清華大學 TSINGHUA UNIVERSITY

清華大學 智能产业研究院 Institute for AI Industry Research, Tsinghua University

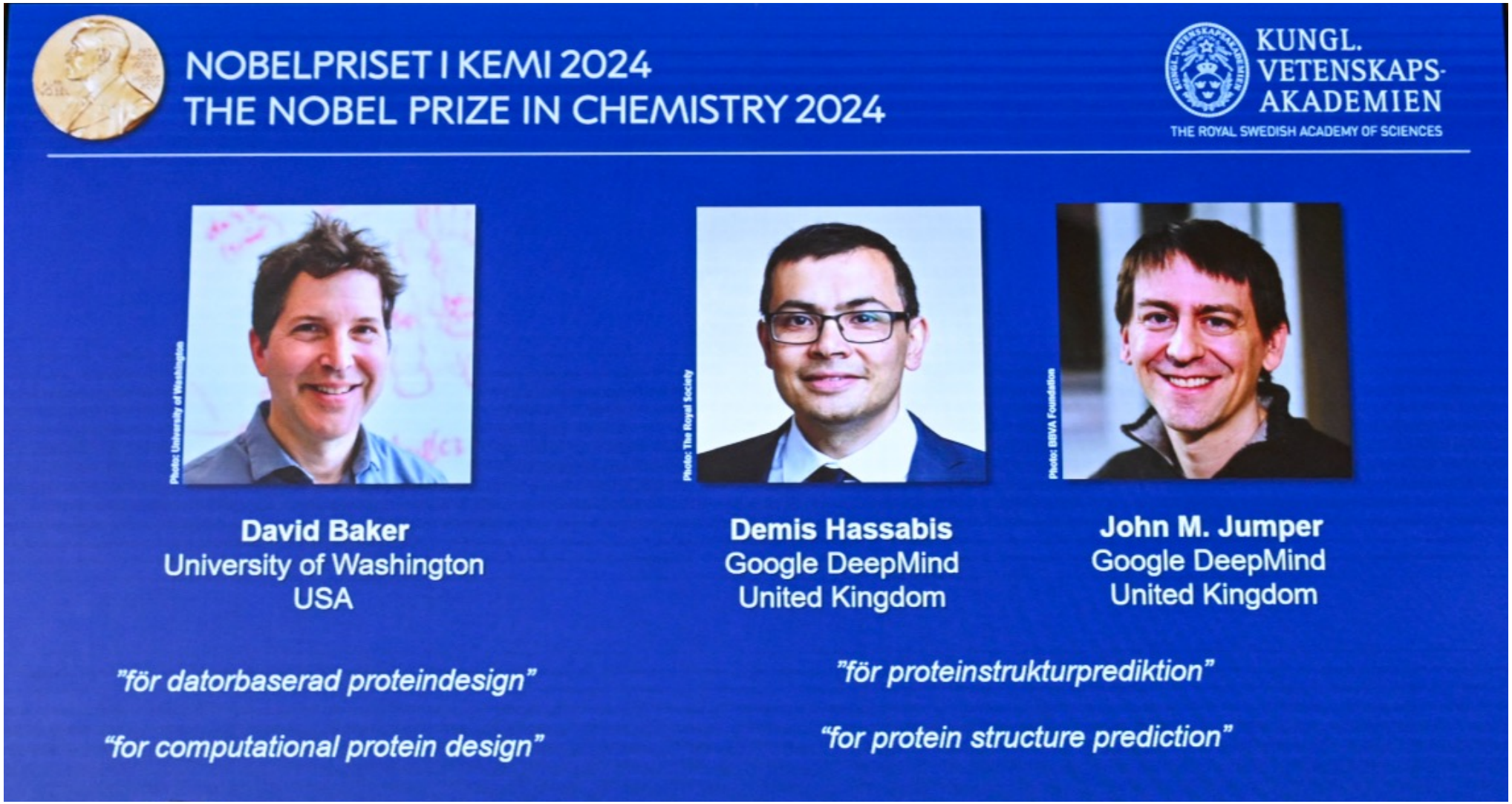
AI for Science
——概念、现状与展望

兰艳艳

清华大学智能产业研究院，首席研究员

人工智能学院，博士生导师

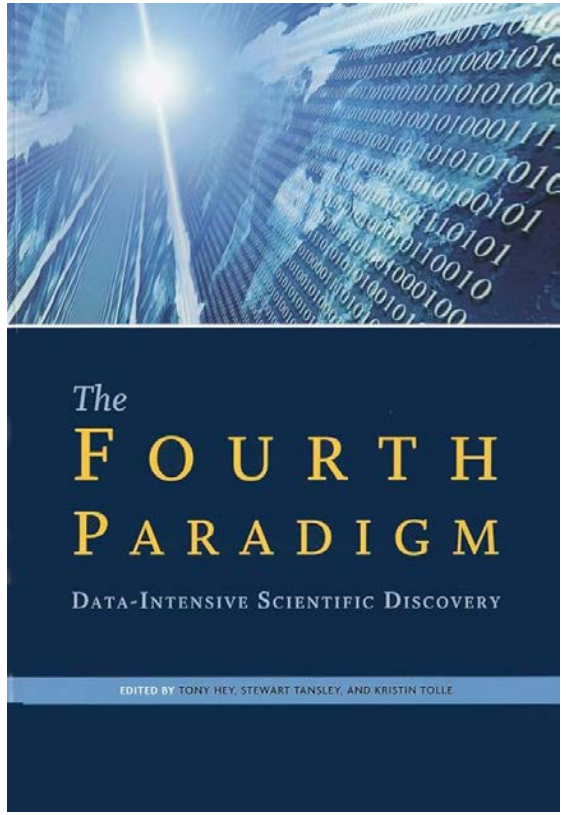
AI for Science：获得诺贝尔奖



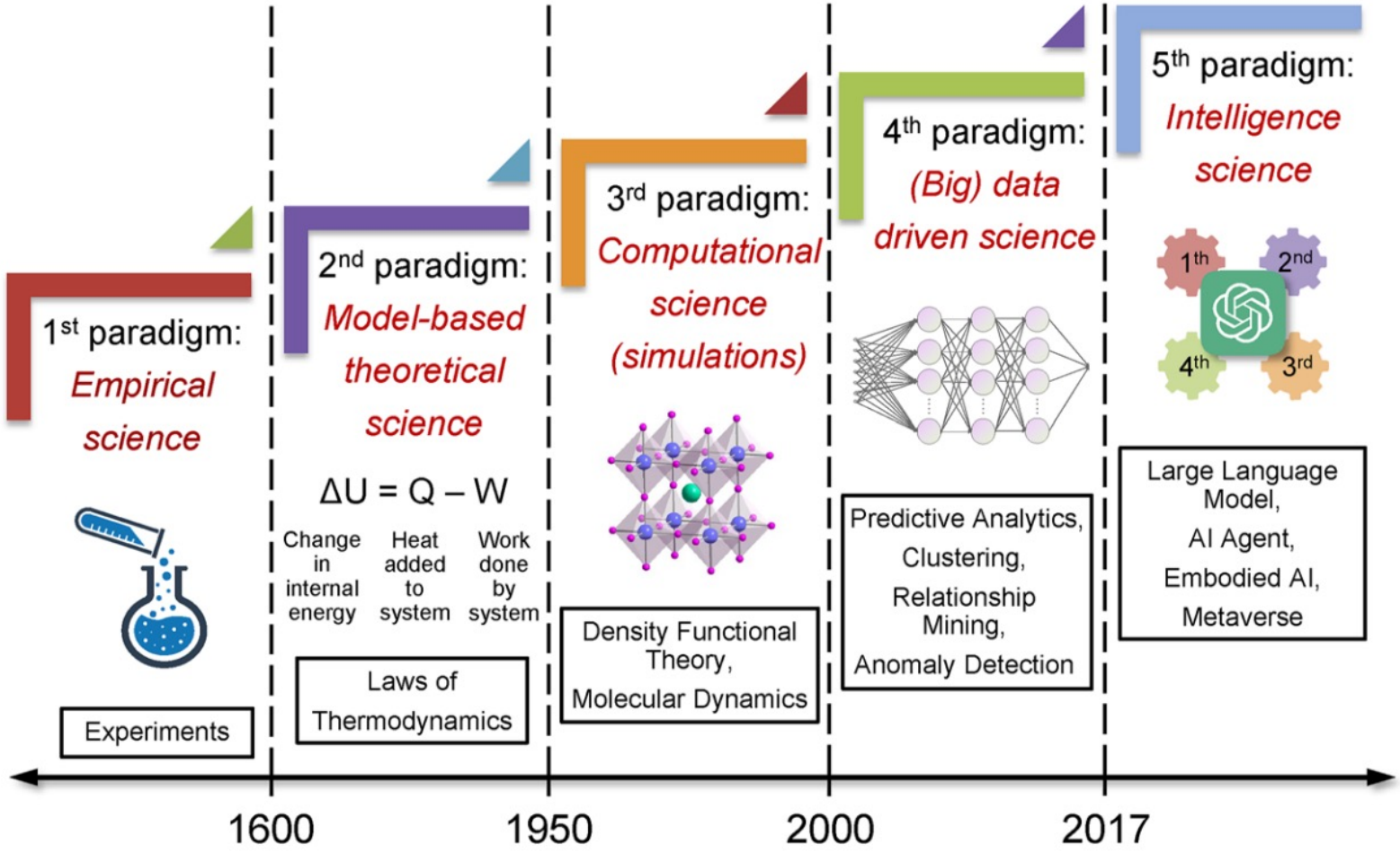
AI for Science的定义

- **一般定义：**利用人工智能技术解决科学研究中复杂问题和挑战的新兴领域。
- **Science:** systematic endeavor that builds and organizes knowledge in the form of testable explanations and predictions about the universe.
- **New Science:** systematic endeavor that builds and organizes knowledge and **models** for both **AI** and human in a **new cooperative manner**.
  - AI is good at understanding high dimensional data and extracting underlying patterns/structures/relationship from lage data
  - Such knowledge and models are often beyond human comprehension (human intelligence), e.g. LLM for biology and chemistry
  - AI first, of AI, by AI, for AI (AI as a Science) ——马维英教授（清华AIR）

AI for Science：科学变革的第五范式



Jim Gray (1944-2007)
图灵奖得主



新的要求：

- 不是已有AI技术的应用
- AI融合人的理论洞察
- 数据与知识的开放共享和安全
- 新一代科学家培养模式

APL Mater. 13, 020601 (2025)

AI for Science的战略地位

中华人民共和国中央人民政府
www.gov.cn

首页 | 简 | 繁 | EN | 登录 | 邮

科技部启动"人工智能驱动的科学研究"专项部署工作

2023-03-27 20:09 来源：新华社

字号：默认 大 超大 | 打印 |

新华社北京3月27日电（记者 宋晨）为贯彻落实国家《新一代人工智能发展规划》，科技部会同自然科学基金委近期启动“人工智能驱动的科学研究”（AI for Science）专项部署工作，紧密结合数学、物理、化学、天文等基础学科关键问题，围绕药物研发、基因研究、生物育种、新材料研发等重点领域科研需求展开，布局“人工智能驱动的科学研究”前沿科技研发体系。

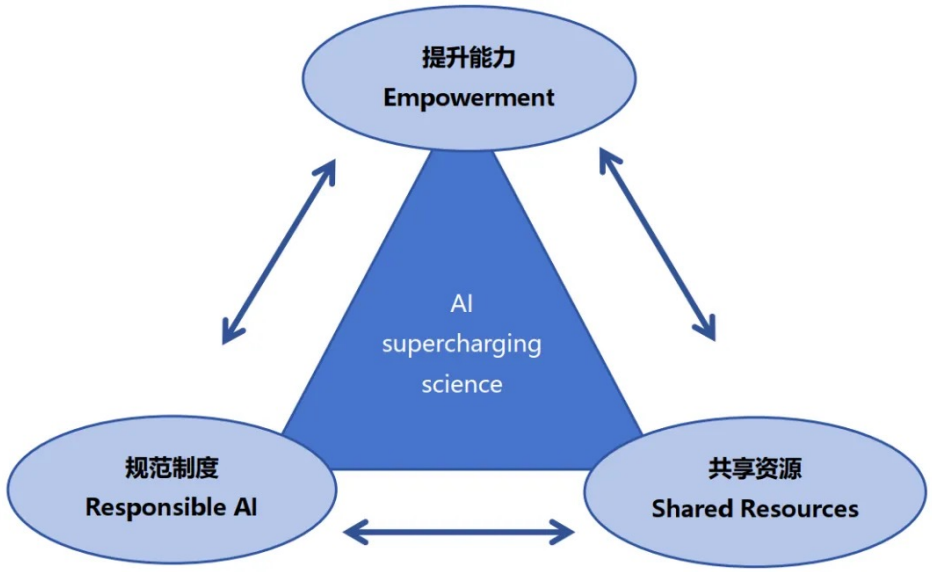
科技部有关负责人表示，当前，“人工智能驱动的科学研究”已成为全球人工智能新前沿。我国在人工智能技术、科研数据和算力资源等方面有良好基础，需要进一步加强系统布局和统筹指导，以促进人工智能与科学研究深度融合、推动资源开放汇聚、提升相关创新能力。

“人工智能驱动的科学研究”是以“机器学习为代表的人工智能技术”与“科学研究”深度融合的产物。中国科学院院士、北京大学国际机器学习研究中心主任鄂维南表示，借助机器学习在高维问题的表示能力，人类可以更加真实细致地刻画复杂系统的机理，同时可以把基本原理以更加高效、更加实用的方式应用于解决实际问题中。

科技创新2030–“新一代人工智能”重大项目实施专家组组长、中科院自动化研究所所长徐波介绍，人工智能技术已经在很多科学研究领域展现出超越传统数学或物理学方法的强大能力，但在“人工智能驱动的科学研究”体系化布局、重大系统设计、跨学科交叉融合、创新生态构建等方面仍有提升空间。

科技部将推进面向重大科学问题的人工智能模型和算法创新，发展一批针对典型科研领域的“人工智能驱动的科学研究”专用平台，加快推动国家新一代人工智能公共算力开放创新平台建设，支持高性能计算中心与智算中心异构融合发展，鼓励绿色能源和低碳化，推进软硬件计算技术升级，鼓励各类科研主体按照分类分级原则开放科学数据。

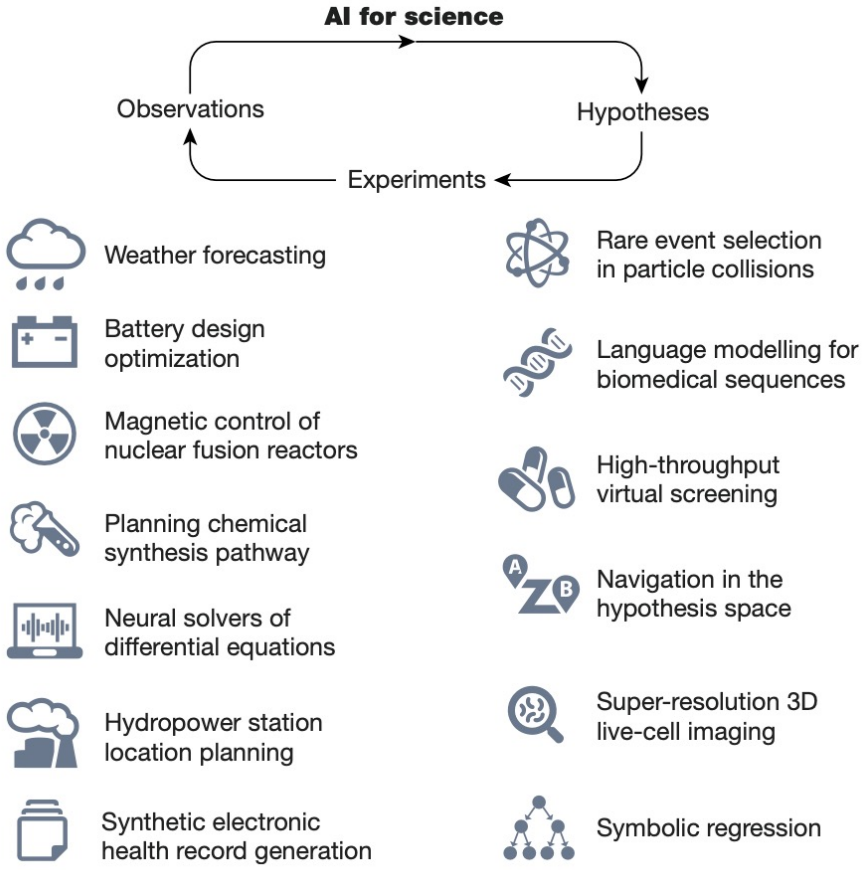
2024年4月29日，美国总统科学技术顾问委员会（PCAST）组织美国领域内权威专家，发布了《赋能研究：利用人工智能应对全球挑战》报告。

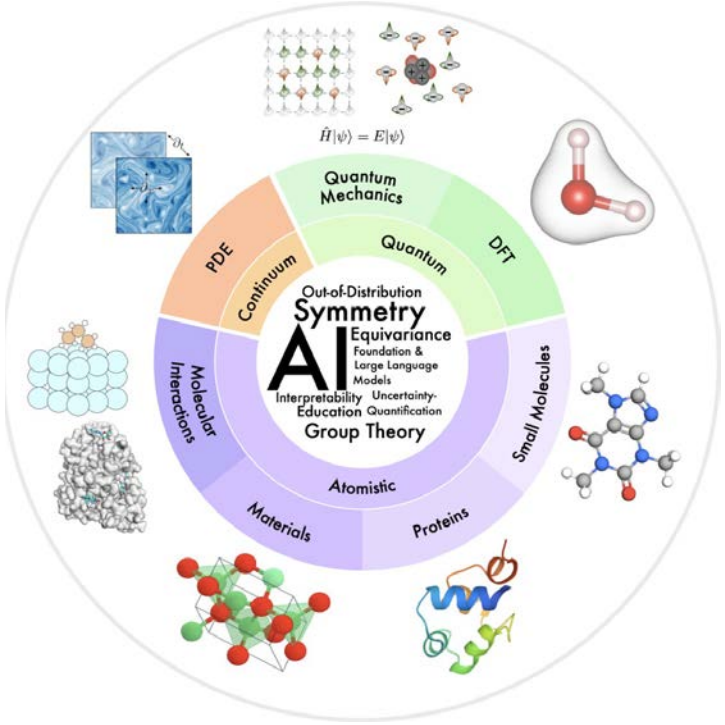


AI for Science的系统化技术体系与应用

五个部分：

- 智能化的数据收集与处理
- AI学习科学数据有意义的表示
- 生成式AI创造科学假设推动边界
- AI在复杂实验与模拟的作用
- 挑战与未来方向





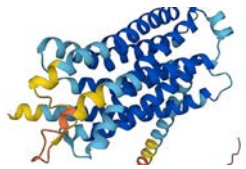
AI4S子领域

- 小尺度（亚原子尺度，波函数，电子密度）
- 中尺度（原子尺度，蛋白质，材料，相互作用）
- 大尺度（宏观系统，流体，气候，地下）

Scientific discovery in the age of artificial intelligence, Nature 2023.
Artificial intelligence for science in quantum, atomistic, and continuum systems, arxiv 2023.

