清華大學 药学院
School of Pharmaceutical Sciences, Tsinghua University

Improving the accuracy and generalizability of molecular property regression models with a substructure-substitution-rule-informed framework

Boxue Tian
Tsinghua University (Beijing, China)
boxuetian@mail.tsinghua.edu.cn
https://tianlab-tsinghua.cn/

The real problem

How to reduce the dependence of AI models on data?

Chemical space is vast, while most molecules are not drugs.

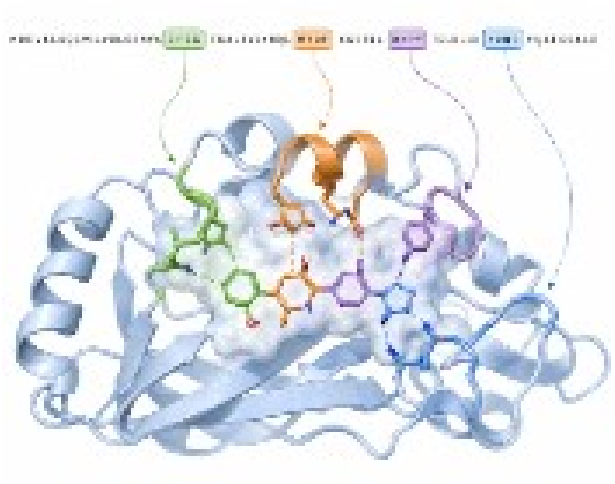
This leads to:

i) The process of collecting data to support AI in biomedical research is inefficient.

ii) The current data volume and data quality are insufficient to support high accuracy.

Outline

$$\sigma_r \leq 3e$$



AI Theory

Mathematically, the "Error Bound Theorem" is proved: the quality of substructure substituition rules determines the upper bound of the model's prediction accuracy.

Why we study “rules”

Drug discovery with PWRules

Large/high-quality rules are automatically extracted from protein sequences to serve as the knowledge base for pharmaceutical science

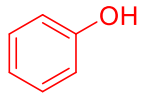
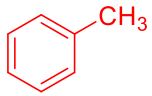
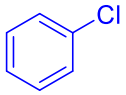
How to extract rules and how to use them?

1 "Error Bound Theorem"

--why we study rules

QSAR

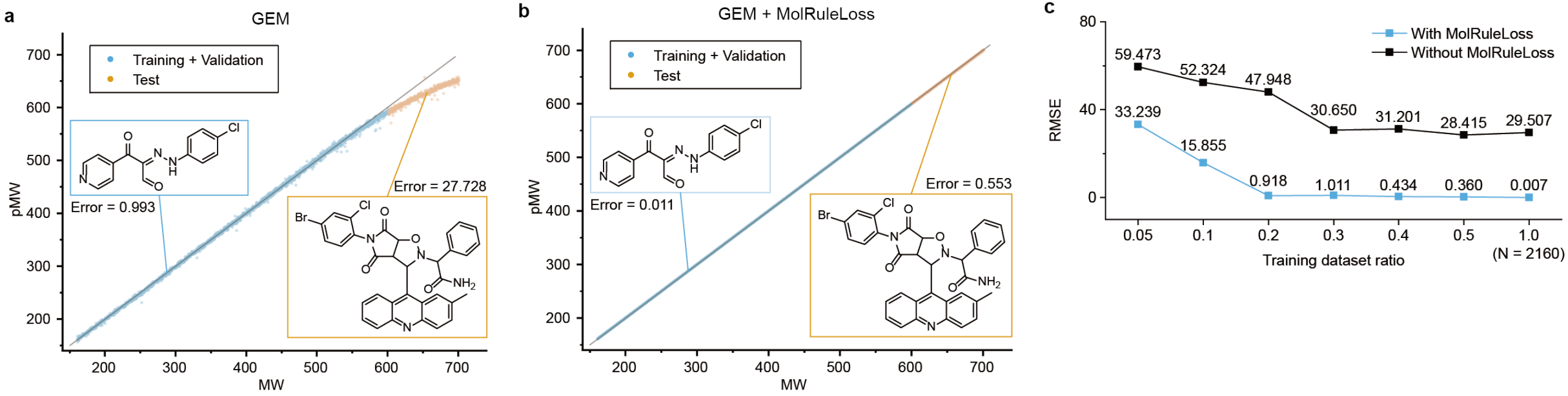
$$Y = f(X)$$

Property	Structure
Y1	
Y2	
…	…
?	

Unexpected finding: the molecular weight experiment

1) State-of-the-art molecular representation learning models fail to learn molecular weight.

2) We developed MolRuleLoss, a substructure-substitution-rule-informed framework that improves the accuracy and generalizability of multiple molecular property regression models (MPRMs) such as GEM and UniMol for diverse molecular property prediction tasks.

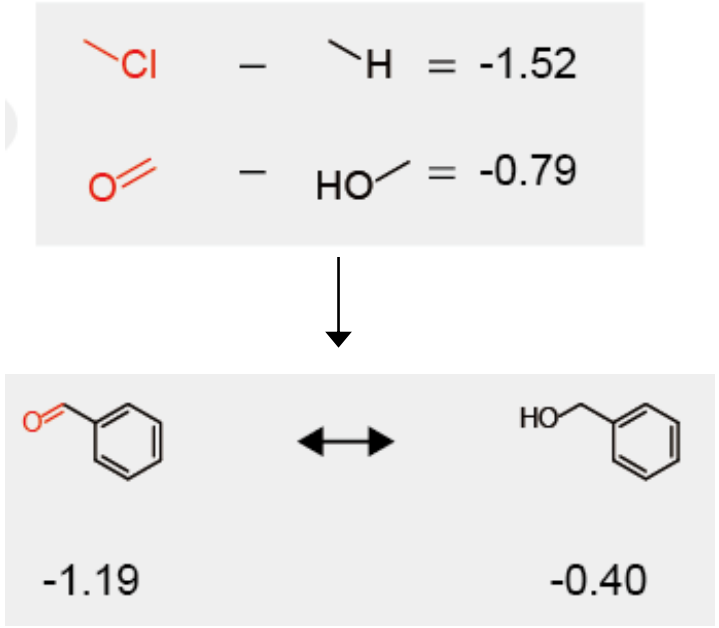


X. Fan … B.Tian* *Chemical Science* **2026**

MolRuleLoss

- We inject substructure substitution rules (SSRs) into the model, and the quality of these rules (SSRs) is determined by their standard deviation.

