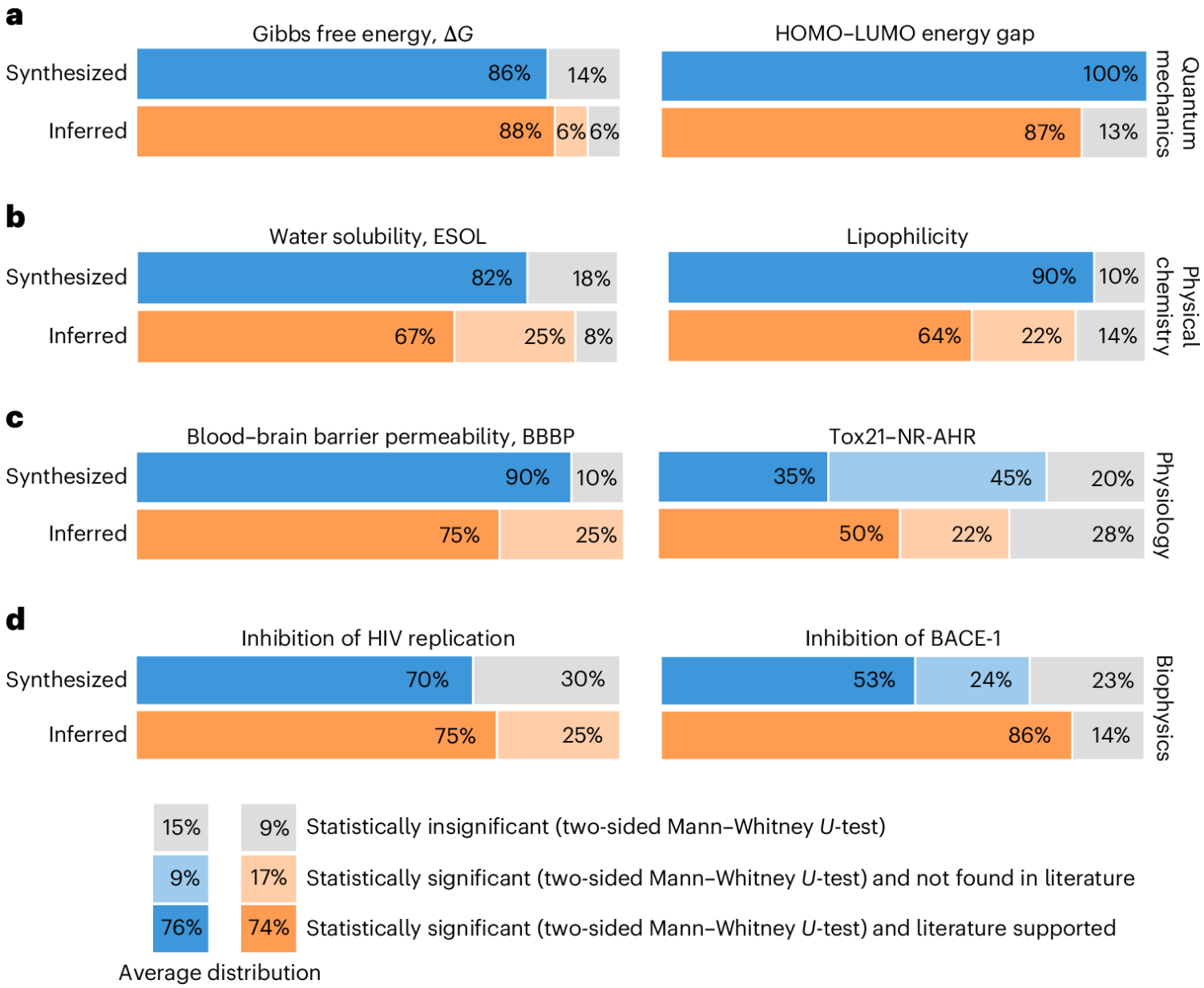
➢ **Experimental Results:** validation of established rules



3. An average of 91.3% of the inferred rules were statistically significant, higher than synthesized, among which an average of 74% rules were already documented in existing scientific literature.

4. six out of eight tasks have statistically significant rules that could not be identified in the existing literature, reflecting a genuine capability of LLM4SD to derive meaningful rules from data.

➢ **Experimental Results:** validation of established rules



5. LLM4SD was able to infer second-order-rules.

The carbonyl functional group and fragment rings are predicted as key determinants of a molecule's BBBP, which is not identified in literatures.

However, they can influence a molecule's cross-sectional area, furtherly affects its orientation in lipid–water interfaces—factors vital for membrane partitioning and permeation.

Discussion

➢ Discussion

1. Despite focused on molecular property prediction in this study, LLM4SD can be extended to complex tasks like protein and gene sequence analysis.
   - Biological sequences are much longer and more complex than SMILES, posing challenges for LLMs due to limited context handling and domain-specific knowledge requirements.

2. Improvements could include:
   - Pretraining on large biological datasets
   - Using retrieval-augmented generation with databases like UniProt and GenBank
   - Developing better tokenization for biological sequences

Conclusion

➢ Conclusion

1. LLM4SD is proposed as a framework designed to harness LLMs for driving scientific discovery in molecular property prediction by synthesizing knowledge from literature and inferring knowledge from scientific data.

2. LLM4SD has outperformed GNNs in molecular property prediction in physiology, biophysics, quantum mechanics, physical chemistry.

3. Despite their smaller scales, domain-specific LLMs such as the Galactica models outperformed the general LLMs like Falcon series, underscoring the pivotal role of proper domain-specific pretraining.

4. LLM4SD not only validates well-established scientific principles but also uncovers potentially new rules, enhancing both the quality and trustworthiness of the research output.

Questions? Comments?

Thank You