AI for Science研究进展

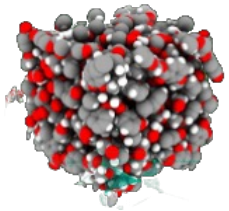
AI+生物：AlphaFold2



预测数量
提升200倍

《Science》年度科学突破榜首

AI+分子动力学： DeepMD-kit



计算耗时
100ms量级

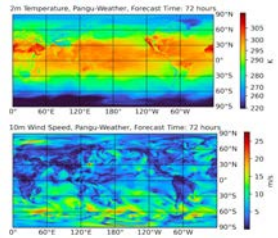
高性能计算最高奖戈登贝尔奖

AI+数学： AlphaTensor

打破矩阵乘法
50年速度记录

Nature封面文章，自动设计算法

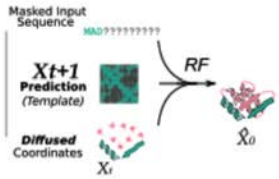
AI+气象： Pangu



秒级预测达到
最高精度

登顶Nature正刊，破解气象预测难题

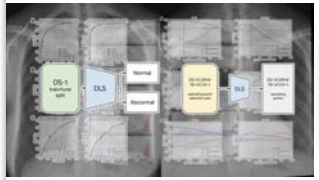
AI+生物： RFdiffusion



基于AI的蛋白
质从头设计

Nature正刊，最强扩散式蛋白设计

AI+医疗： DLS



阅片时间
节省30%

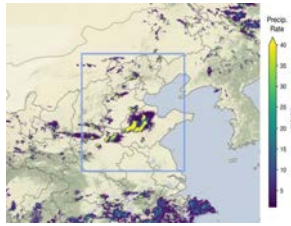
《Nature》正刊，提高放射医生效率

AI+化学： ROBO Chemist

持续工作的AI
化学家

Nature封面文章，研发全新催化剂

AI+气象： NowcastNet



准确预测未
来6小时天气

Nature正刊，突发极端预报成现实

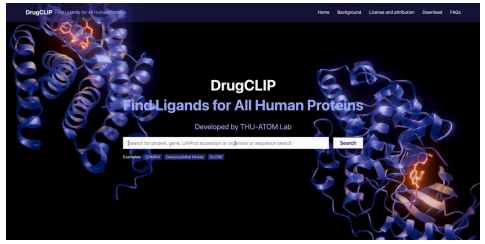
AI for Science适合做什么问题?

- **三个标准**
  1. 是否能够将问题描述为在一个庞大的组合搜索空间或解空间中的搜索问题
  2. 是否有一个明确的目标函数或评价指标，可以用来优化和不断改进
  3. 是否拥有大量的数据，或者至少是一个高效准确的模拟器，可以生成来自正确分布的合成数据

Demis Hassabis
DeepMind创始人及CEO

智能药物研发为例

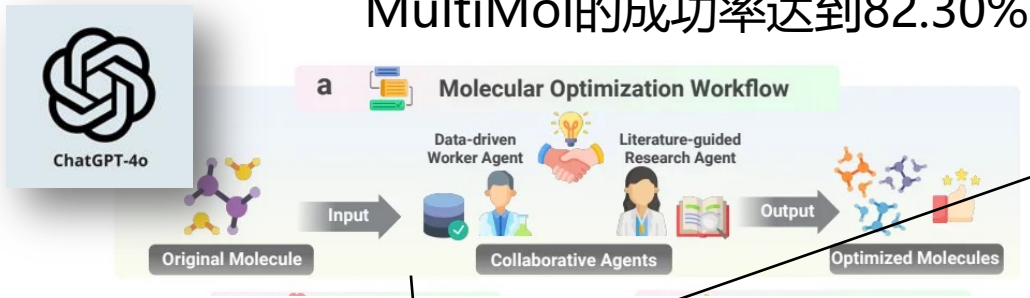
虚拟筛选



DrugCILP, NeurIPS 2023
加速10^6倍，一天完成全基因组蛋白筛选5亿小分子

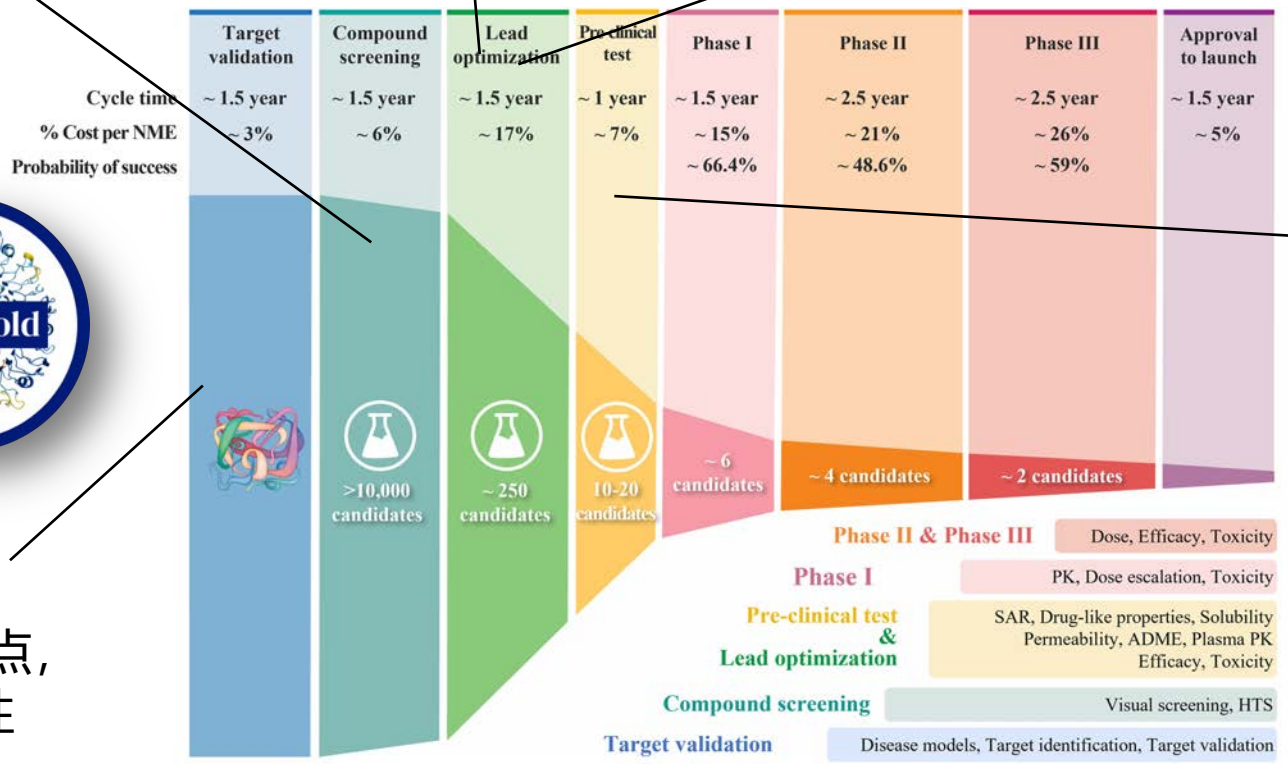
药物优化

MultiMol, 2025
在六个多目标优化任务中，MultiMol的成功率达到82.30%

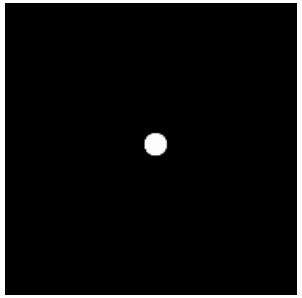


靶点发现

PandaOmics, JCIM 2024
识别28个潜在的ALS治疗靶点，验证了其中8个基因的有效性

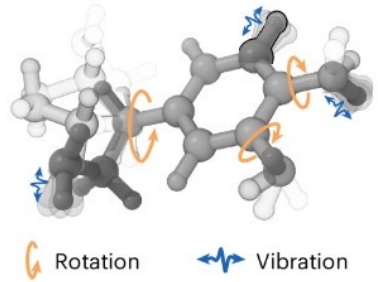


De novo生成



SLDM, 2025
分子生成效率提升100倍

性质预测

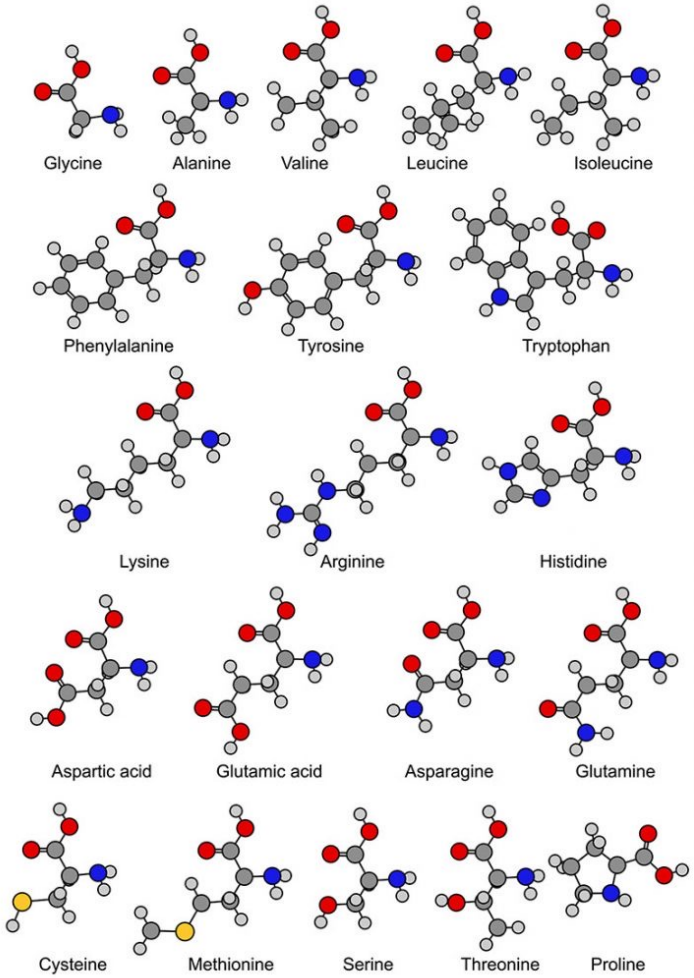


Frad, NMI 2024
接近量子化学精度
比传统量子化学（DFT）快2-3个数量级

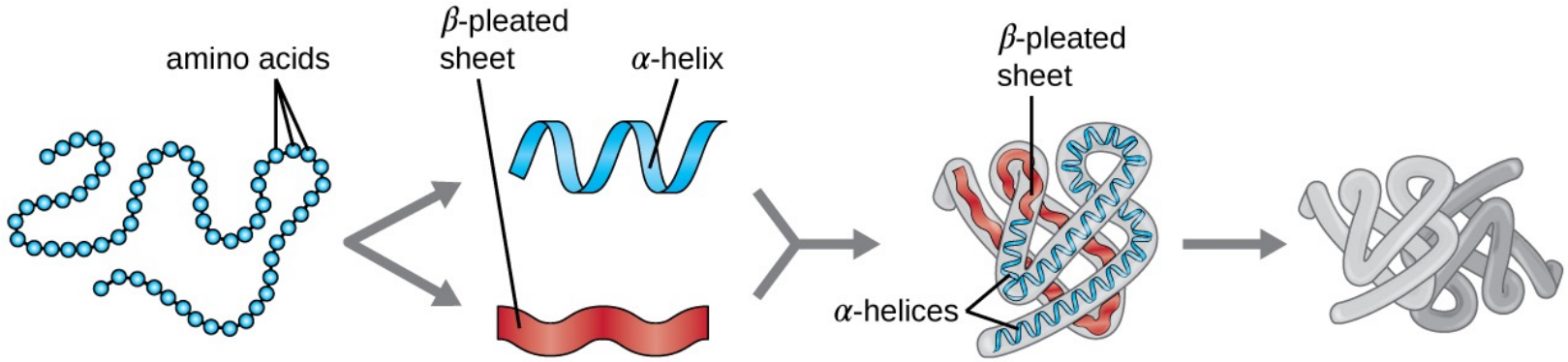
智能药物研发中的AI4S问题

- 蛋白质结构预测 —— 回归/生成
- 大规模药物虚拟筛选 —— 信息检索
- 生成式药物设计 —— 生成式人工智能
- 分子大模型及药物性质预测 —— 大模型
- 药物发现智能体 —— 智能体

蛋白质序列、结构和功能



20种常见氨基酸



蛋白质一级结构

氨基酸序列

蛋白质二级结构

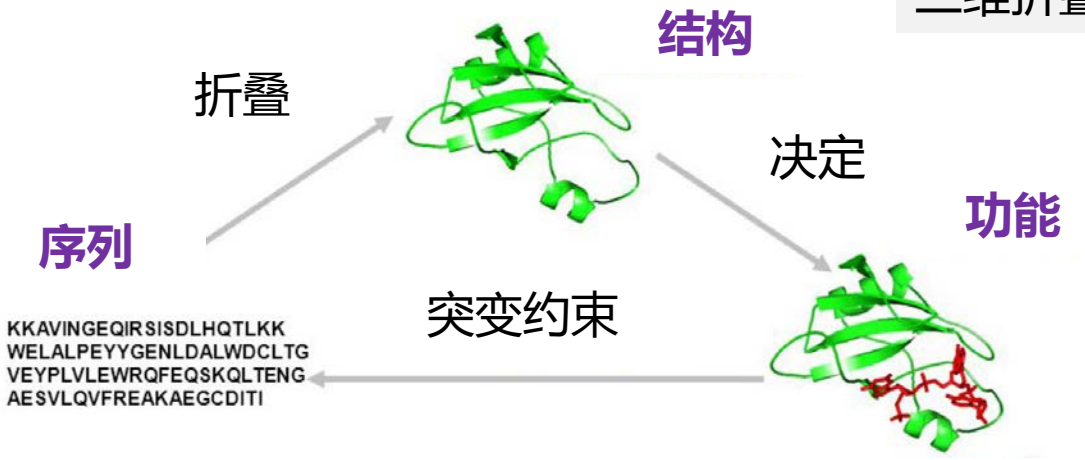
多肽链局部折叠形成螺旋或折叠片

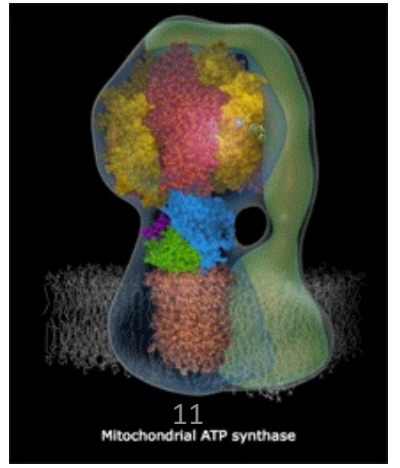
蛋白质三级结构

由于侧链相互作用，蛋白质形成三维折叠结构

蛋白质四级结构

由多个氨基酸链组成的蛋白质





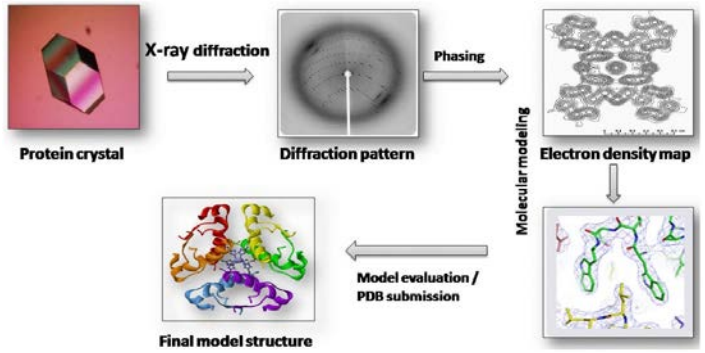
“分子马达”
ATP合成酶

蛋白质结构预测问题

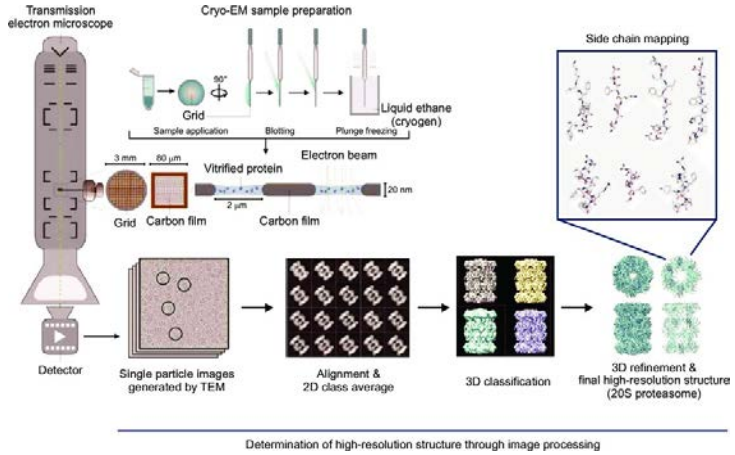
蛋白质3D结构解析成本高、限制多

蛋白质序列易获得

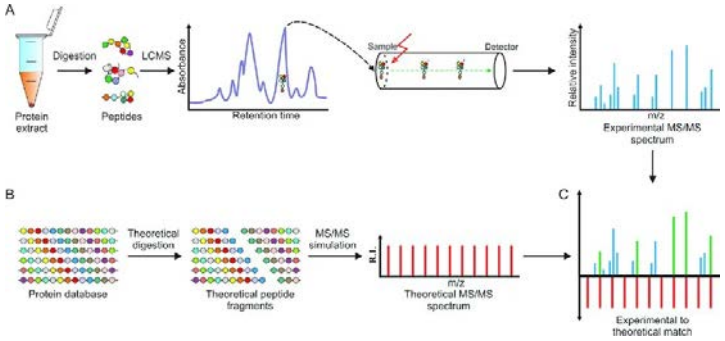
X射线晶体衍射



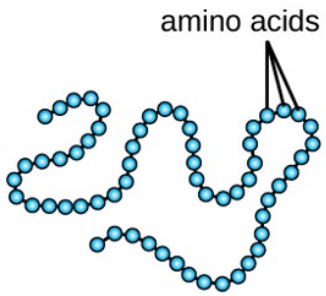
冷冻电镜



蛋白质谱法测序

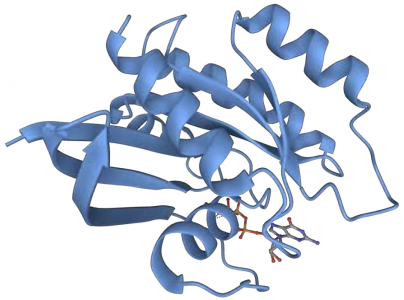


如何根据蛋白质的**氨基酸序列**预测蛋白质的**三维结构**?



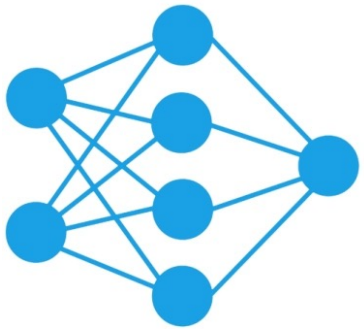
序列

?