Solubility rules

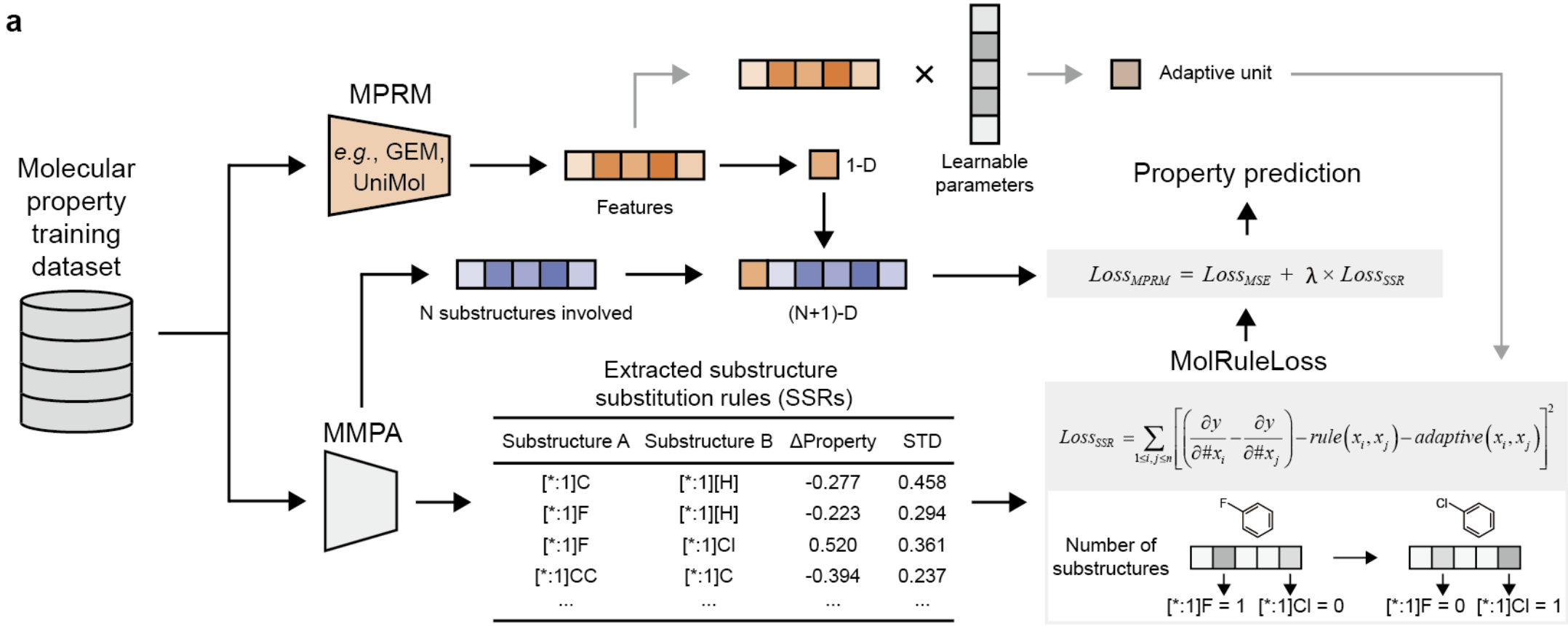


Extracted substructure substitution rules (SSRs)

Substructure A	Substructure B	ΔProperty	STD
[*:1]C	[*:1][H]	-0.277	0.458
[*:1]F	[*:1][H]	-0.223	0.294
[*:1]F	[*:1]Cl	0.520	0.361
[*:1]CC	[*:1]C	-0.394	0.237
...	...	...	...

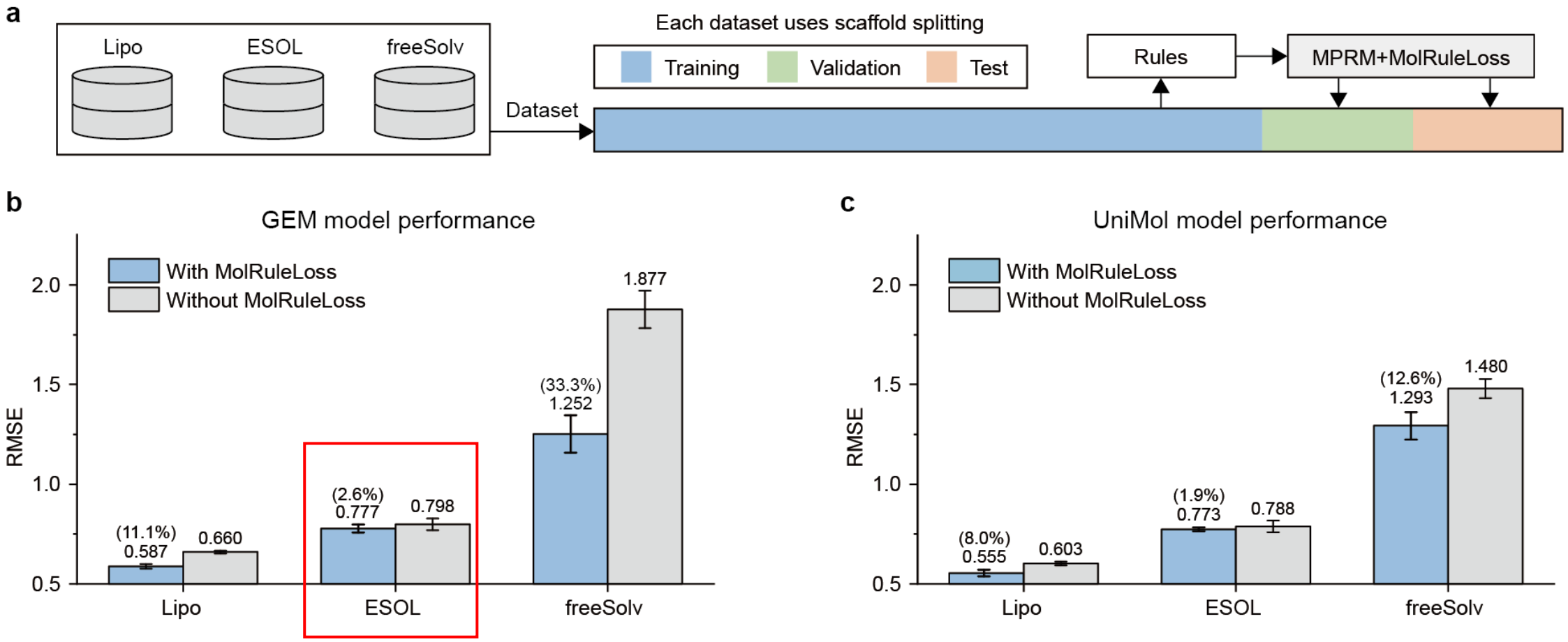
X. Fan … B.Tian* *Chemical Science* **2026**