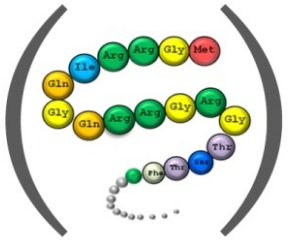


三维结构

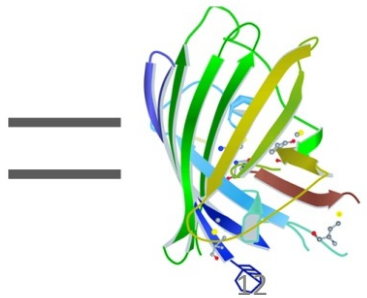
深度神经网络



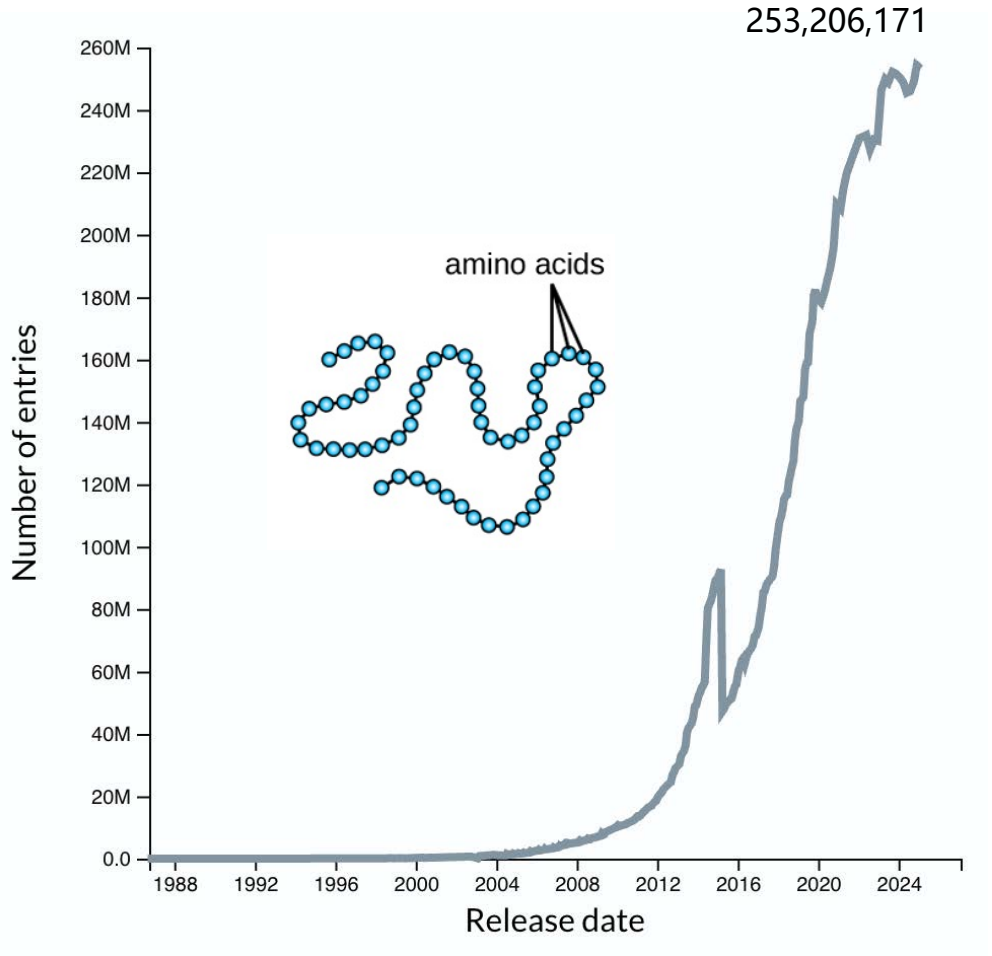
氨基酸序列



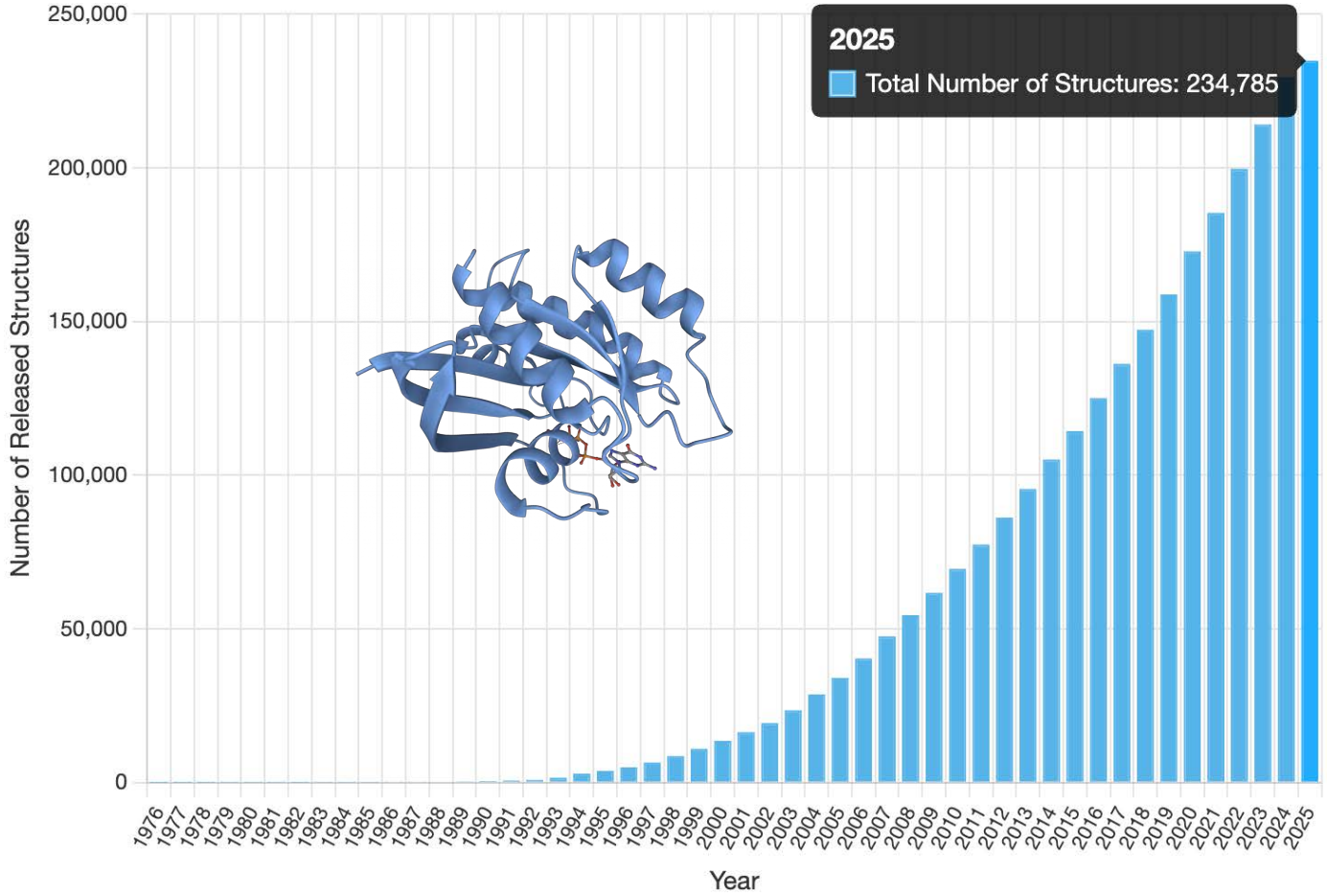
蛋白质结构



高质量数据的大规模增长

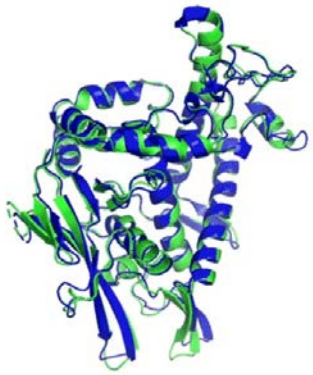


截止2025.04，**Uniprot蛋白质序列数据库**
共包含253,206,171条蛋白序列数据

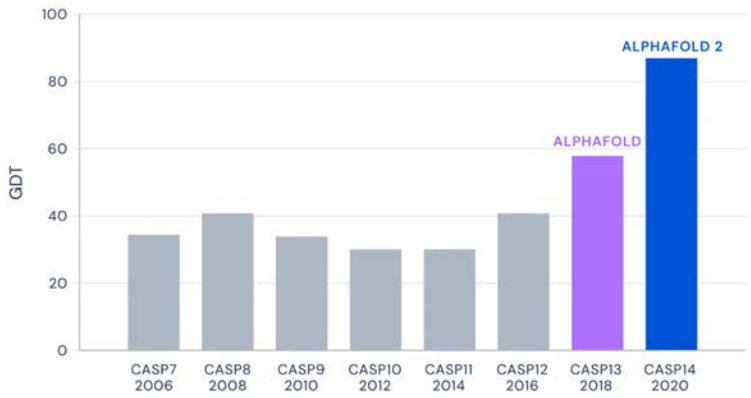


截止2025.04，**PDB蛋白质结构数据库**
共包含234,785个蛋白质结构数据

评价指标及Benchmark

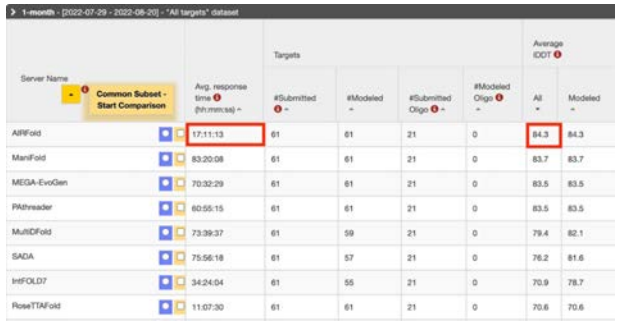


指标	英文全称	评估目标	公式	是否需要对齐	适用场景
RMSD	Root Mean Square Deviation	原子坐标偏差	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i^{\text{pred}}-\mathbf{x}_i^{\text{true}}\|^2}$	是	刚性结构比对
GDT-TS	Global Distance Test – Total Score	全局折叠准确性	$\frac{1}{4}\left(\frac{N_{d\le 1\text{Å}}}{N}+\frac{N_{d\le 2\text{Å}}}{N}+\frac{N_{d\le 4\text{Å}}}{N}+\frac{N_{d\le 8\text{Å}}}{N}\right)$	是	单体结构整体评估
lDDT	local Distance Difference Test	局部几何精度	$\frac{1}{N_{\text{pairs}}}\sum_{i,j}\mathbb{I}\left(\frac{\|d_{ij}^{\text{pred}}-d_{ij}^{\text{true}}\|}{\text{threshold}(d_{ij}^{\text{true}})}\le 0.5\right)$	否	局部构象评估
DockQ	Docking Quality Score	复合物界面质量	$\frac{1}{3}\left(F_{\text{nat}}+F_{\text{nonnat}}+\frac{1}{2}\left(1+\frac{\text{RMSD}_{\text{interface}}}{\text{RMSD}_{\text{max}}}\right)^{-1}\right)$	是	蛋白质对接、复合物评估



蛋白质结构预测技术的关键测试（CASP）

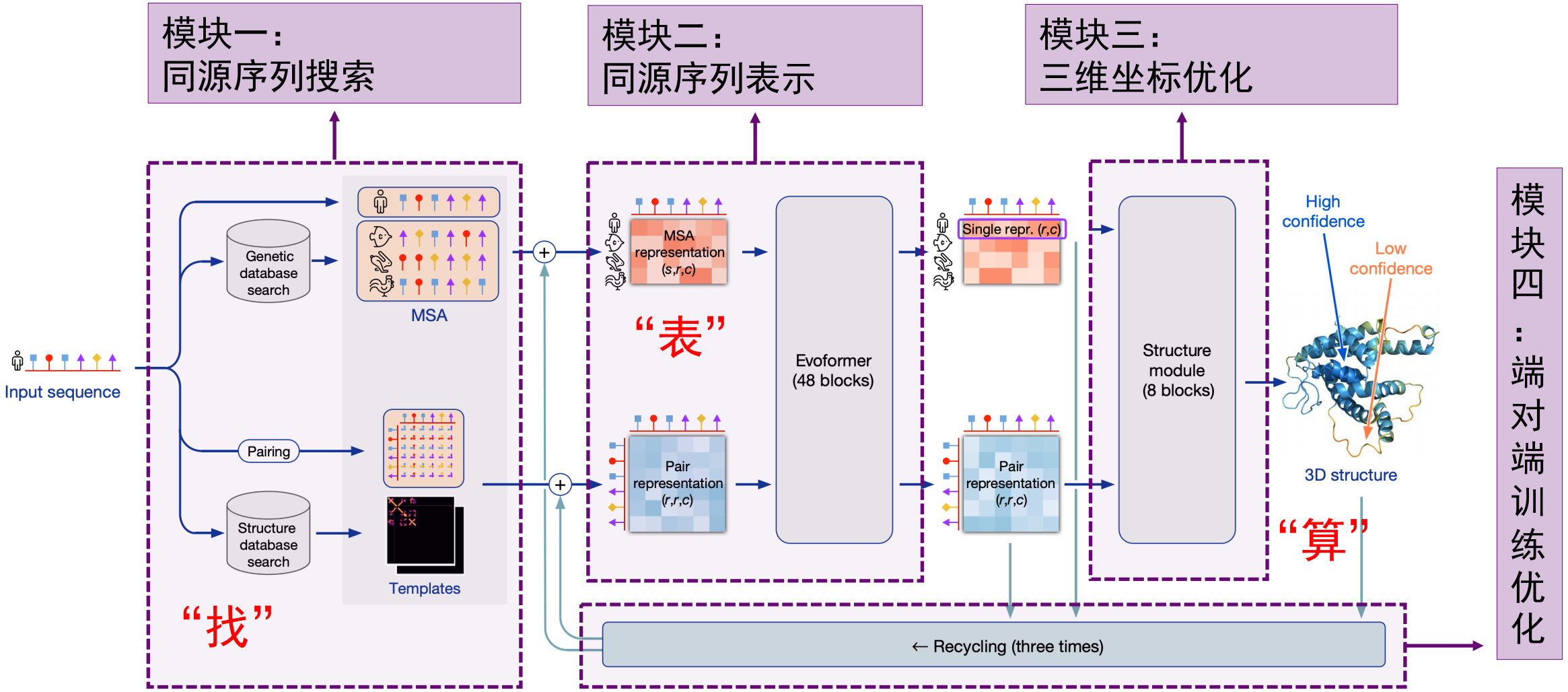
自1994年以来每两年进行一次的全球范围内的蛋白质结构预测竞赛。



全球持续蛋白质结构预测竞赛（CAMEO）

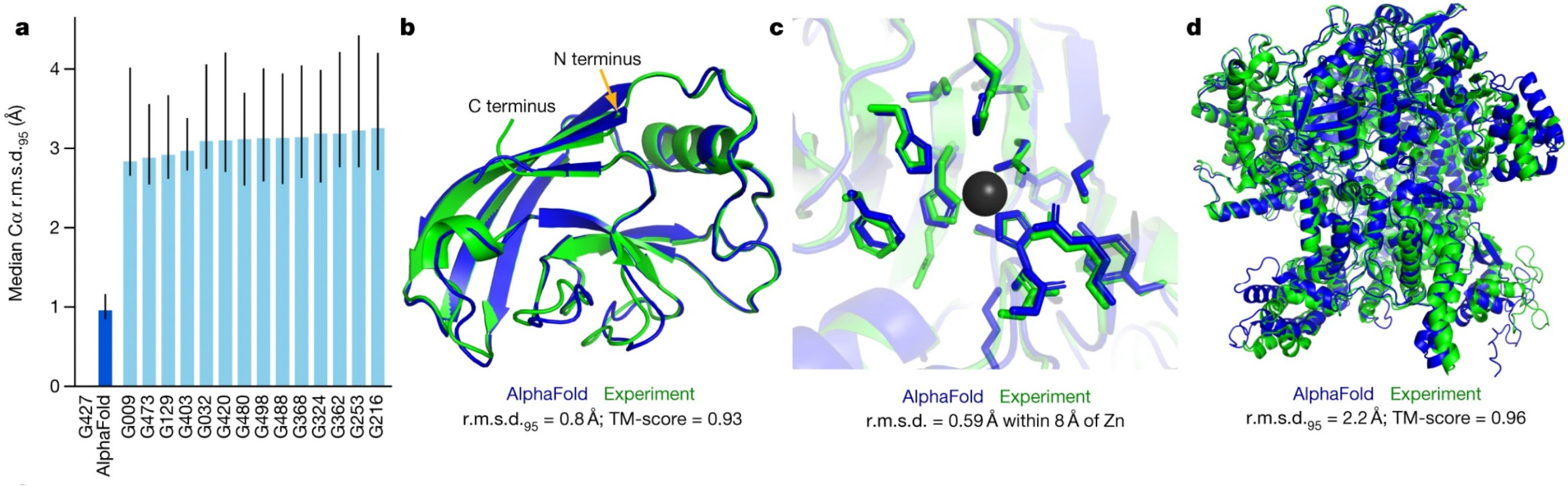
持续收集最新公开的蛋白质序列，每周从中挑选部分序列作为赛题。

AlphaFold2结构预测模型



Highly accurate protein structure prediction with AlphaFold. Nature, 2021. (Google DeepMind)