MolRuleLoss



X. Fan … B.Tian* *Chemical Science* **2026**

MolRuleLoss improves models' accuracy

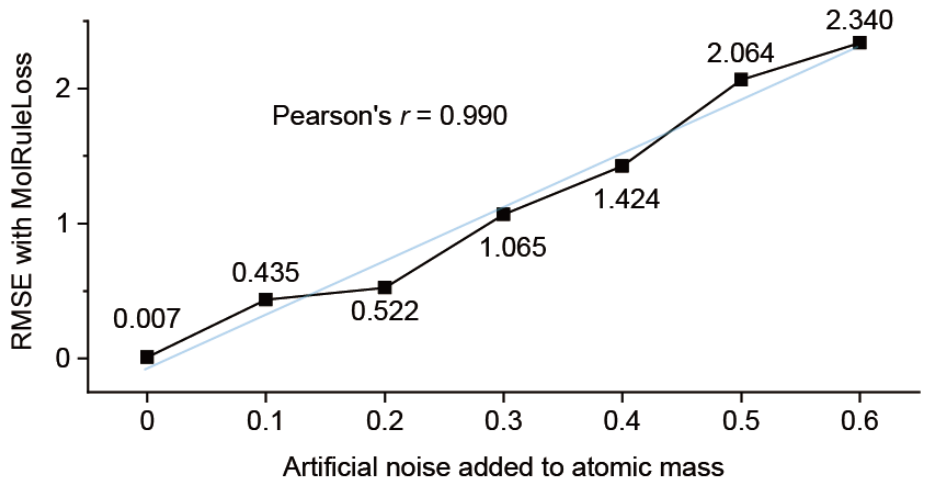


X. Fan … B.Tian* *Chemical Science* **2026**

“Error Bound Theorem”

The upper bound of the variation for property change of SSRs are positively correlated with a model’ s error

$$\sigma_r \leq 3e$$



$$\sigma_r = |P(r(c)) - P(c)| = |P(r(c)) - F(r(c)) + F(r(c)) - F(c) + F(c) - P(c)|$$