AlphaFold2预测结构达到实验精度



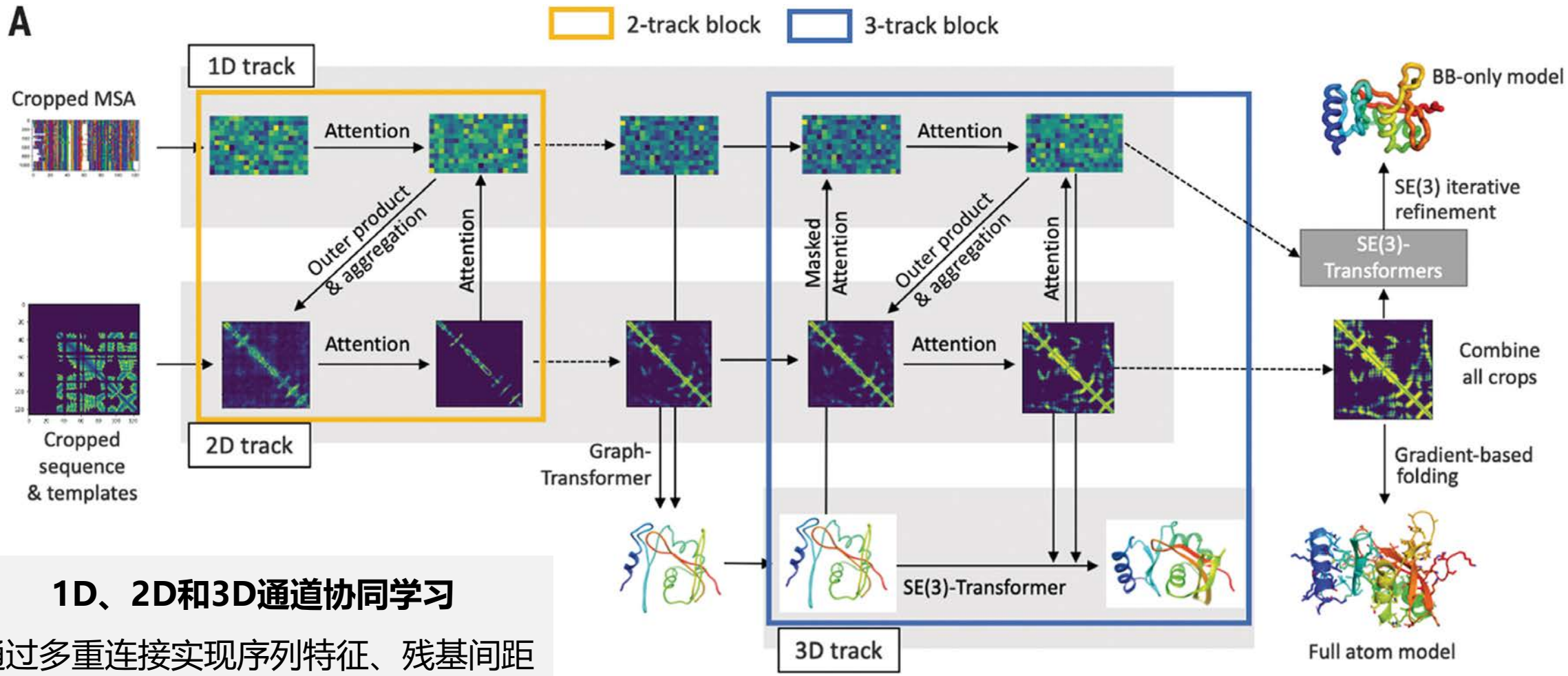
CASP14比赛总体结果

在CASP14测试集87个蛋白质结构域上，预测误差RMSD断崖式领先前15名其他模型

CASP14预测结果实例

单链蛋白、锌离子结合位点、超长链蛋白均达到实验水平

RoseTTAFold结构预测模型

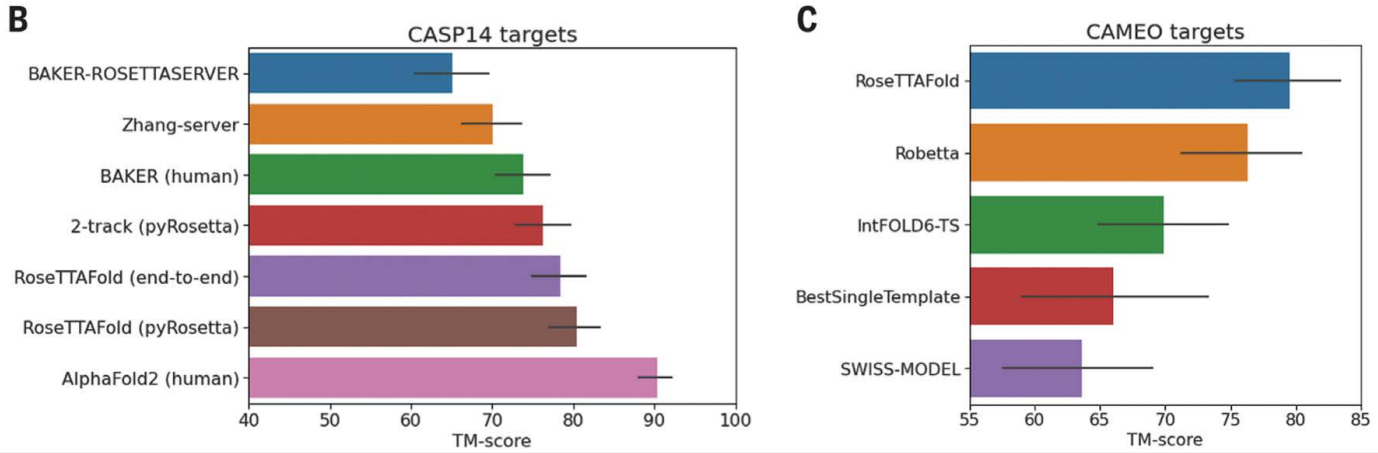


1D、2D和3D通道协同学习

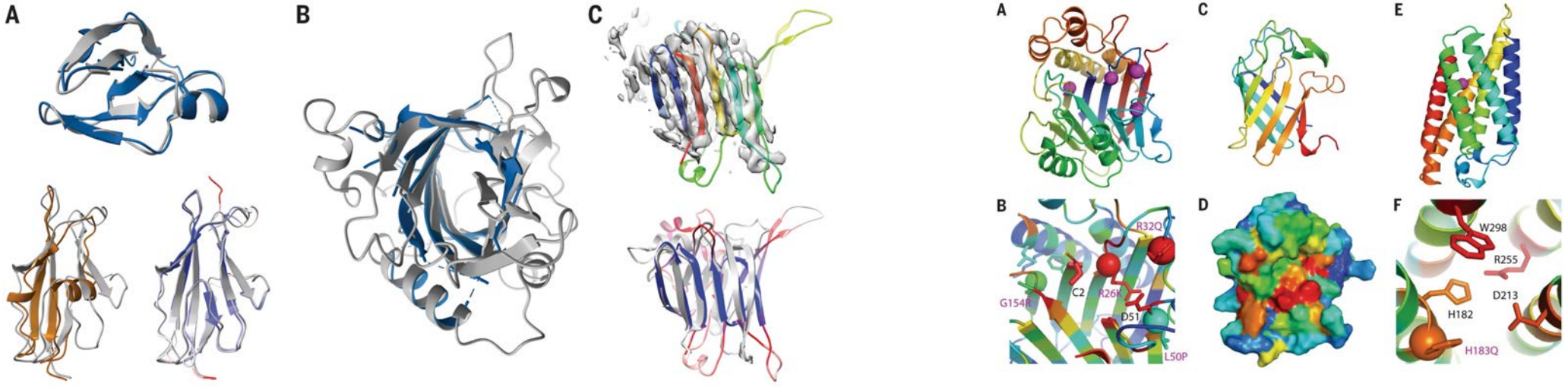
通过多重连接实现序列特征、残基间距和空间坐标关联关系的协同学习

Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021. (Baker Lab)

RoseTTAFold实验结果

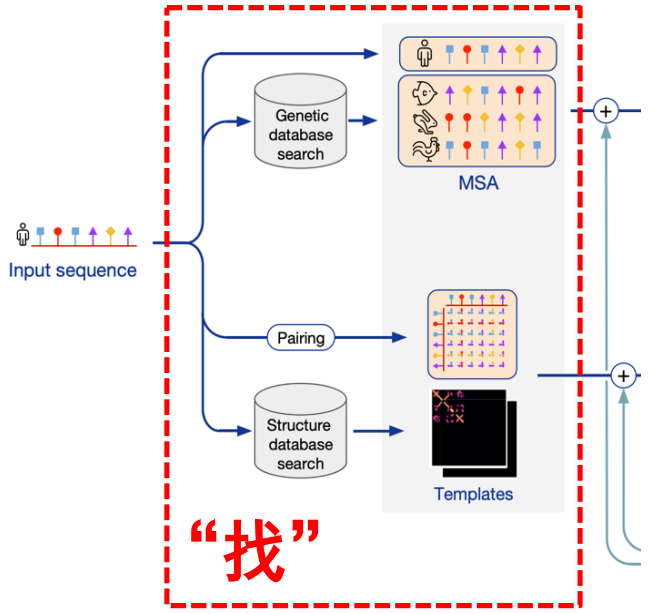


RoseTTAFold在CASP14和CAMEO评测上的预测精度领先

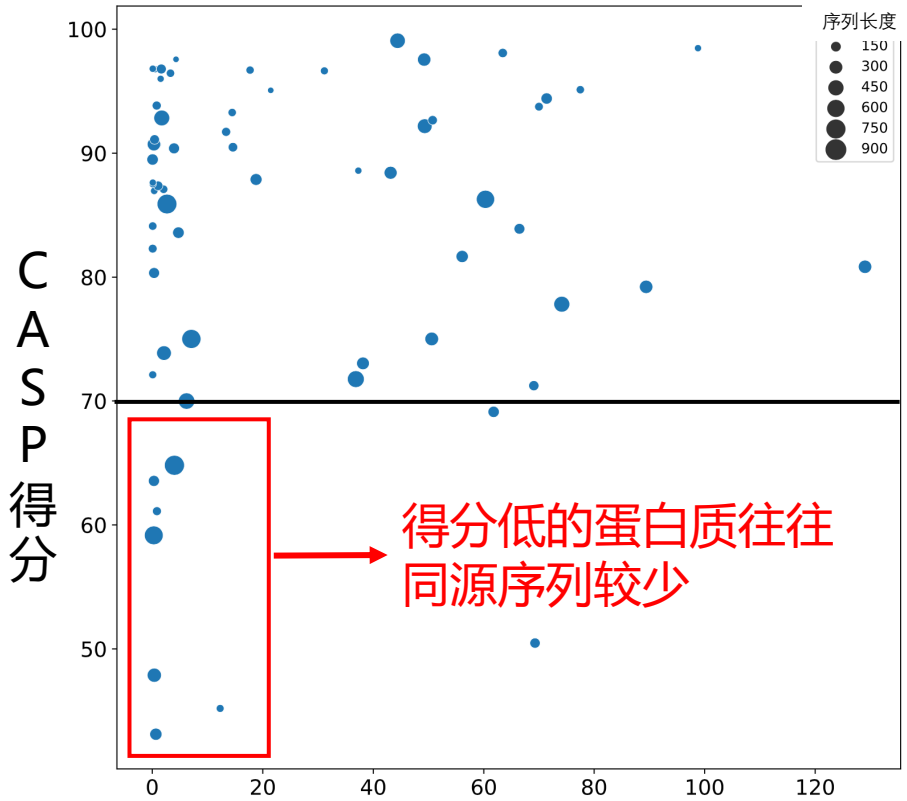


RoseTTAFold预测结构接近实验精度，并揭示蛋白质功能机制

AlphaFold2在同源低或无同源场景下失准

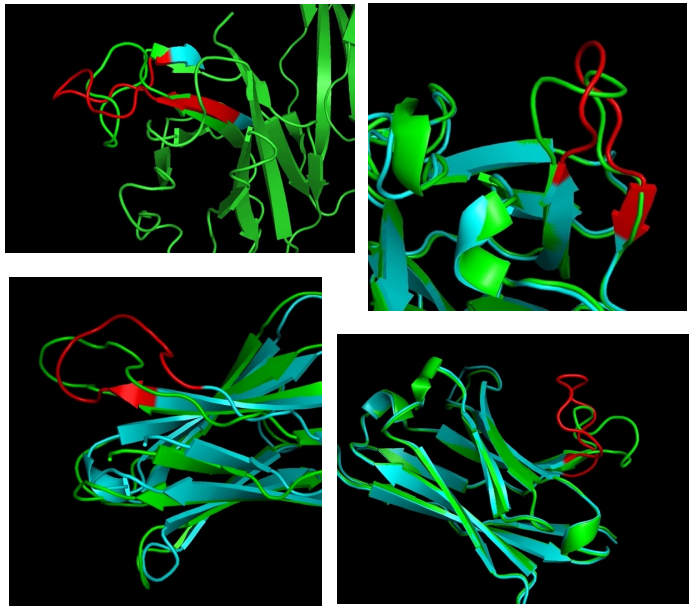


AlphaFold2的同源检索模块



有效序列个数

AlphaFold2输出 真实结构

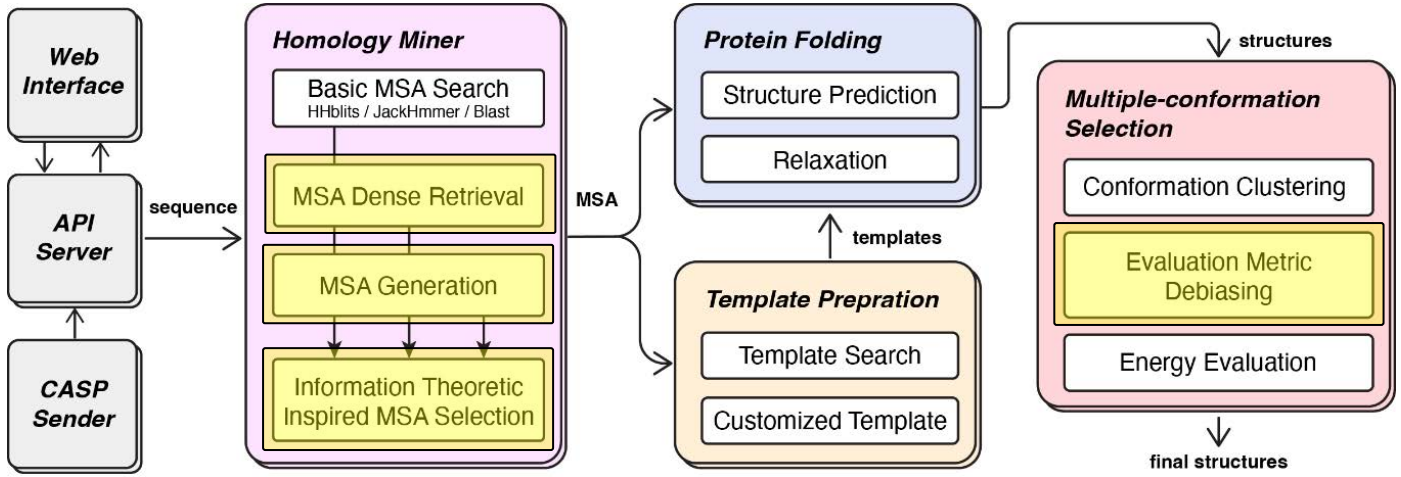


对抗体功能可变区（CDR3）的结构预测精度欠佳

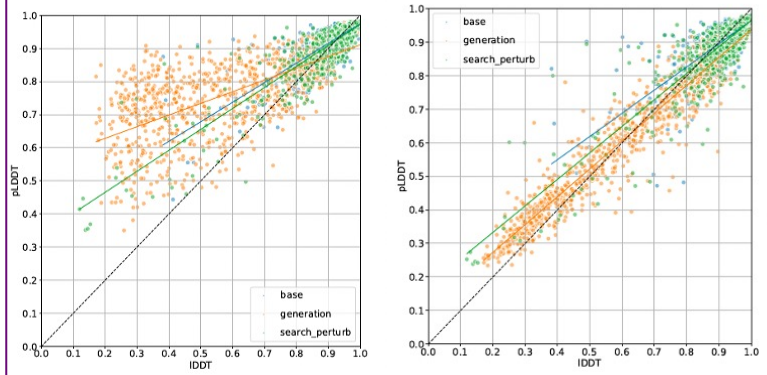
AlphaFold2对缺乏有效同源蛋白的蛋白质结构预测误差较大

AIRFold同源挖掘结构预测系统

AIRFold框架图

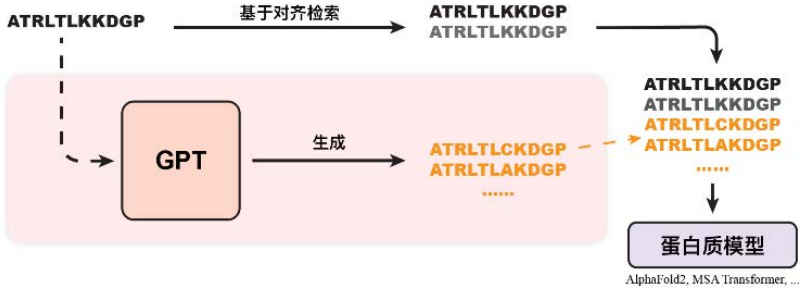


创新点四： pLDDT纠偏



创新点一：同源序列检索

创新点二：同源序列生成



创新点三：同源序列筛选

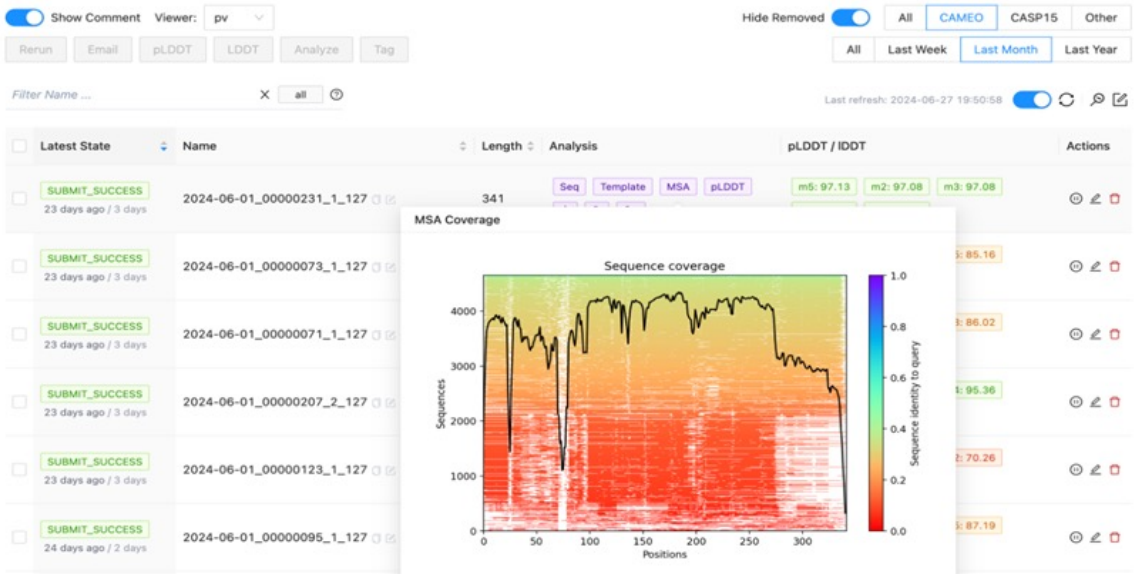
AIRFold 通过同源挖掘和pLDDT纠偏提升蛋白质结构预测模型精度

AIRFold, 2022. (Tsinghua AIR)

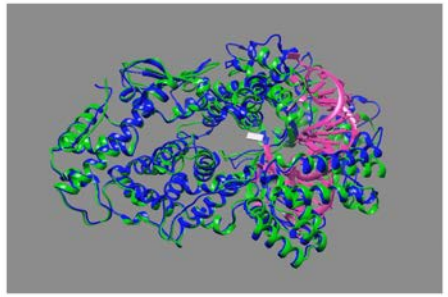
AIRFold实验结果

Server Name	Avg. response time (hh:mm:ss)	Targets #Submitted	#Modeled	#Submitted Oligo	#Modeled Oligo	lDDT All	lDDT Modeled
AIRFold	17:11:13	61	61	21	0	84.3	84.3
ManiFold	83:20:08	61	61	21	0	83.7	83.7
MEGA-EvoGen	70:32:29	61	61	21	0	83.5	83.5