$$\leq |P(r(c)) - F(r(c))| + |F(r(c)) - F(c)| + |F(c) - P(c)|$$

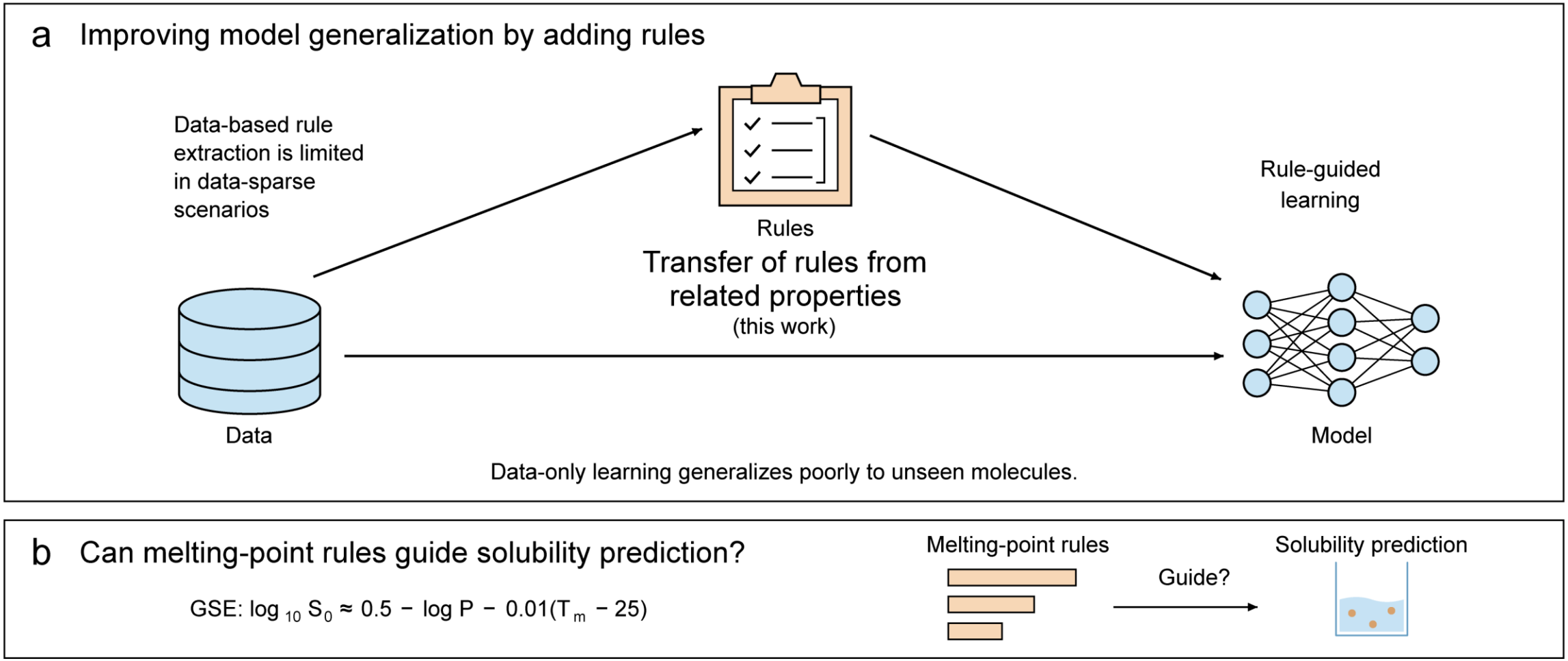
$$\leq e + e + e$$

$$= 3e$$

X. Fan … B.Tian* *Chemical Science* **2026**

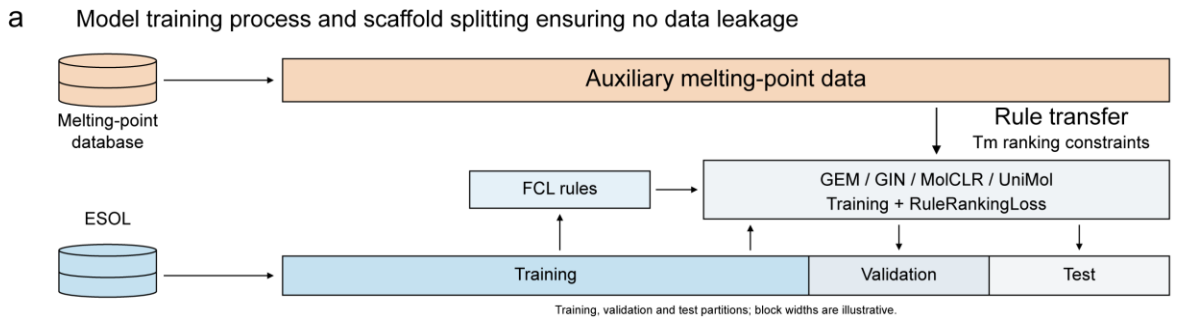
Proof by Dr. Weihao Li

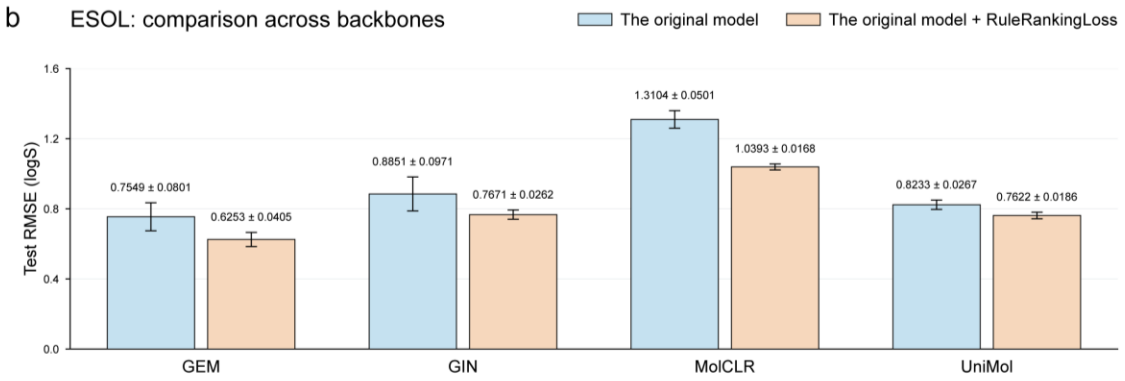
RuleRankingLoss for solubility prediction

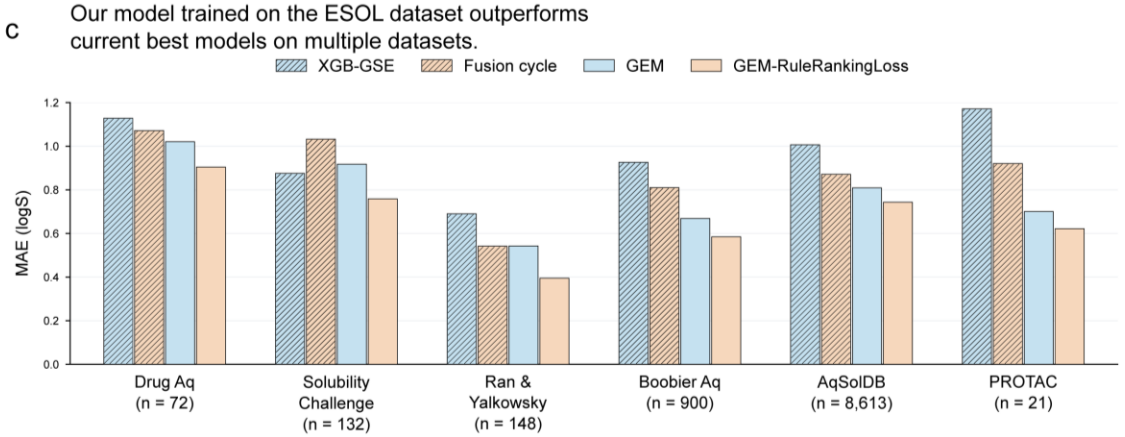


Z. Yang … B.Tian* *unpublished data*

RuleRankingLoss for solubility prediction

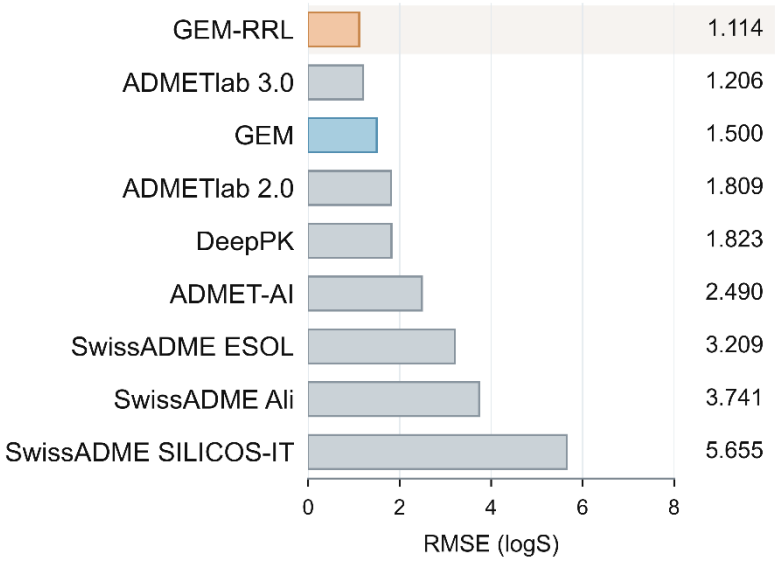






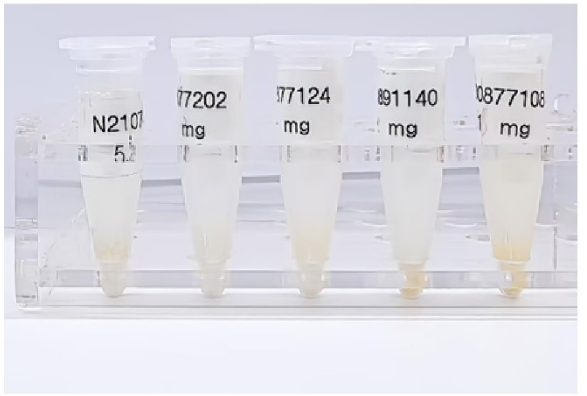
Z. Yang … B.Tian* *unpublished data*

PROTAC solubility prediction



d Wet-lab experiments





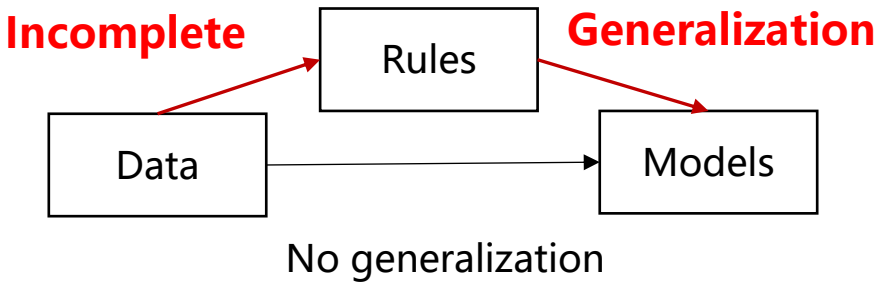


Z. Yang … B.Tian* *unpublished data*

AI theory

Trends : "big data, high-quality data" + "big rules, high-quality rules"

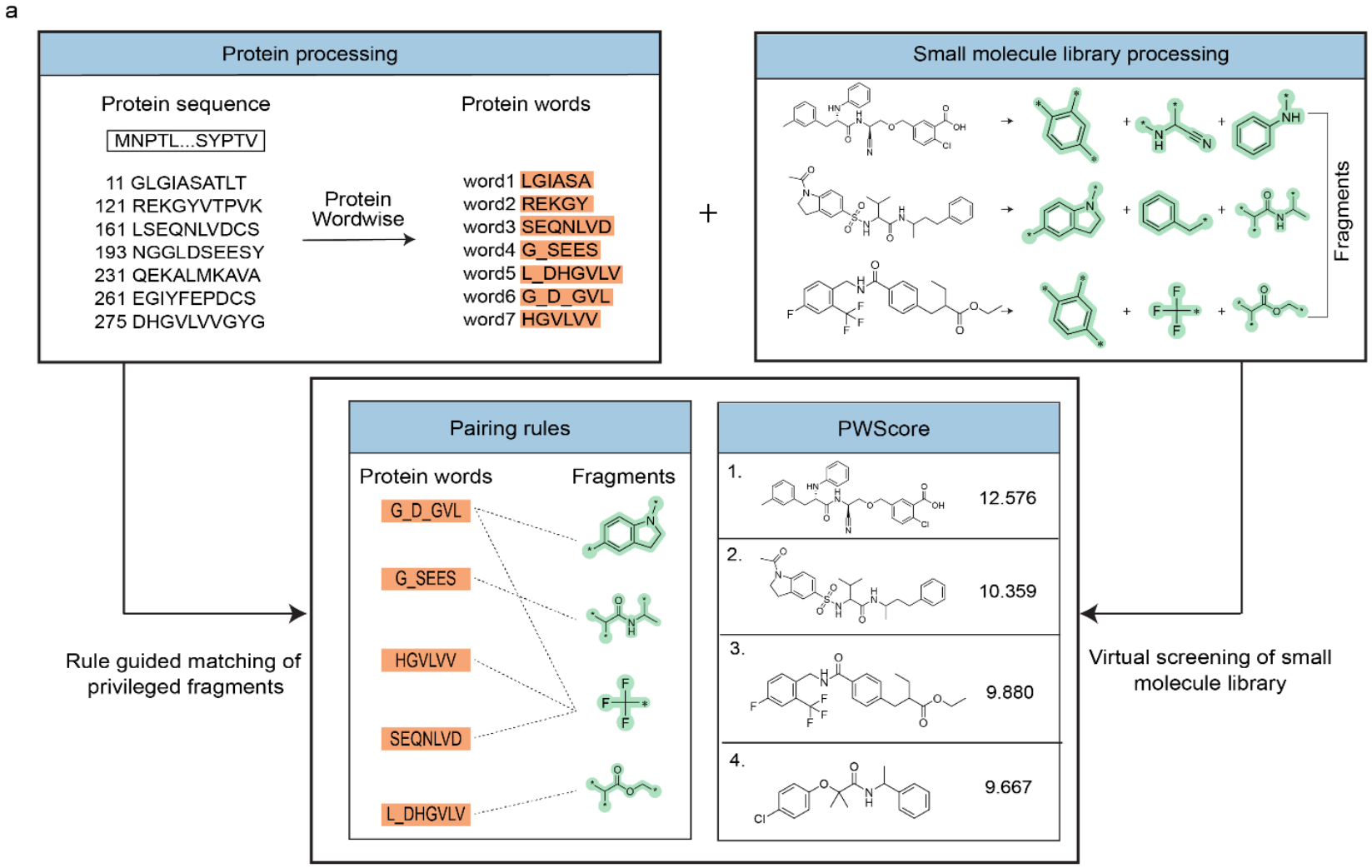
New questions:
How should protein substructure rules be defined?
How can we obtain a large number of such rules?