2022年7月29日至2022年8月20日一连续4周获得CAMEO冠军

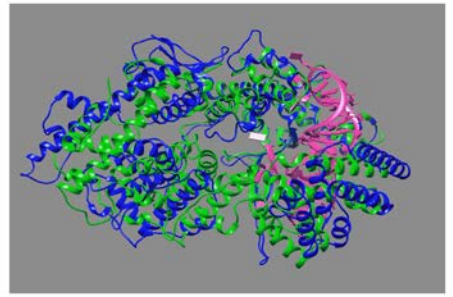


AIRFold具有用户友好的交互界面和便捷的结果可视化系统



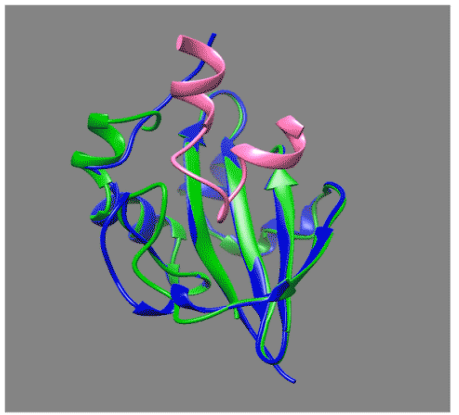
AIRFold Pred vs. PDB: 7VTI_A

TM-score:0.91 IDDT: 81.74



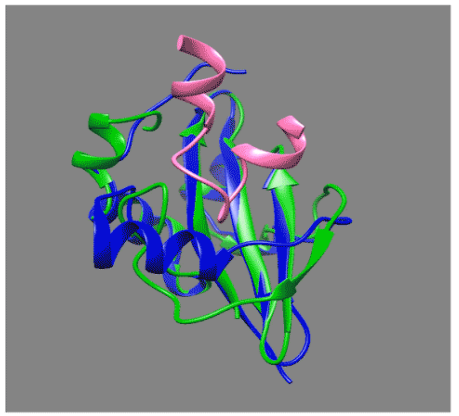
AlphaFold2DB vs. PDB: 7VTI_A

TM-score: 0.64 IDDT: 60.11



AIRFold vs. PDB: 7ENR_C

RMSD:0.891 IDDT: 85.73

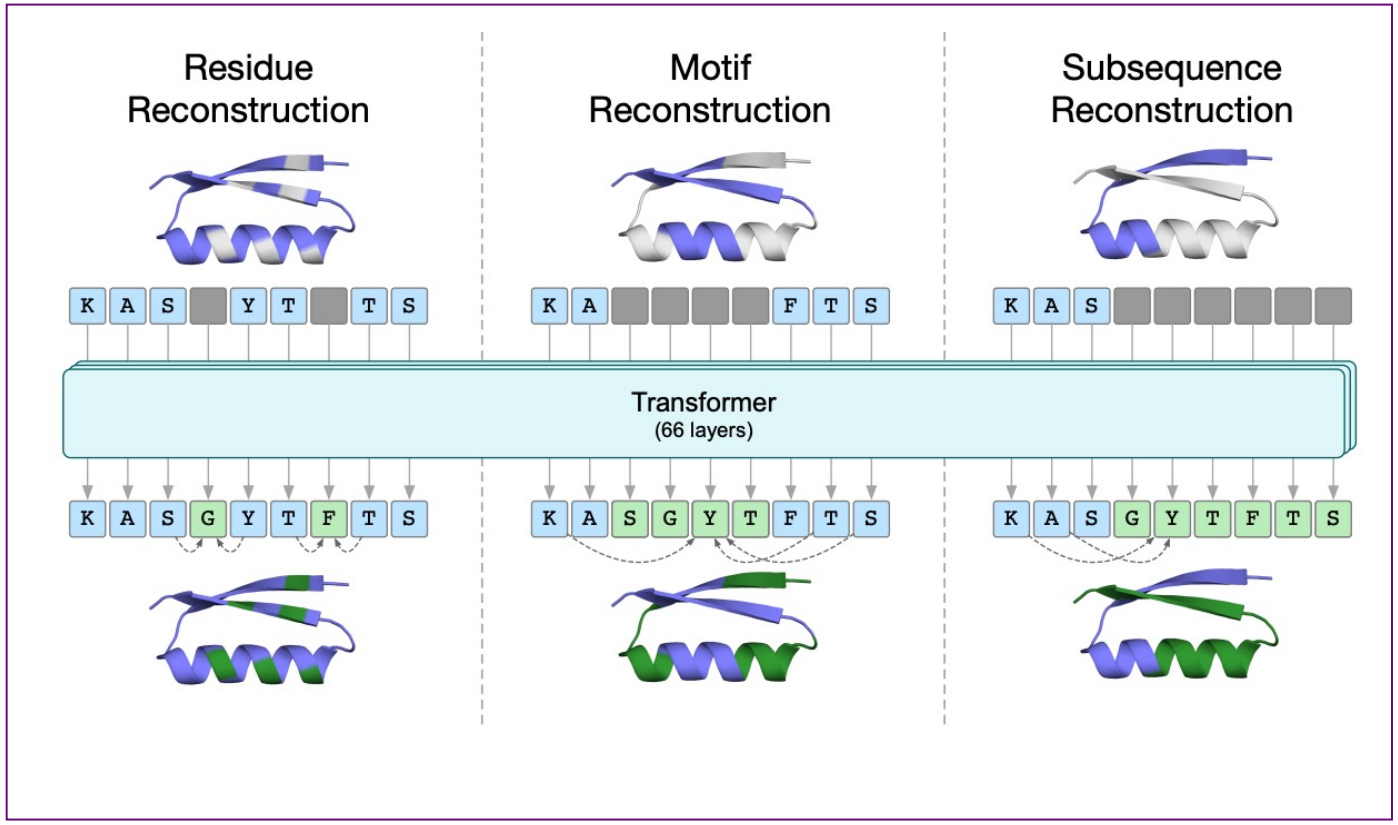


AlphaFold2 vs. PDB: 7ENR_C

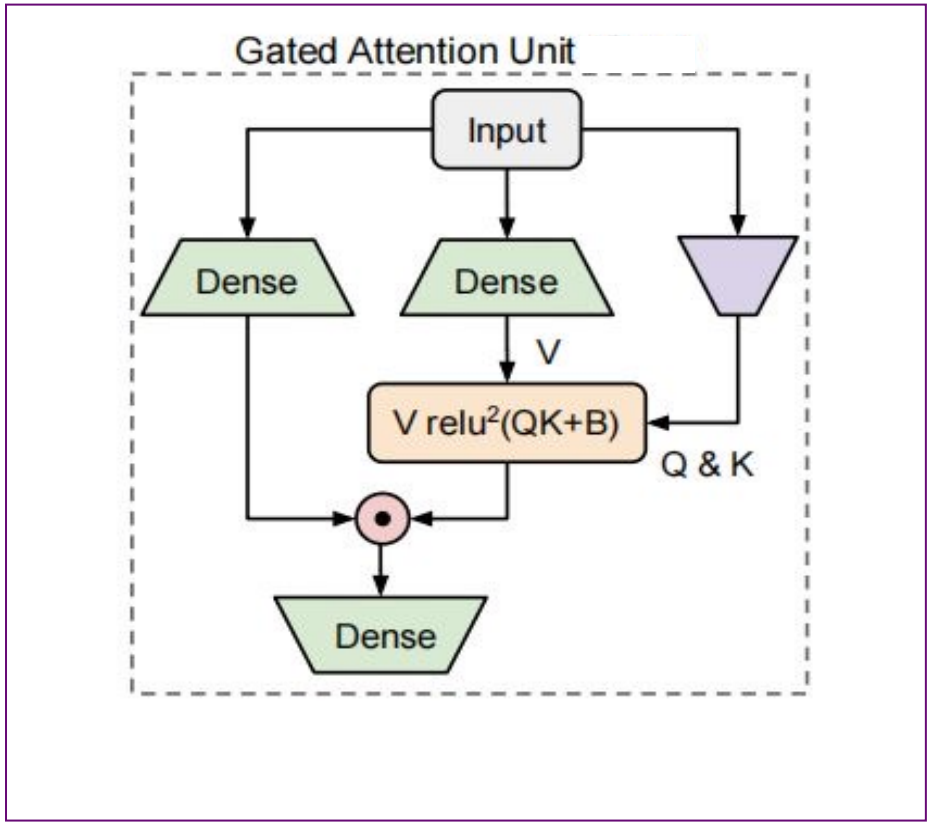
RMSD: 1.552 IDDT: 69.22

在国际蛋白质结构预测大赛 CAMEO 上**连续四周获得冠军**，在**缺乏同源蛋白序列**的基因治疗关键蛋白Cas13和**孤儿蛋白**AcrIIA14上预测结果**准确性显著优于AlphaFold2。**

OmegaPLM：通过蛋白质预训练捕捉同源信息



OmegaPLM预训练的三种损失



OmegaPLM的基本单元GAU

OmegaPLM 包含66层GAU网络，参数量670M

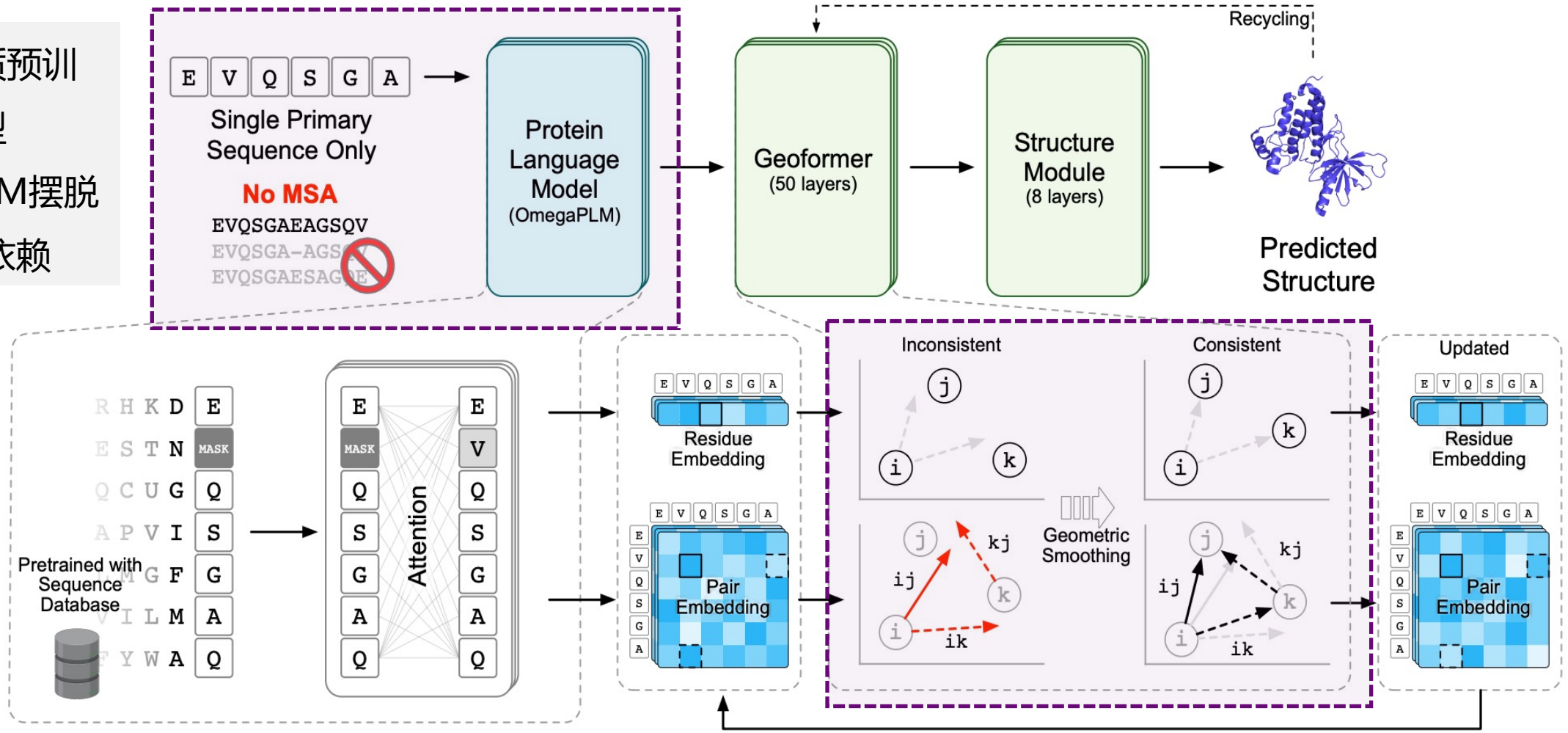
OmegaPLM 采用多层次，多粒度的训练目标函数对模型进行优化

GAU 具有参数量小、运算速度快等特点

High-resolution de novo structure prediction from primary sequence. bioRxiv, 2022. (Helixon)

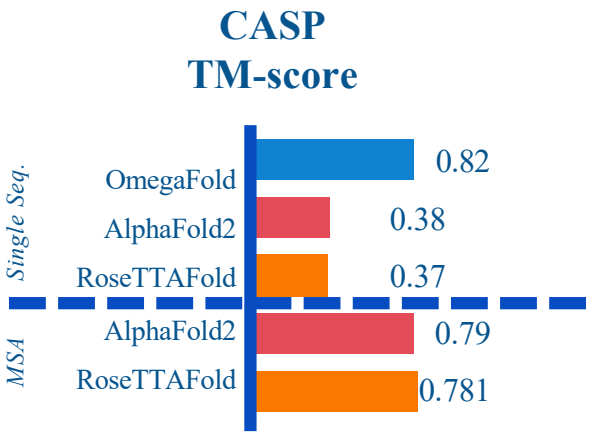
OmegaFold单序列结构预测模型

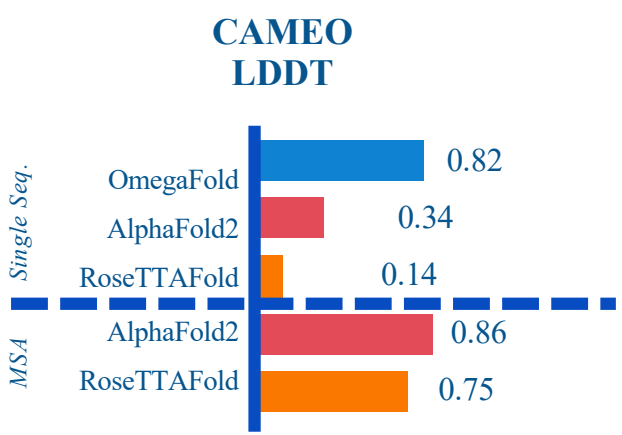
通过蛋白质预训练语言模型OmegaPLM摆脱对MSA的依赖

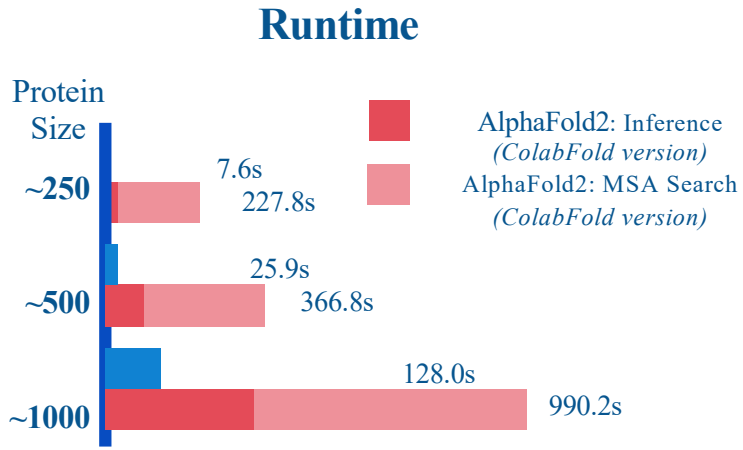


Geoformer 的通过约束高维空间的表示和三维空间的表示的一致性，来减少损失的信息

OmegaFold实验结果

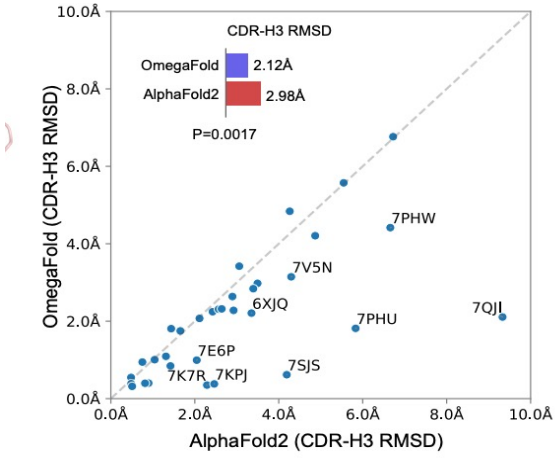


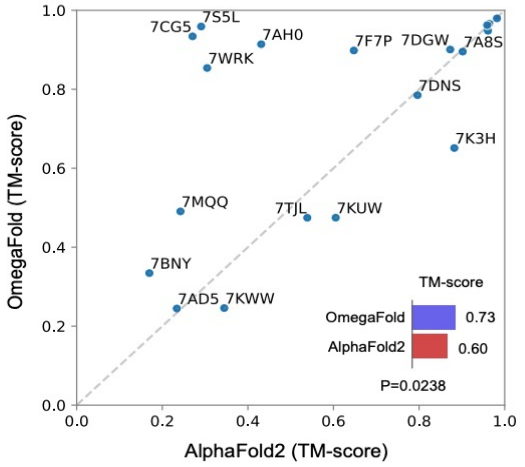




OmegaFold方法优于不使用MSA的AlphaFold2方法
和使用MSA的AlphaFold2方法性能相当

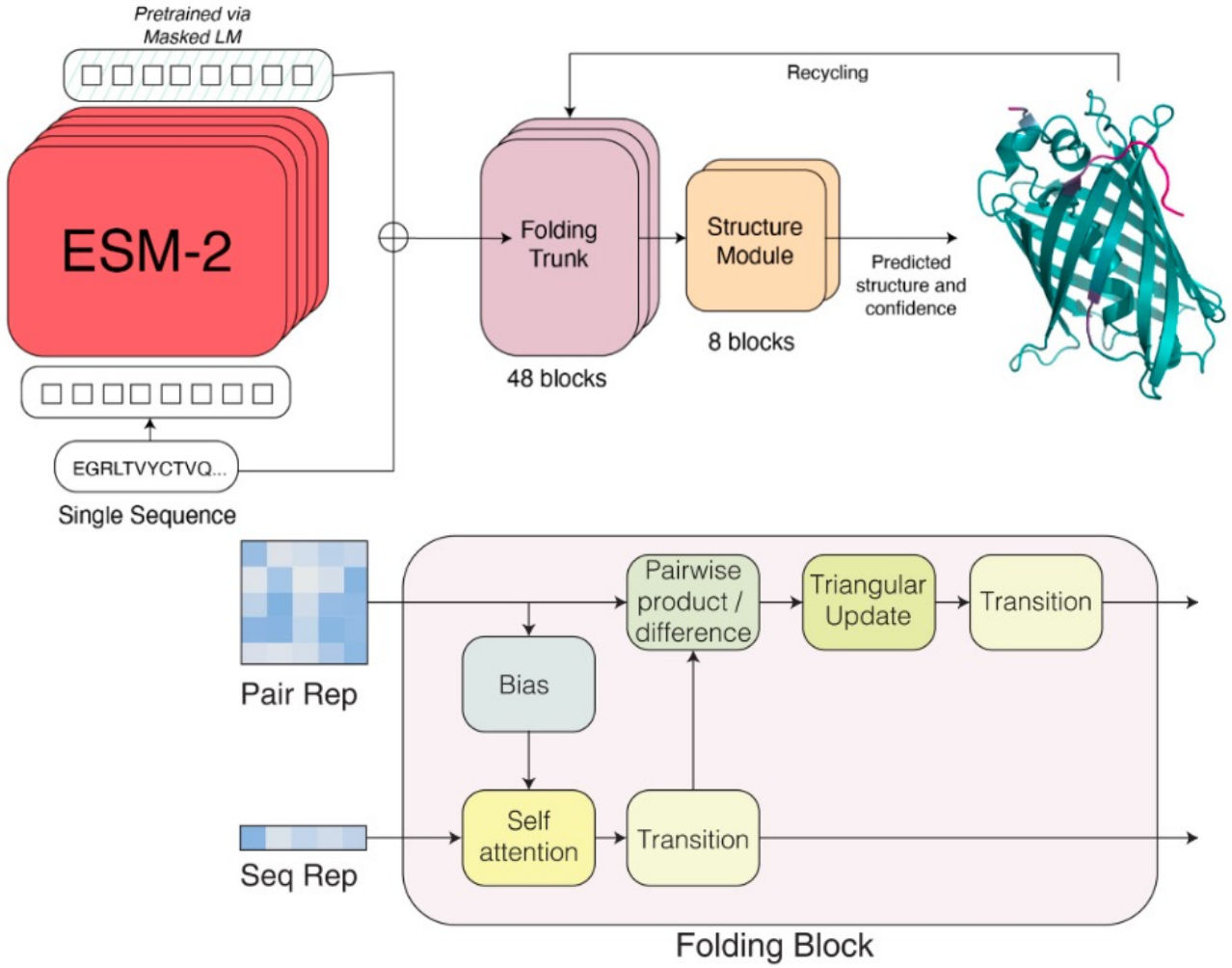
OmegaFold相比于AlphaFold2在
推理速度上具有优势



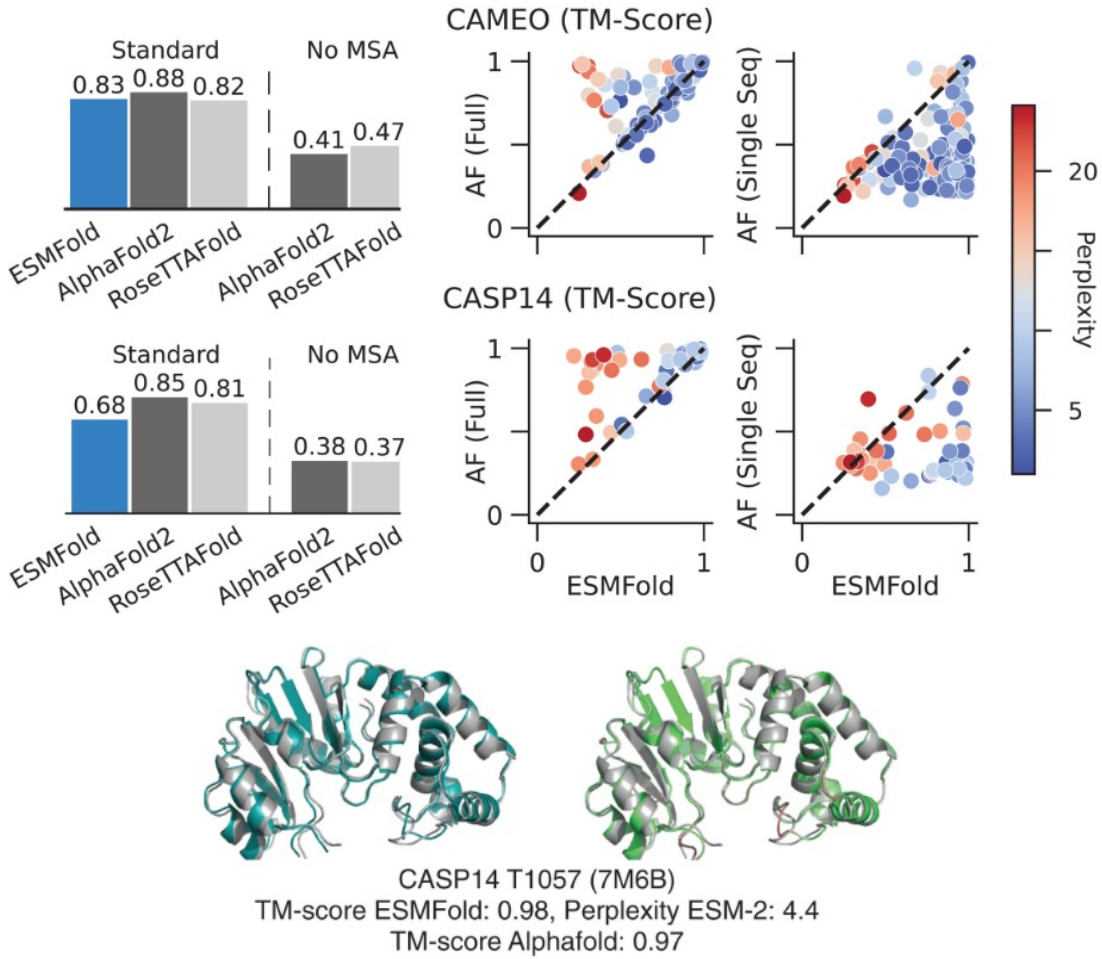


OmegaFold在抗体和孤儿蛋白上大幅度超越AlphaFold2

ESMFold单序列结构预测模型



基于ESM-2的单序列结构预测模型ESMFold



CASP14 T1057 (7M6B)
TM-score ESMFold: 0.98, Perplexity ESM-2: 4.4
TM-score Alphafold: 0.97