2 PWRules for drug discovery

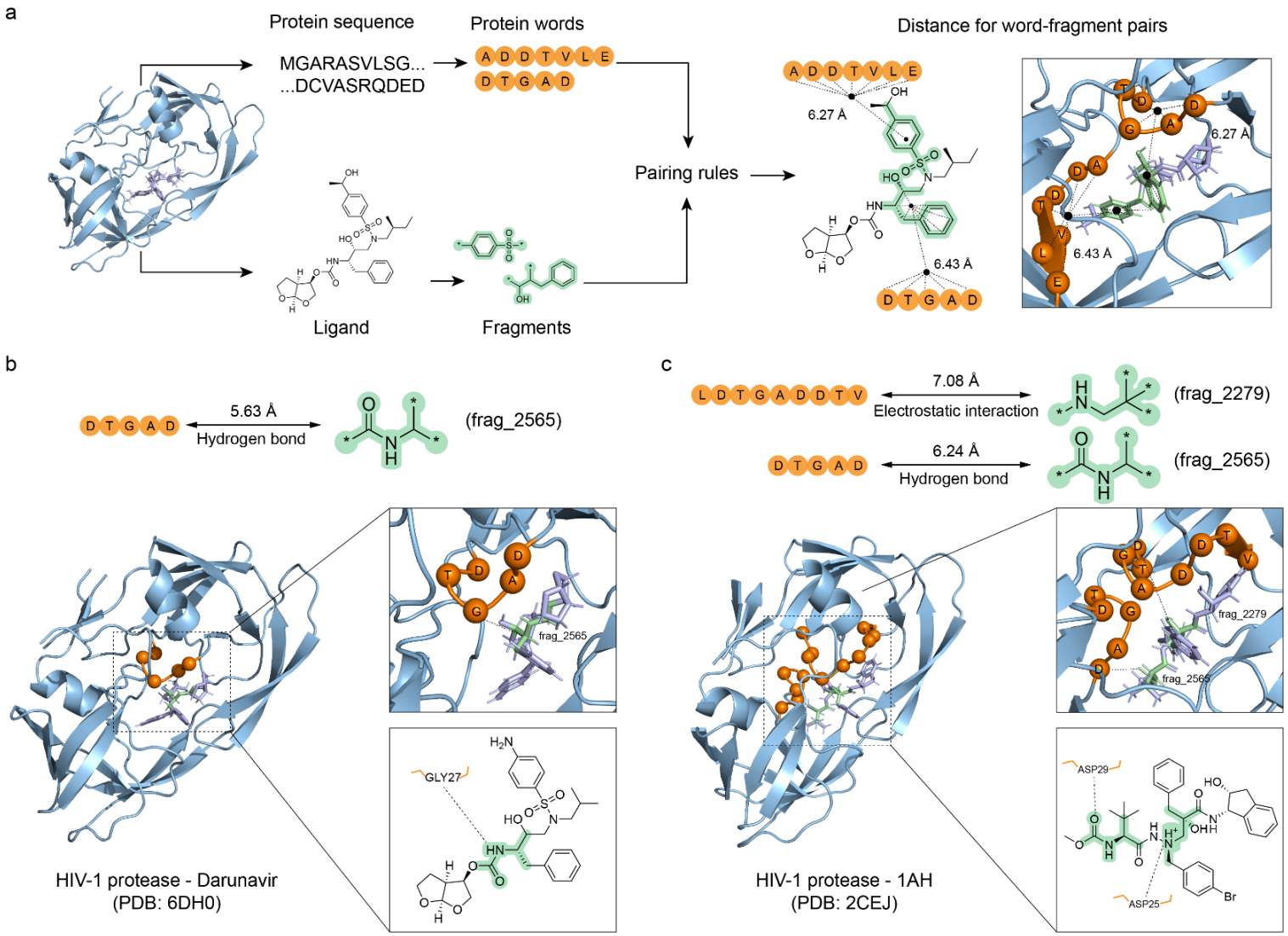
--how we define substructures in proteins and how to define rules for drug discovery

Protein Word-fragment rules



J. Chen … B.Tian* *Chemical Science* 2026

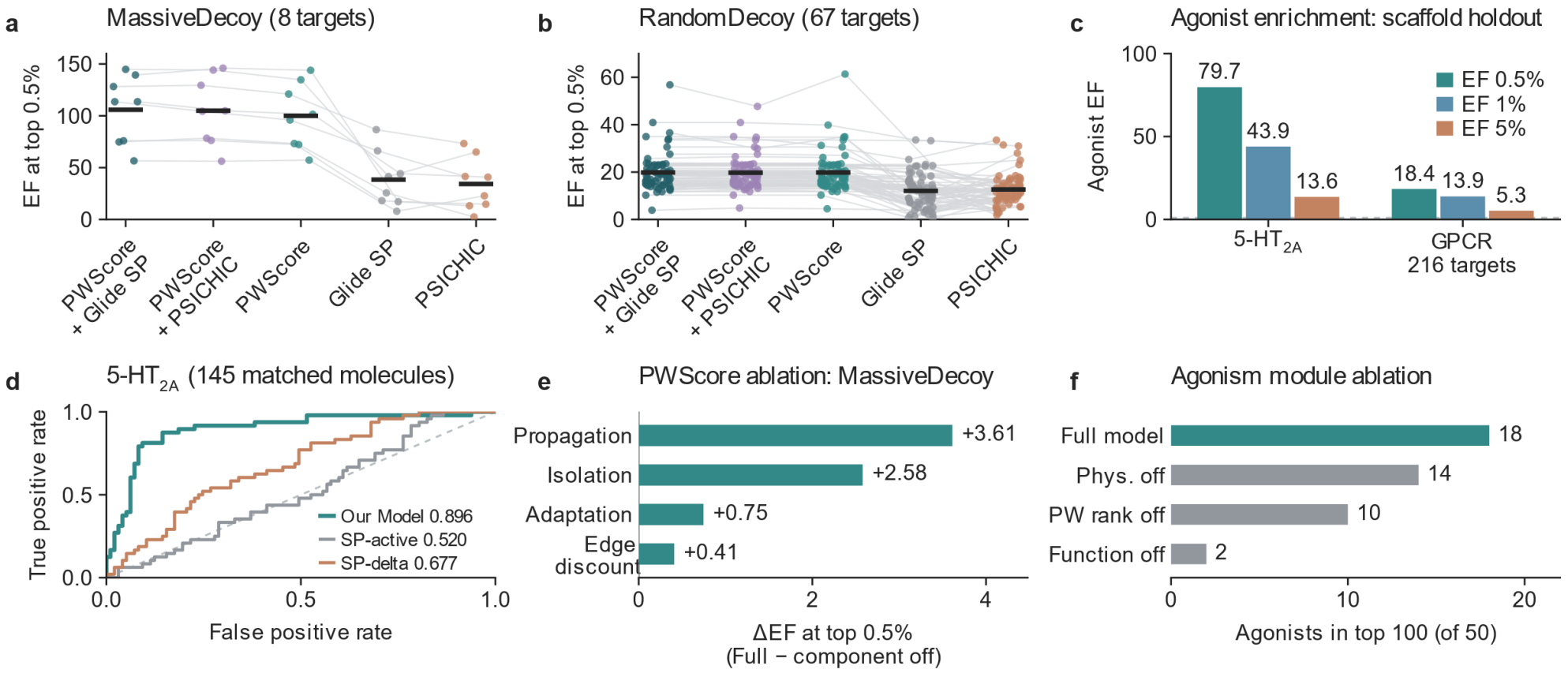
Structural validation of word–fragment rules



J. Chen … B.Tian* *Chemical Science* 2026

Evaluation of Word-fragment rules in virtual screening tasks

Word-fragment rules outperform Glide and PSICHIC.