有MSA条件下，ESMFold与AlphaFold2相当

无MSA条件下，ESMFold大幅度领先AlphaFold2

Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023. (Meta)

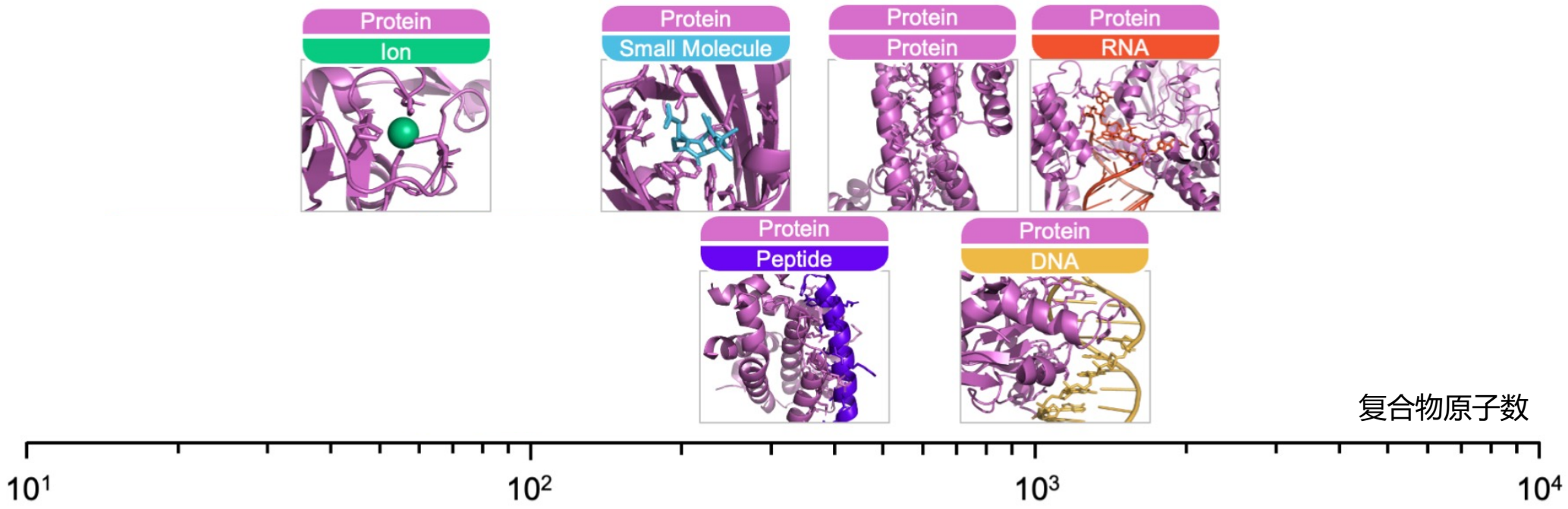
复合物结构预测

单体蛋白质结构预测

单链蛋白质如何折叠?

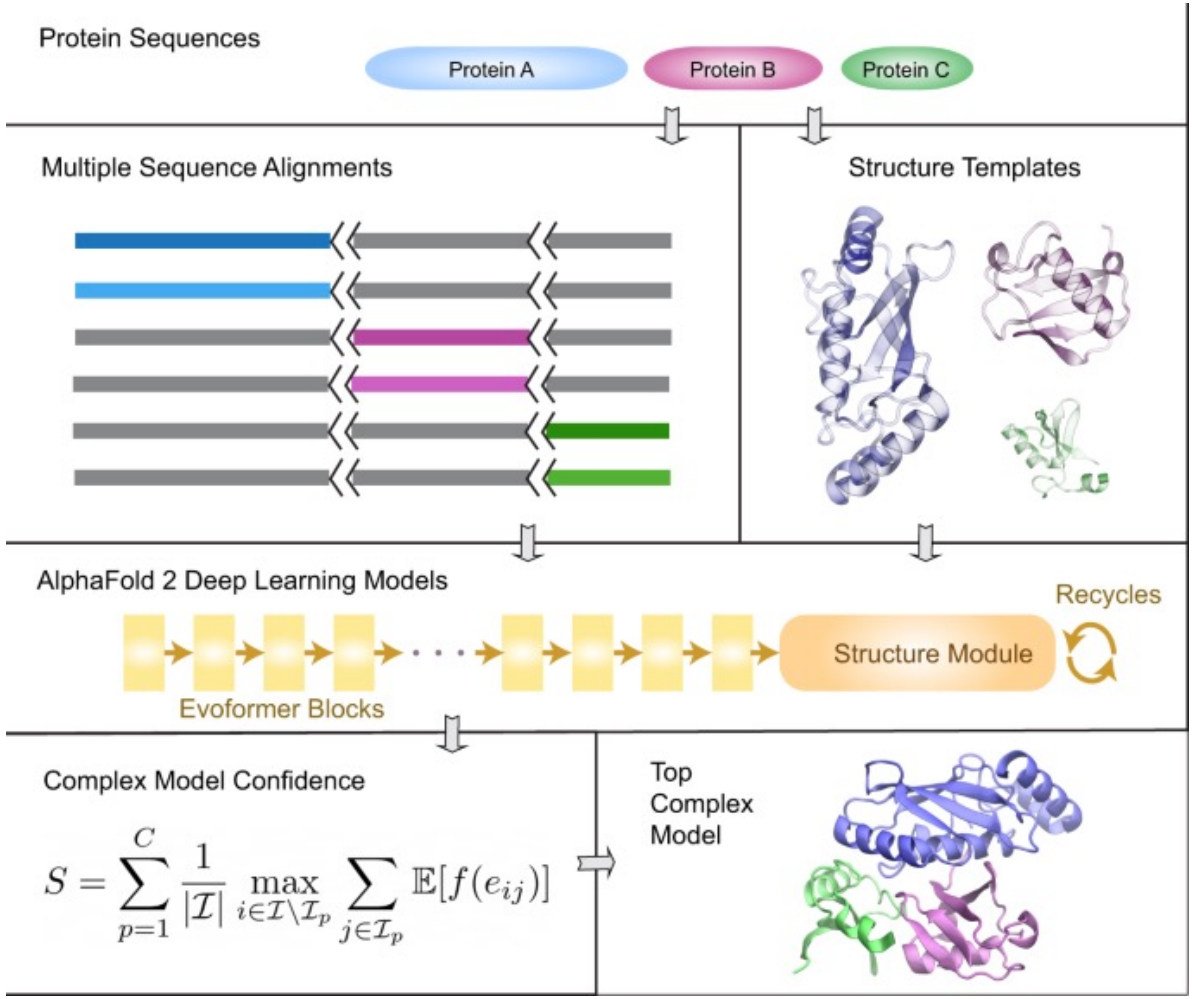
蛋白质复合物结构预测

蛋白质与配体如何结合?



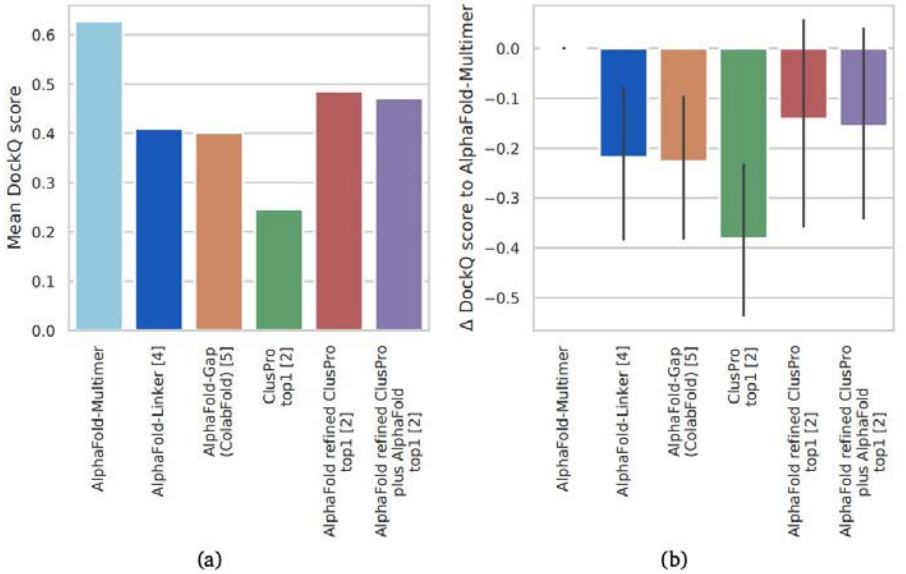
常见配体: 金属离子、小分子、多肽、蛋白质、DNA、RNA

蛋白复合物结构预测模型AlphaFold-Multimer

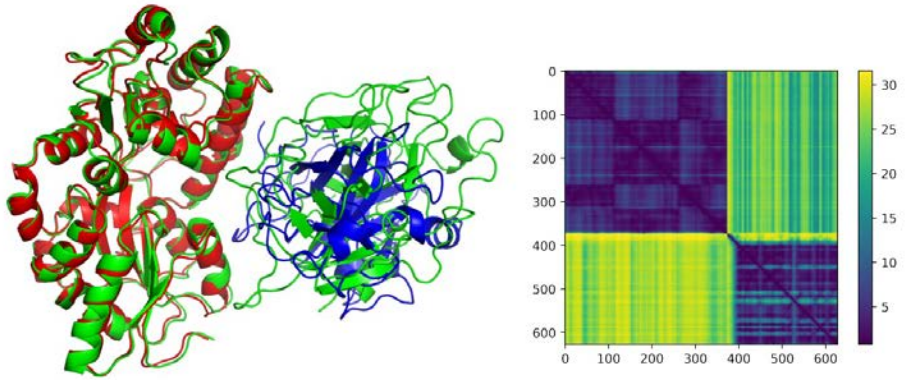


模型示意图

*注：此为AF2Complex模型结构图，AlphaFold-Multimer流程与此一致



AlphaFold-Multimer 在异源二聚体复合物结构预测上性能领先

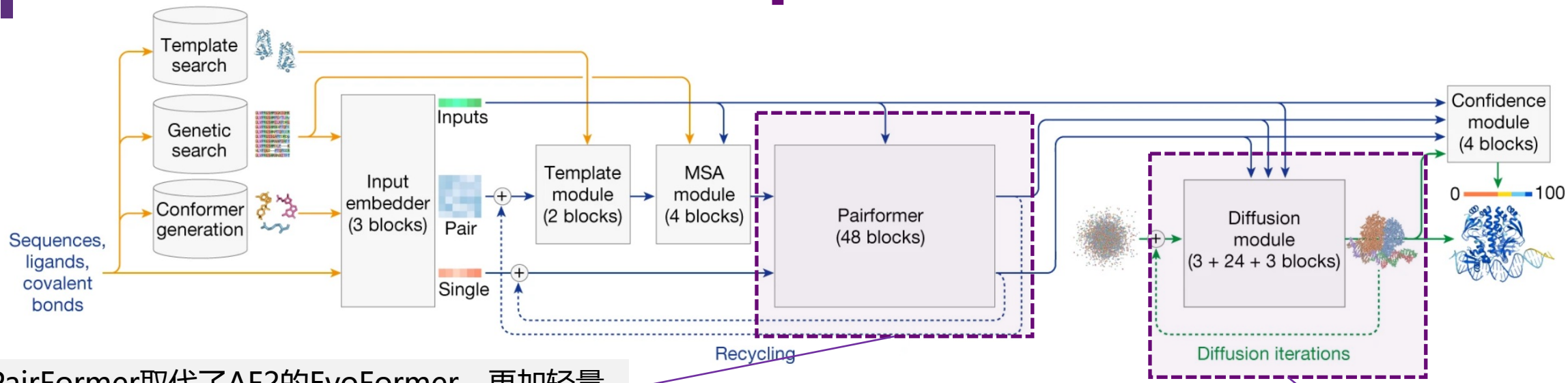


TM-score = 66.9, DockQ = 0.035, N_{res} = 628, PDB ID = 6QF7

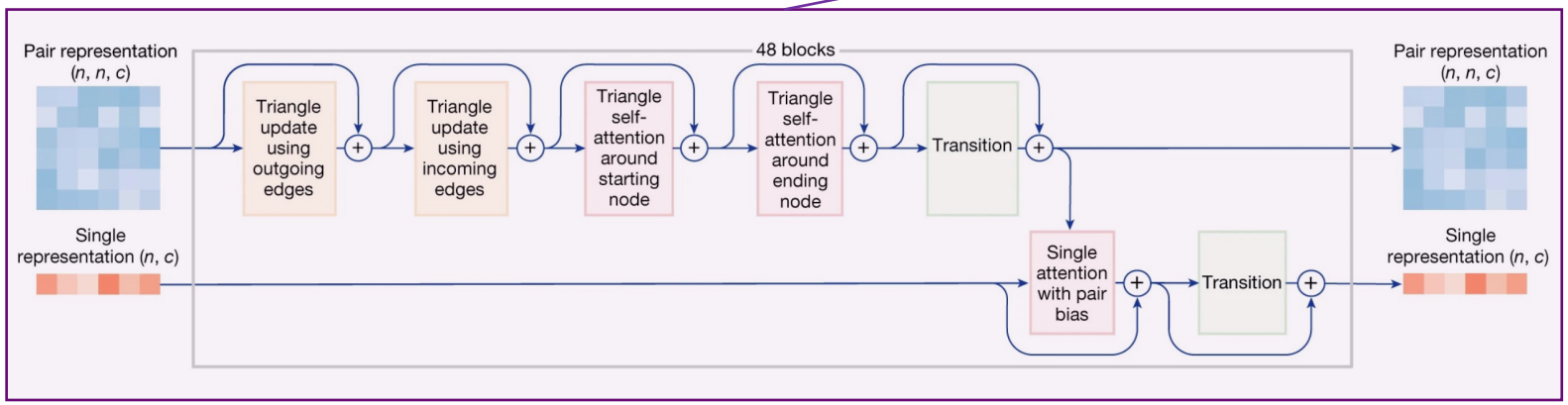
模型输出预测对齐误差PAE用于指示残基结合置信度

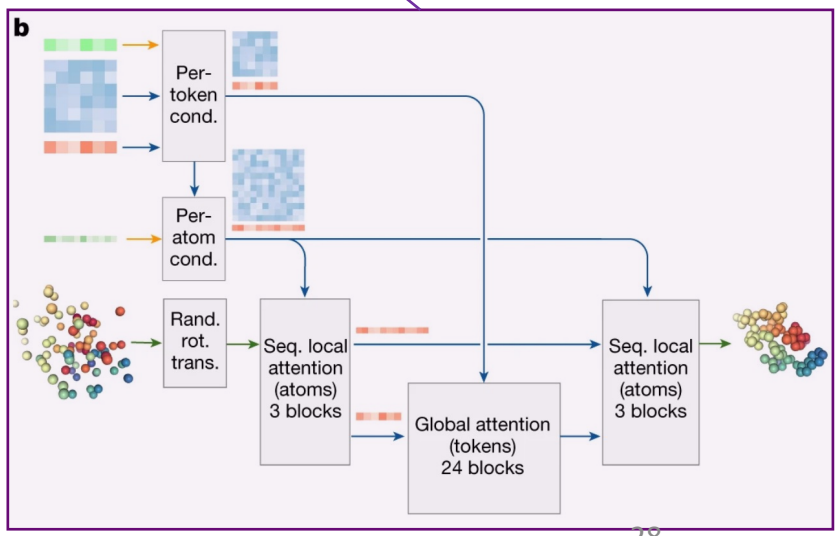
Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2023. (Google DeepMind)

复合物结构预测模型AlphaFold3



PairFormer取代了AF2的EvoFormer，更加轻量



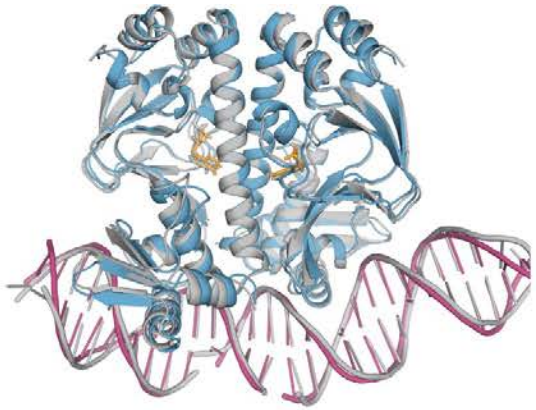


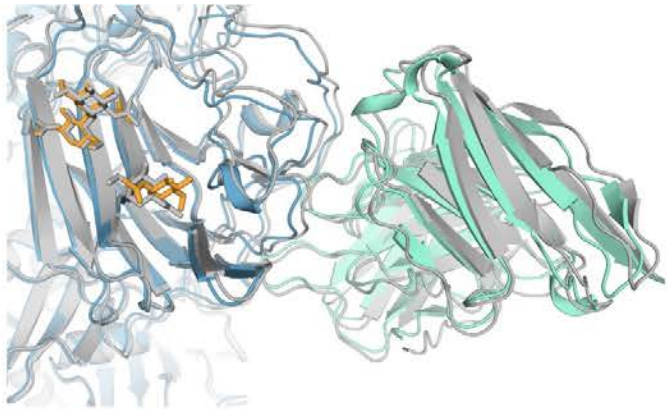
扩散模块直接处理原始原子坐标和结构表征，
取代了AF2的rotational frames和等变处理

Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024. (Google DeepMind)

AlphaFold3实验结果

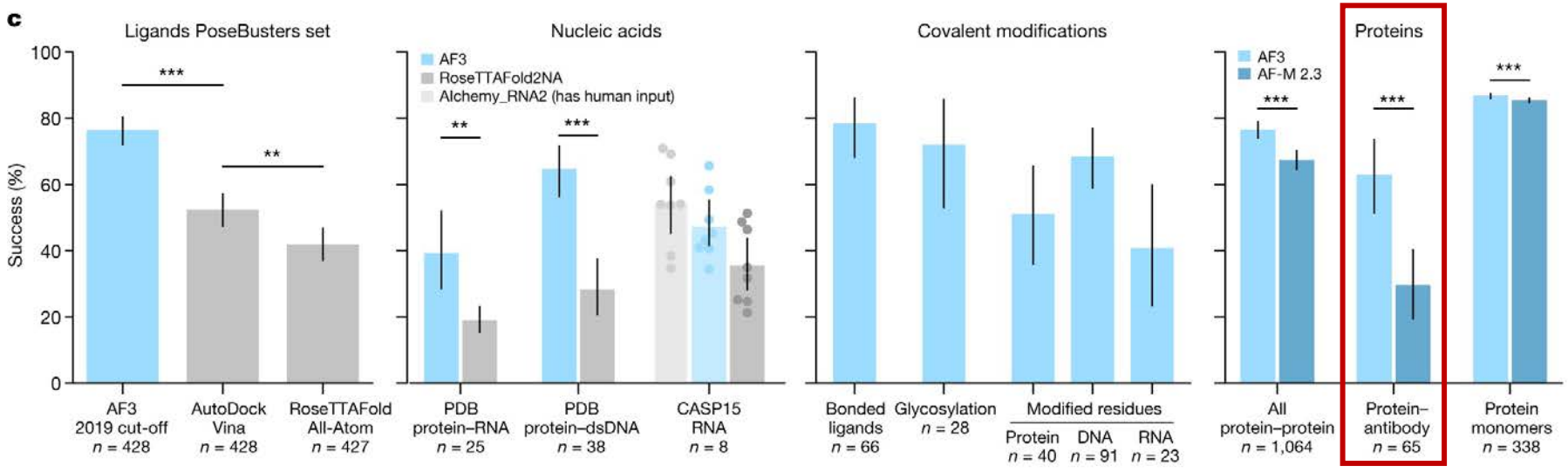
细菌CRP/FNR家族转录调控蛋白与DNA以及cGMP的复合物结构





人类冠状病毒OC43刺突蛋白，4,665个残基的糖基化蛋白与中和抗体结合

AlphaFold 3性能比传统方法提升50%，达到或超过了基于物理的分子动力学模拟的方法



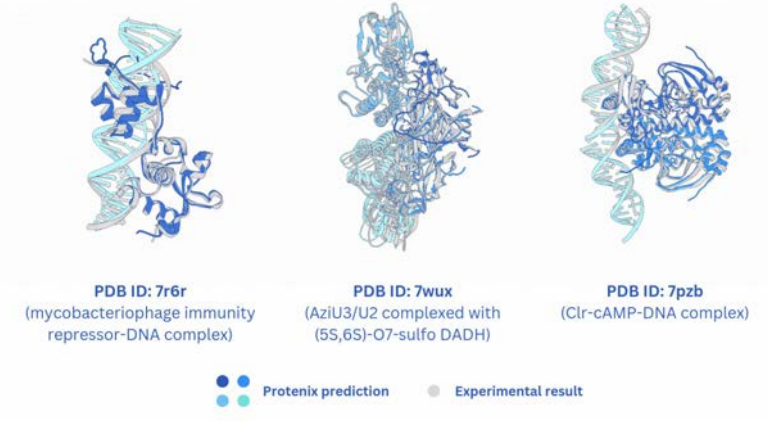
小分子　核酸　共价修饰　蛋白复合物

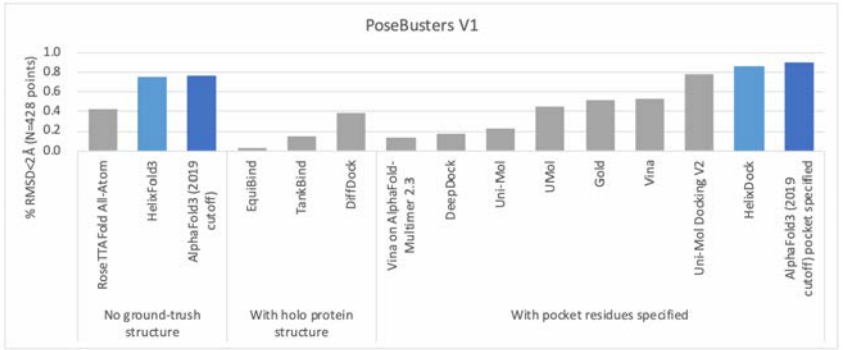
AF3对蛋白质-蛋白质预测有提升，特别在抗体-蛋白质相互作用的预测准确率达62.9%；对蛋白质单体LDDT的预测则有显著改善

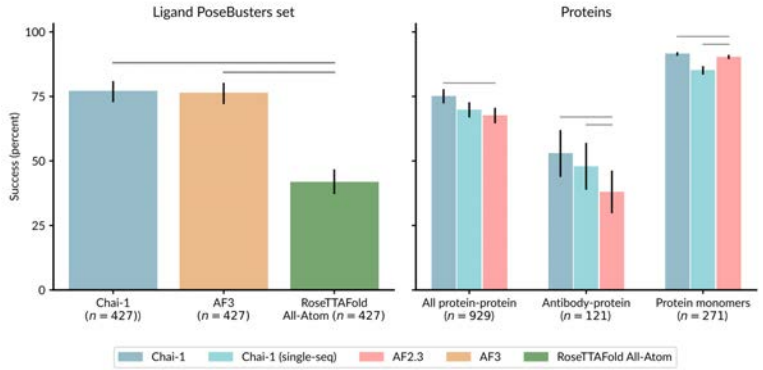
AlphaFold3能够高精度预测包含几乎所有分子类型的复合物结构

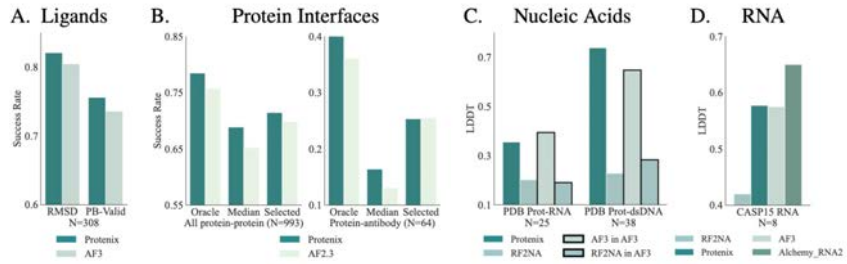
开源AlphaFold3

飞桨 PaddleHelix









HelixFold3

百度

公开推理代码、模型权重

Chai-1

Chai Discovery

公开推理代码、模型权重

Protenix

字节跳动

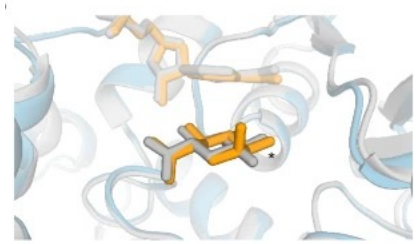
公开推理代码、模型权重、训练代码

Technical Report of HelixFold3 for Biomolecular Structure Prediction. 2024

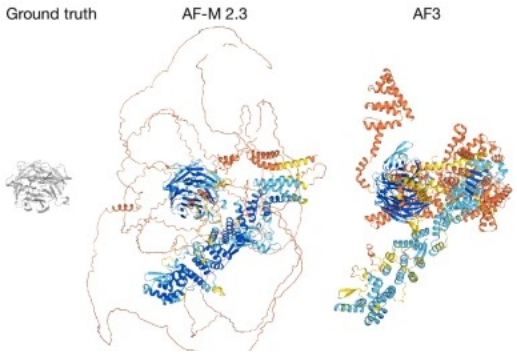
Chai-1: Decoding the molecular interactions of life. 2024

Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. 2025

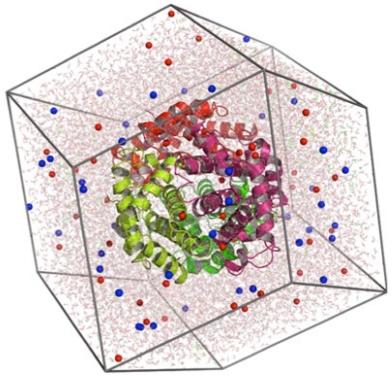
待解决的问题和发展趋势



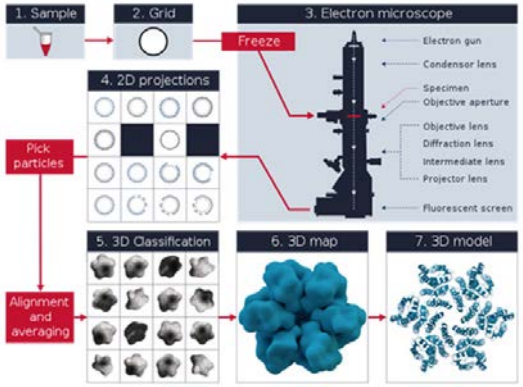
物理合理性缺陷
手性错误、原子重叠



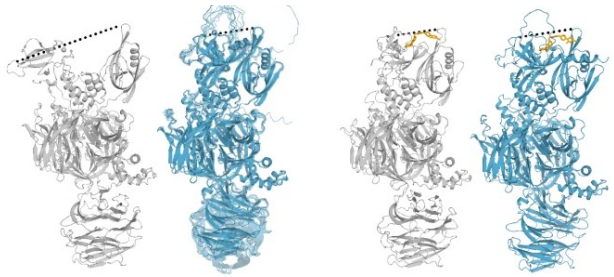
幻觉效应
无序区结构过度折叠



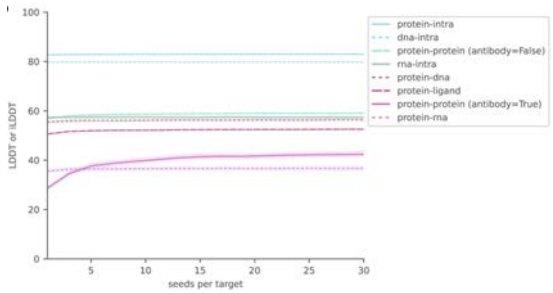
融合物理原理
引入分子动力学模拟等
增强物理感知



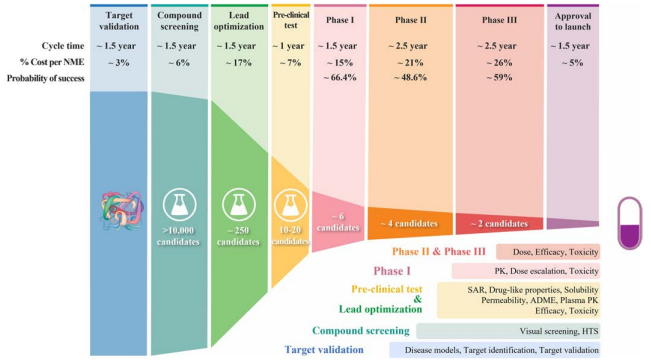
动态与多构象建模
整合冷冻电镜数据，捕捉蛋白质动态行为



动态构象、多构象难题
难以模拟蛋白质动态构象



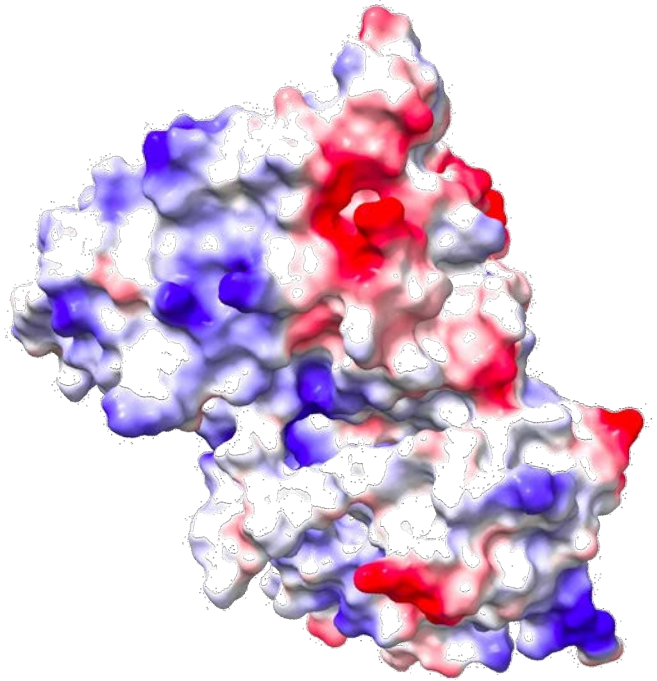
特定结构预测困难
核酸复合物训练数据少、抗体预测依赖大量采样



与实际应用结合
推动AI辅助药物设计（AIDD）的实际应用

*注：每个方向下具体方法仅作为举例示意

基于结构的药物发现

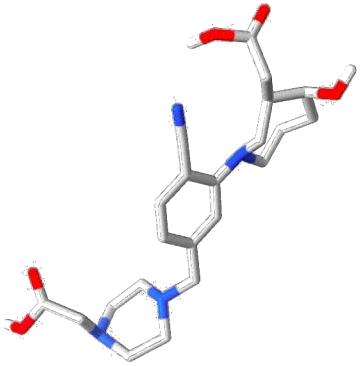


基于靶点的虚拟筛选

从已有化合物库中寻找可能结合的分子

优点：分子性质可控，没有合成问题

缺点：不是全新分子

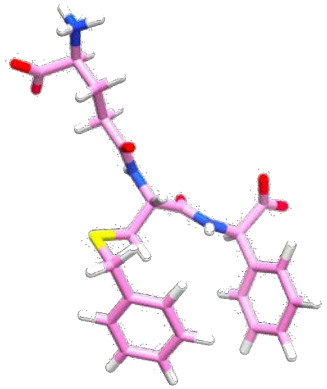


基于靶点的分子生成

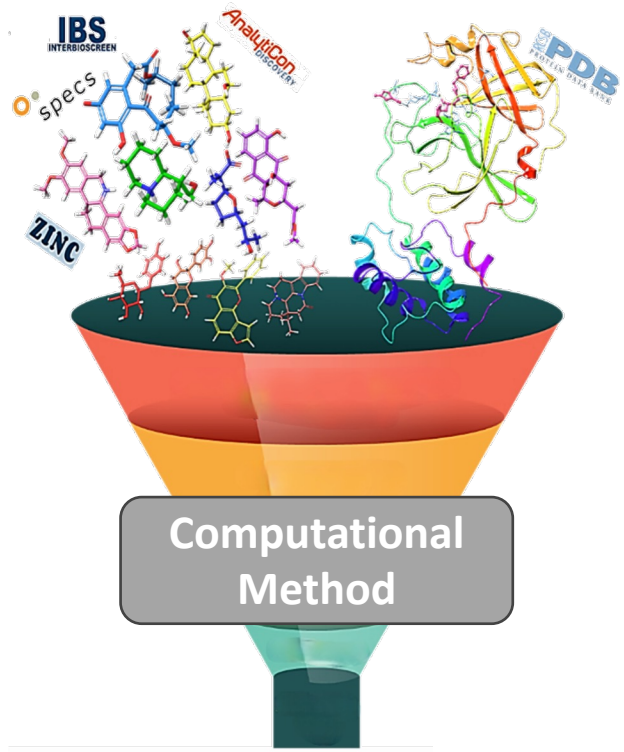
从头设计或者改造已有分子，从而得到可能结合的新分子

优点：生成全新分子，探索化合物空间

缺点：新分子合成困难



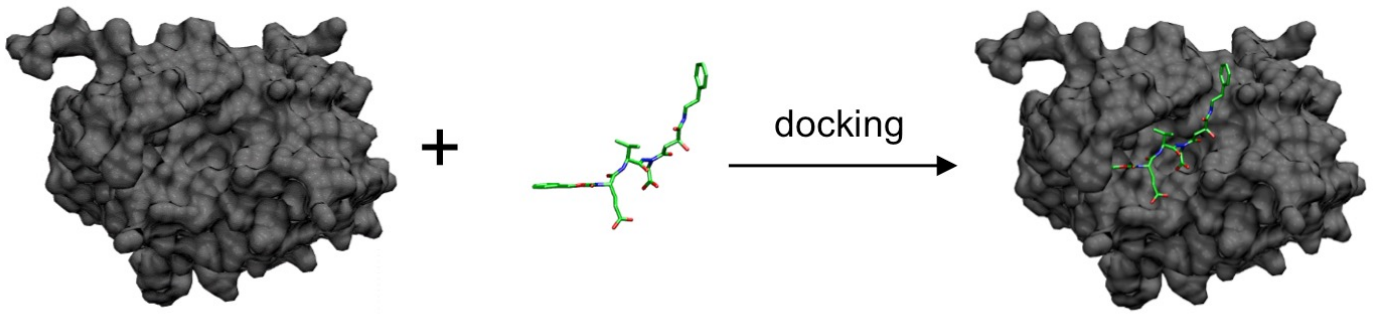
虚拟筛选



虚拟筛选（Virtual Screening）

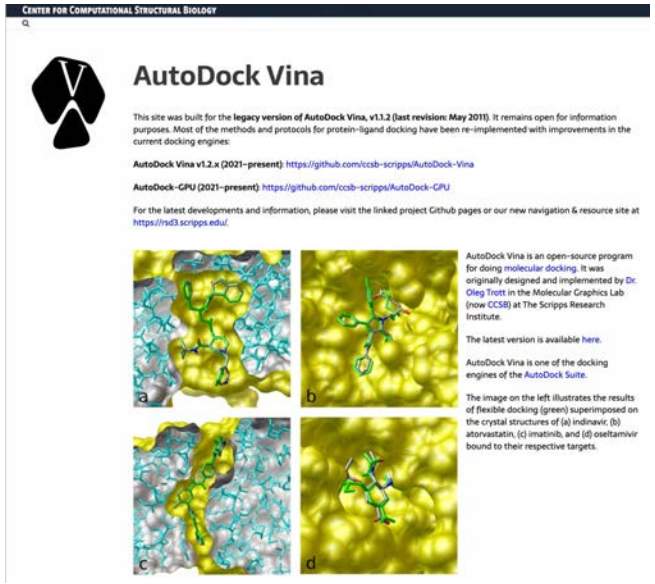
通过搜索分子库，识别能够与蛋白质🔒结合的关键分子🔑。

分子对接：传统虚拟筛选方法

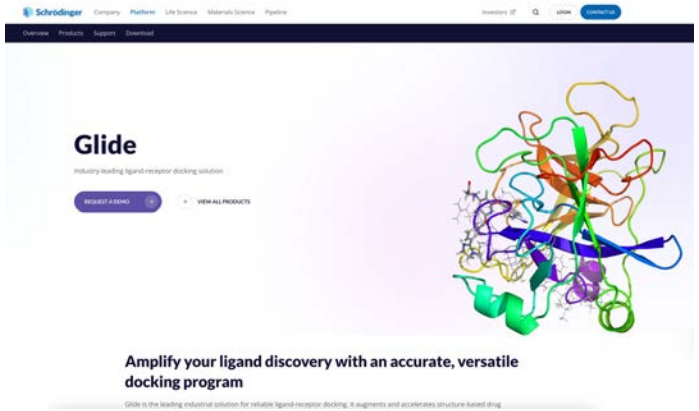


- **分子对接（Docking）**：通过预测小分子与蛋白质靶点的结合位点和结合方式（构象），从而评估其亲和力和生物活性。
- **方法流程：**基于分子对接的虚拟筛选，需要先预测对接构象，然后基于打分函数对分子进行排序 。
- **缺陷：**对接方法速度慢，每个分子都需进行大量构象采样和物理打分计算，计算复杂且难以并行化。