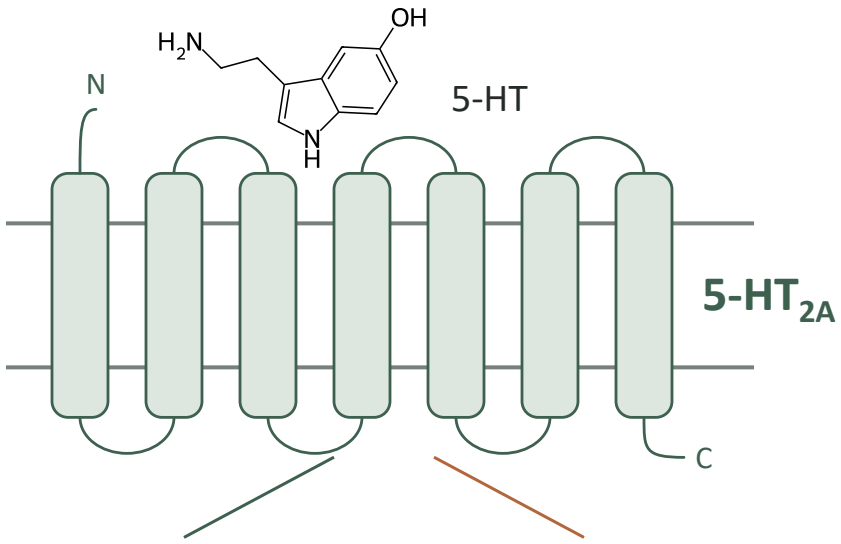


J. Chen … B.Tian* *unpublished data*

Discovery of GPCR agnoists with PWRules

Drug target background

A “Magic Mushroom” Inspires New Directions in Antidepressant Drug Development: Psilocybin has demonstrated antidepressant efficacy, but its hallucinogenic effects limit its therapeutic application.



Gq / PLC / Ca^{2+}

Gi

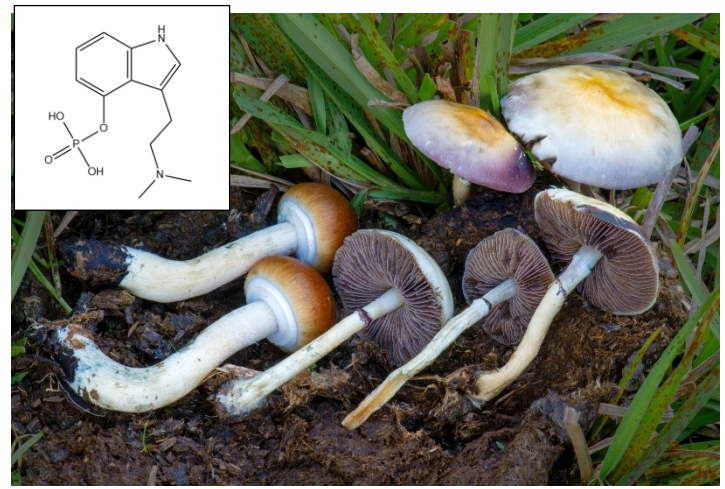
antidepressant efficacy

hallucinogenic effects

1) Find a binder

2) Find a agnoist

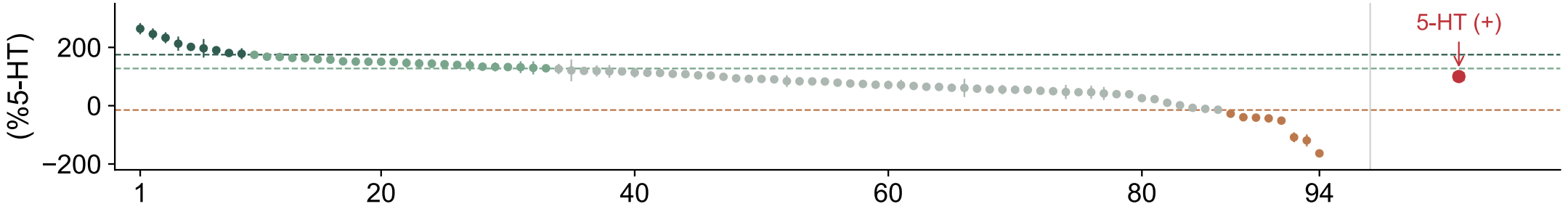
3) Gq bias

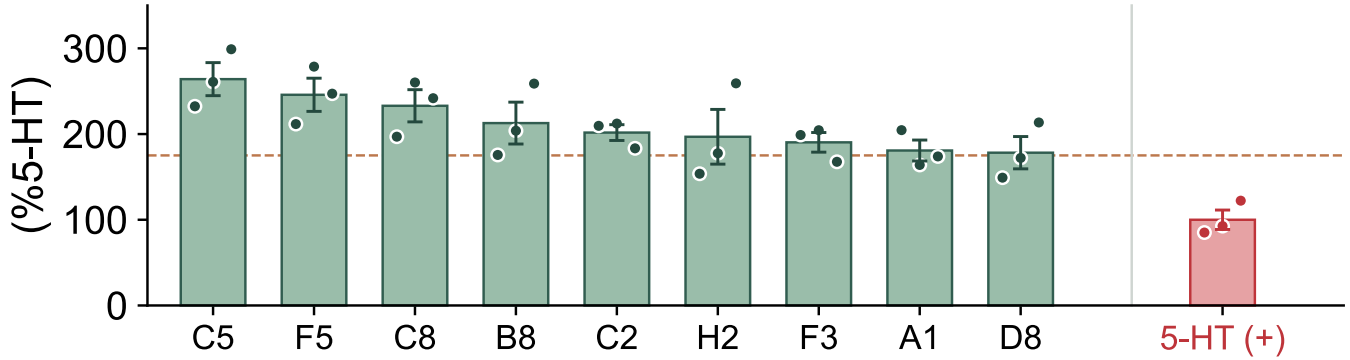


Psilocybin (5HT2A agonist) and Psilocybe

[1] Zhdanava et al., J Clin Psychiatry, 2021. [2] Goodwin et al., NEJM, 2022. [3] Xu et al., Nature, 2026.

Discovery of GPCR agnoists with PWRules





9 Agnoists

24 Partial agnoists

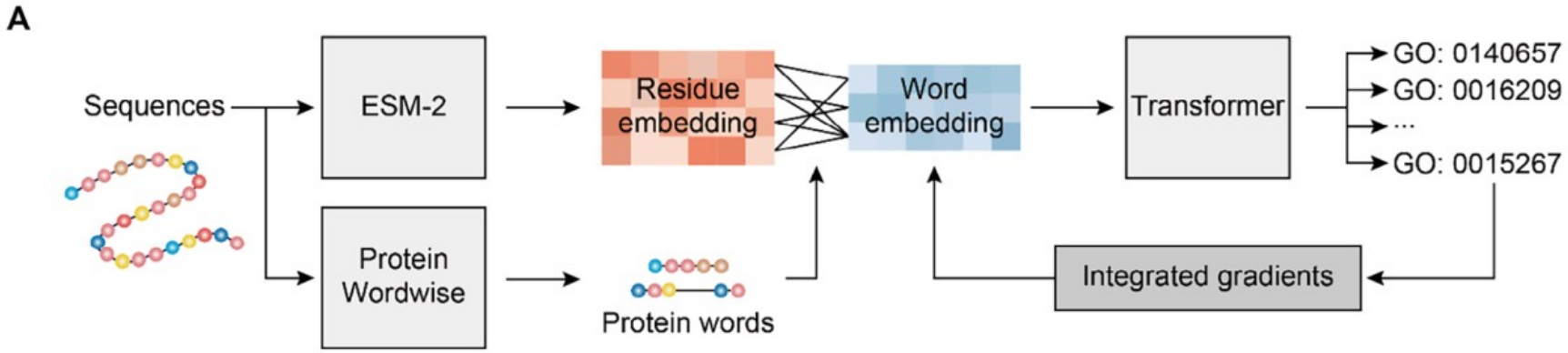
8 Antagonists

53 No activity

J. Chen … B.Tian* *unpublished data*

Protein Wordwise

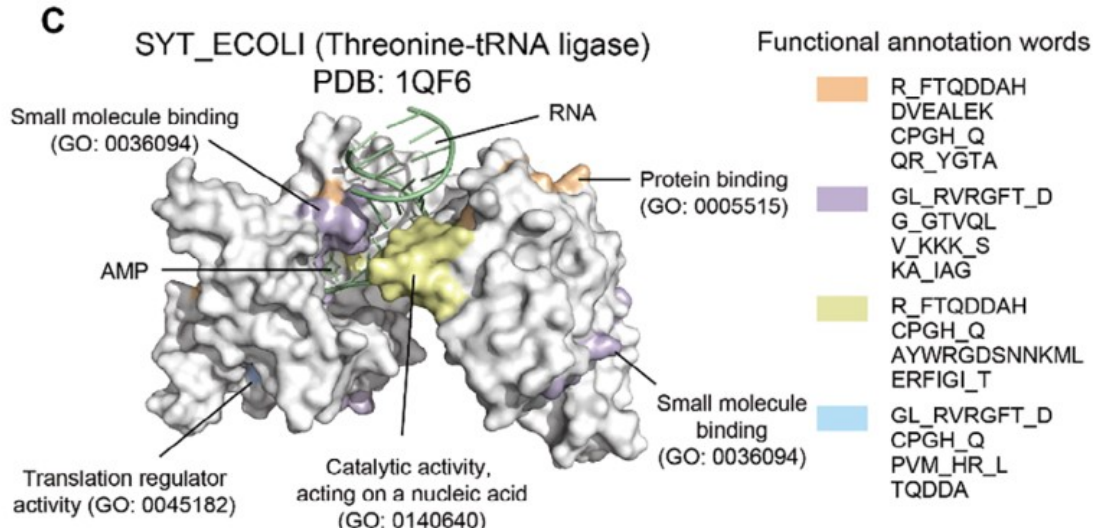
1) Unsupervised protein segmentation to build a dictionary (5–20 residues, contiguous/discontiguous)
2) Supervised learning of word functions



B

WordTableGO65

Function name	Words
DNA binding (GO: 0003677)	ACGLY, ...
Protein binding (GO: 0005515)	R_FTQDDAH, ...
Ligase activity (GO: 0016874)	SSGTTGL, ...
Antioxidant activity (GO: 0016209)	PERVVHARGA, ...
Channel activity (GO: 0015267)	ALG_LIF, ...
...	...



H. Chen ... B.Tian* *Advanced Science* **2026**

清华大学药学院
School of Pharmaceutical Sciences, Tsinghua University

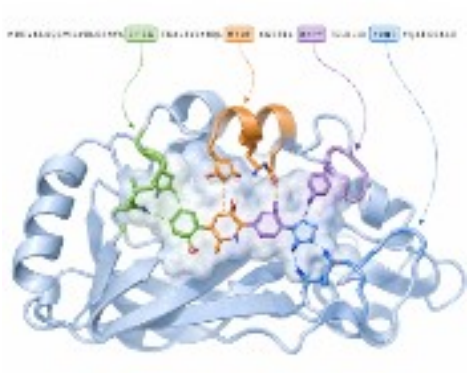
Summary

$$\sigma_r \leq 3e$$

AI Theory

Mathematically, the "Error Bound Theorem" is proved: the quality of substructure substituition rules determines the upper bound of the model's prediction accuracy.

Why we study “rules”



Drug Discovery with PWRules

Large/high-quality rules are automatically extracted from protein sequences to serve as the knowledge base for pharmaceutical science

How to extract rules and how to use them?

Code and Paper

1) Property prediction (MolRuleLoss & RuleRankingLoss)

X. Fan … B.Tian* *Chemical Science* **2026**

Z. Yang … B.Tian* *unpublished data*

2) Virtual screening (PWRules)

J. Chen … B.Tian* *Chemical Science* **2026**

J. Chen … B.Tian* *unpublished data*

3) Protein function prediction (Protein Wordwise)

H. Chen … B.Tian* *Advanced Science* **2026**

清華大學药学院
School of Pharmaceutical Sciences, Tsinghua University

Acknowledgements

Prof. Xiangyu Liu (Tsinghua University)

Prof. Jun Liu (Tsinghua University)

Boxue Tian
Tsinghua University
boxuetian@mail.tsinghua.edu.cn
https://tianlab-tsinghua.cn/



国家自然科学基金委员会 NSFC

2023 2023 CCF-BAIDU OPEN FUND CCF-百度松果基金 助推中国科技领跑世界

生命科学联合中心 CENTER FOR LIFE SCIENCES

KEY LAB OF BIOINFORMATICS MOE

FRCBS 北京生物结构前沿研究中心 Beijing Frontier Research Center for Biological Structure

国药集团 SINOPHARM

中国生物