常见分子对接软件



AutoDock Vina

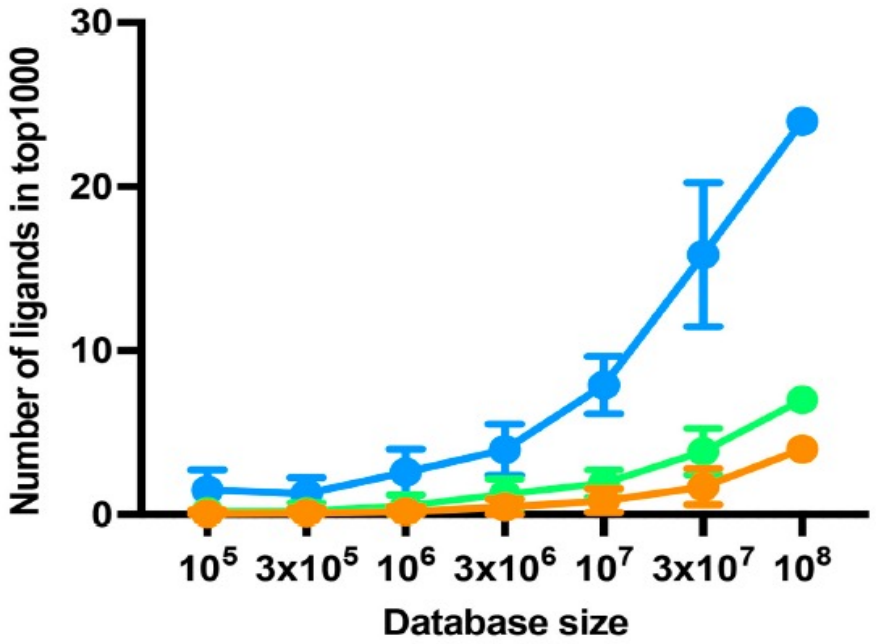


Schrodinger Glide

大规模药物虚拟筛选的意义与困难

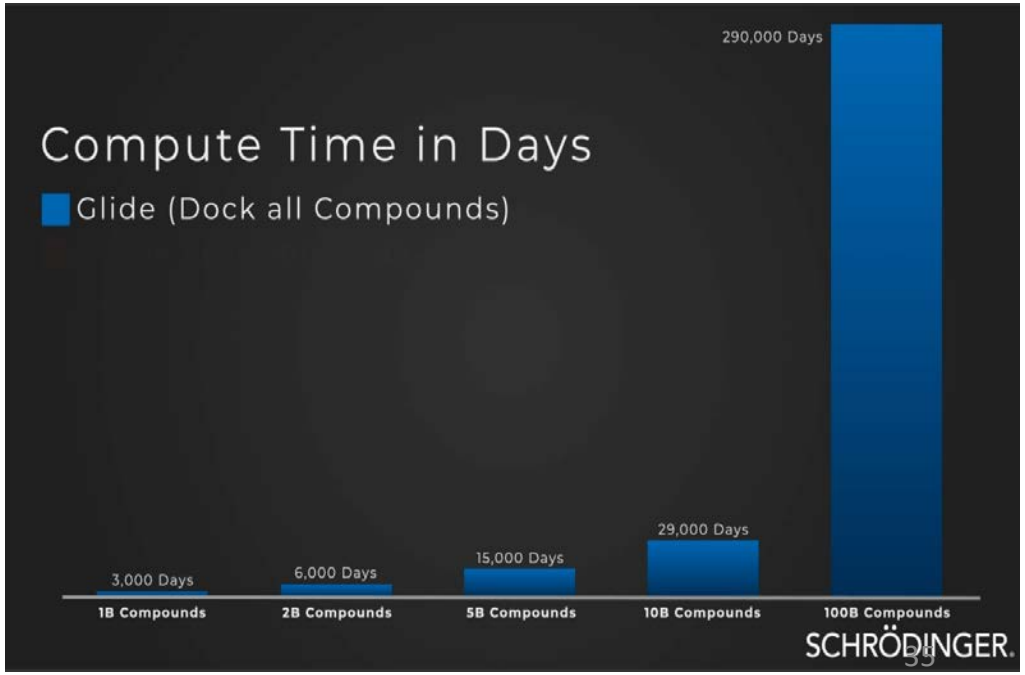
为什么需要大规模虚拟筛选?

Nature正刊论文统计：当候选分子库从10万上升到1亿，通过虚拟筛选得到的最优1000个分子中，有效配体数量提升10倍以上（1-2个提升到了20个以上）



基于对接的大规模虚拟筛选是否可行?

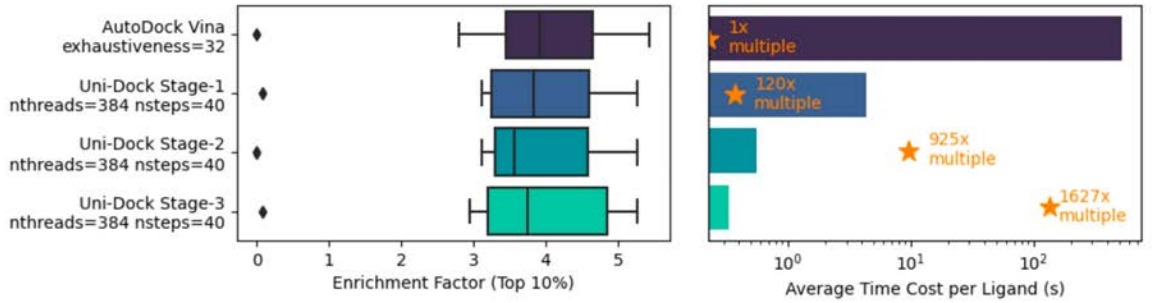
当前主流的商用虚拟筛选软件**Glide**, 筛选10亿分子需要将近10年的时间

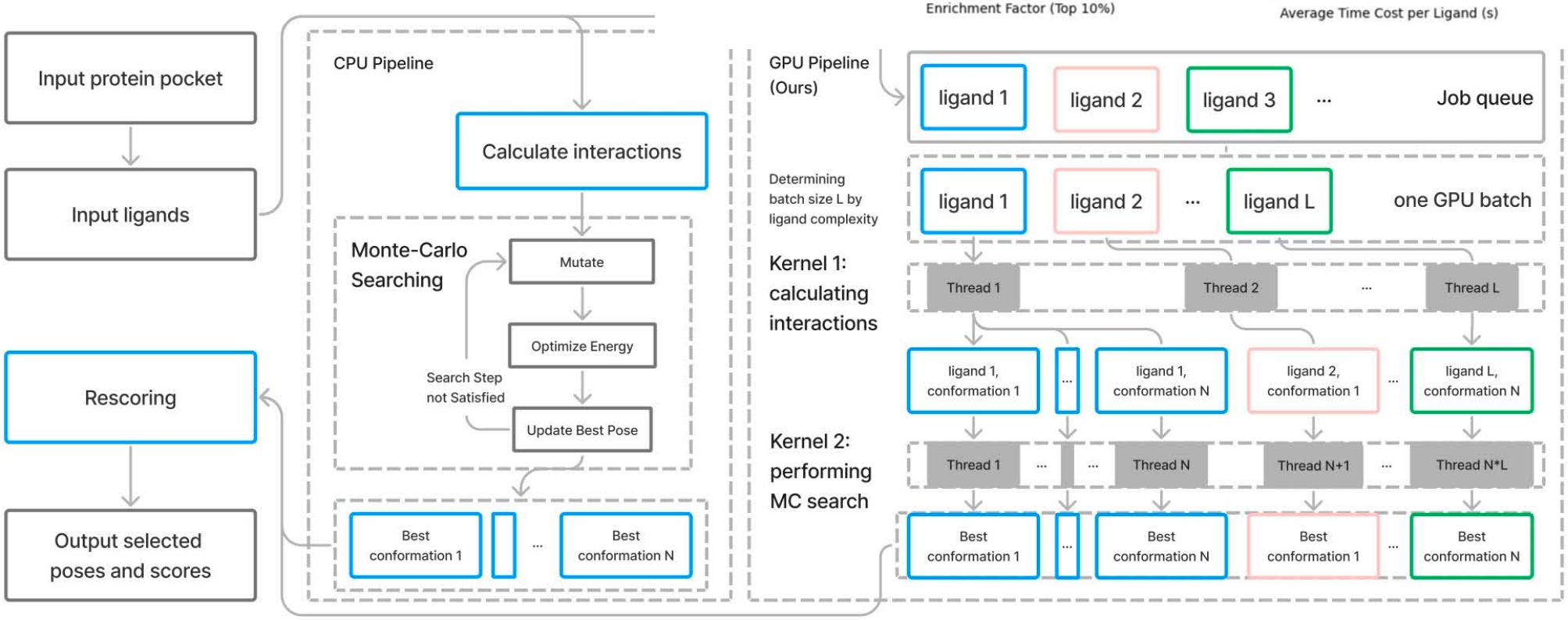


基于GPU加速对接的虚拟筛选

GPU加速模拟对接速度

- Uni-Dock实现1600倍加速



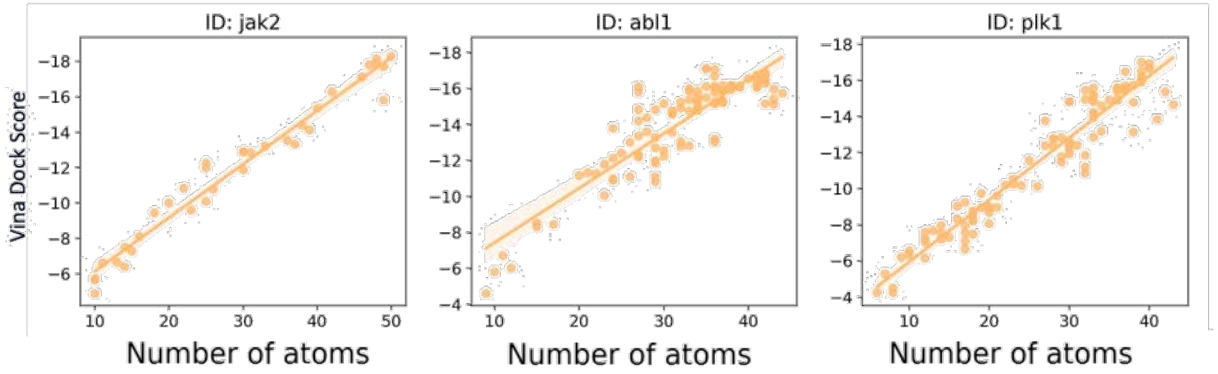


Uni-Dock: GPU-accelerated docking enables ultralarge virtual screening. Journal of chemical theory and computation, 2023. (DP Technology)

分子对接的问题

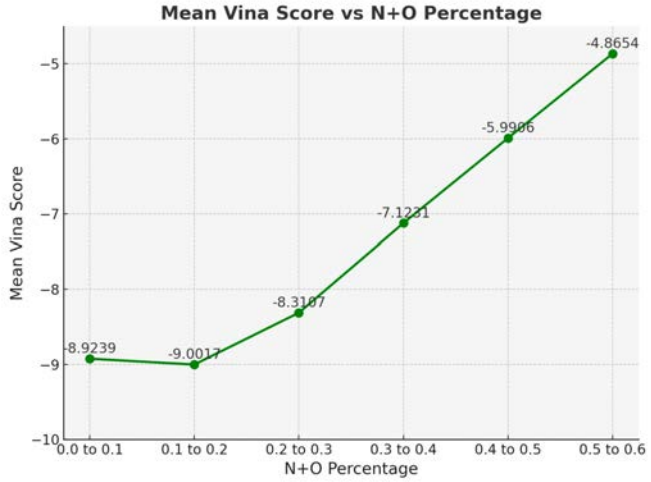
主要问题：

- 打分依赖经验项
- 容易错误偏好特定分子

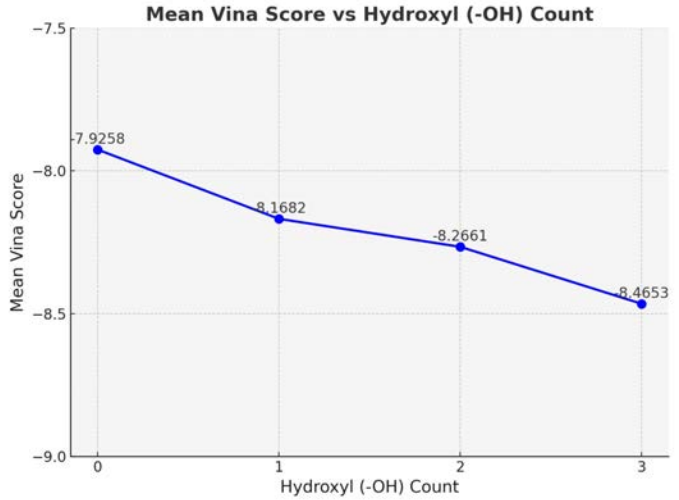


$$\text{GlideScore} = a \cdot \text{vdW} + b \cdot \text{Coul} + \text{Lipo} + \text{Hbond} + \text{Metal} + \text{Rewards} + \text{Penalties}$$

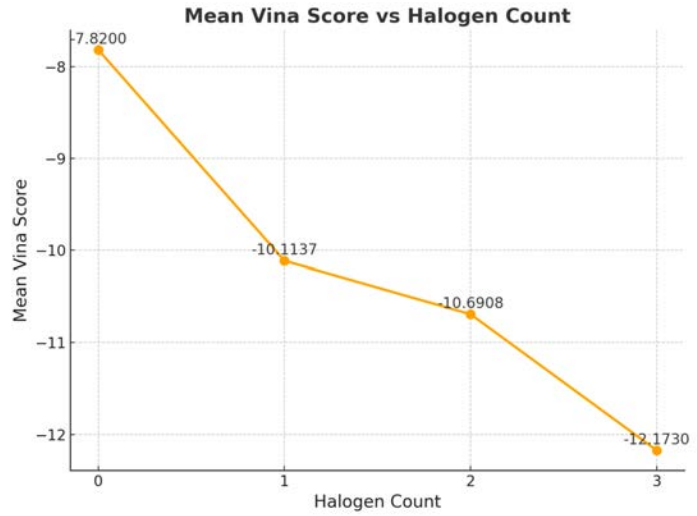
Bias 1: 分子越大docking score越好



Bias 2: N+0比例越小docking score越好



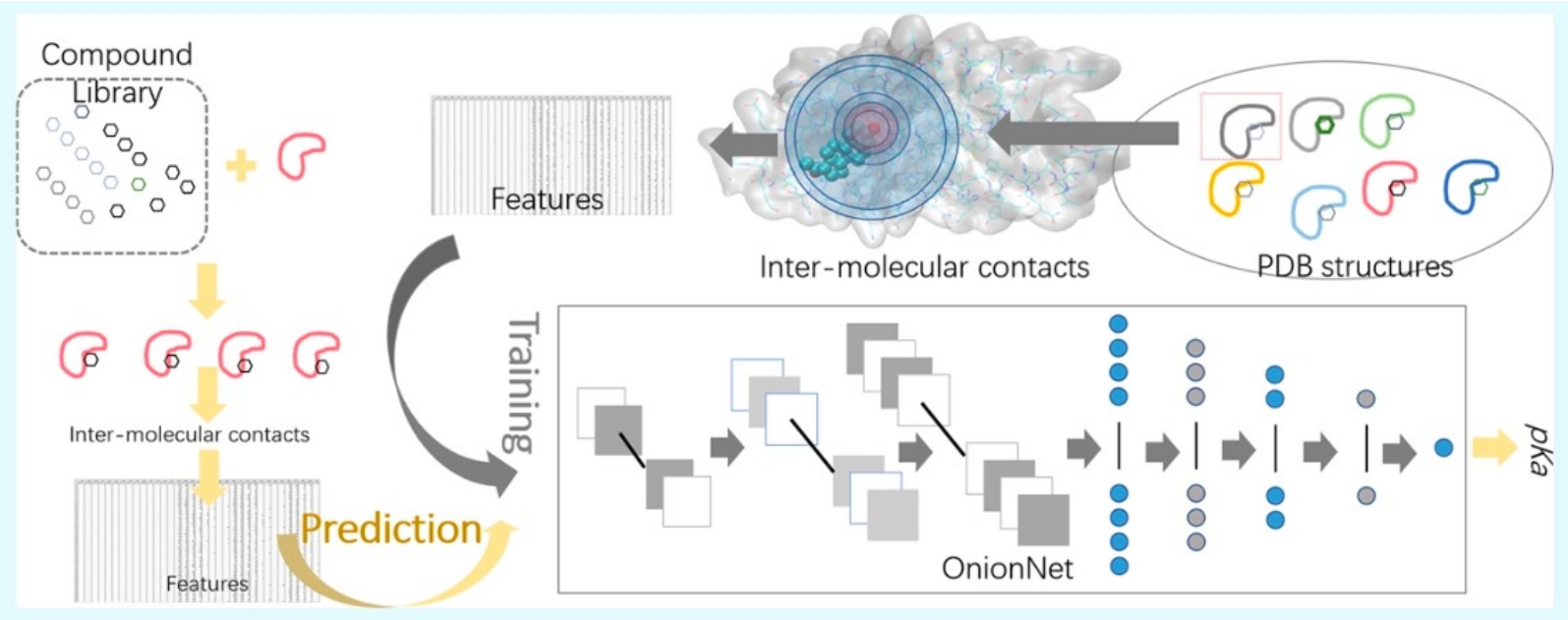
Bias 3: 羟基越多docking score越好

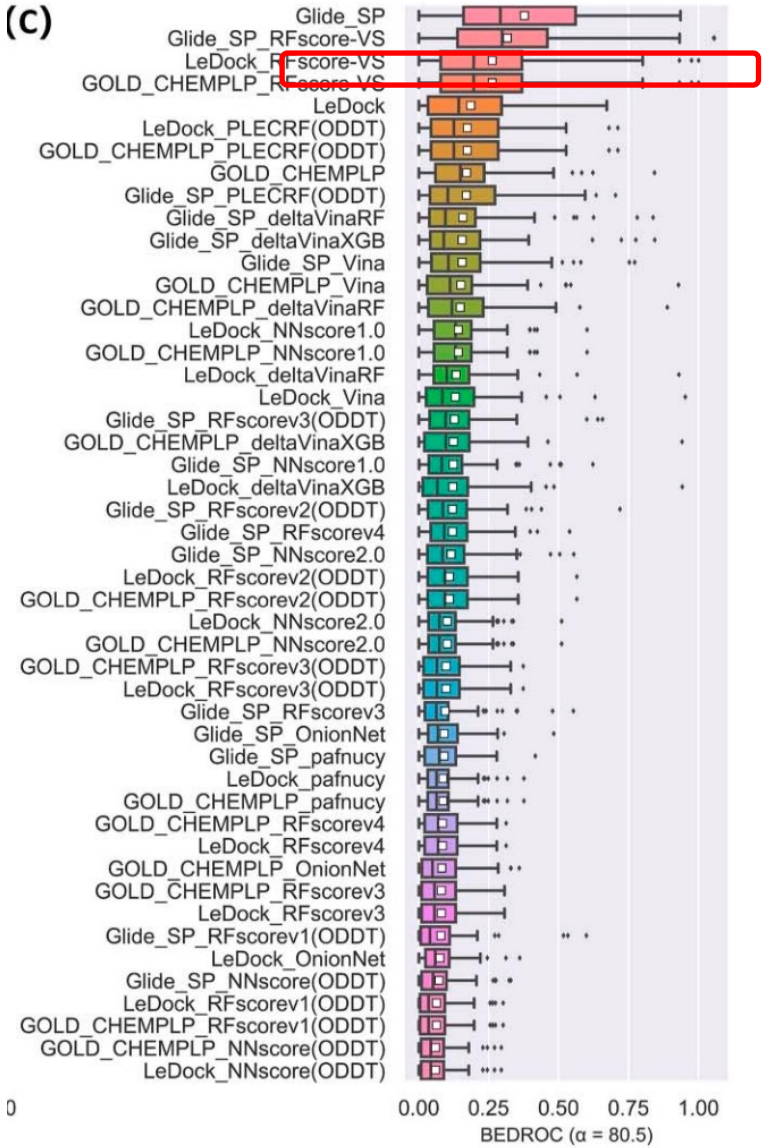


Bias 4: 卤素越多docking score越好

基于深度学习的药物虚拟筛选

改进方案一：依赖分子对接构象，基于深度学习的打分替换对接软件的打分函数。通过建模口袋-分子之间的相互作用预测结合能。





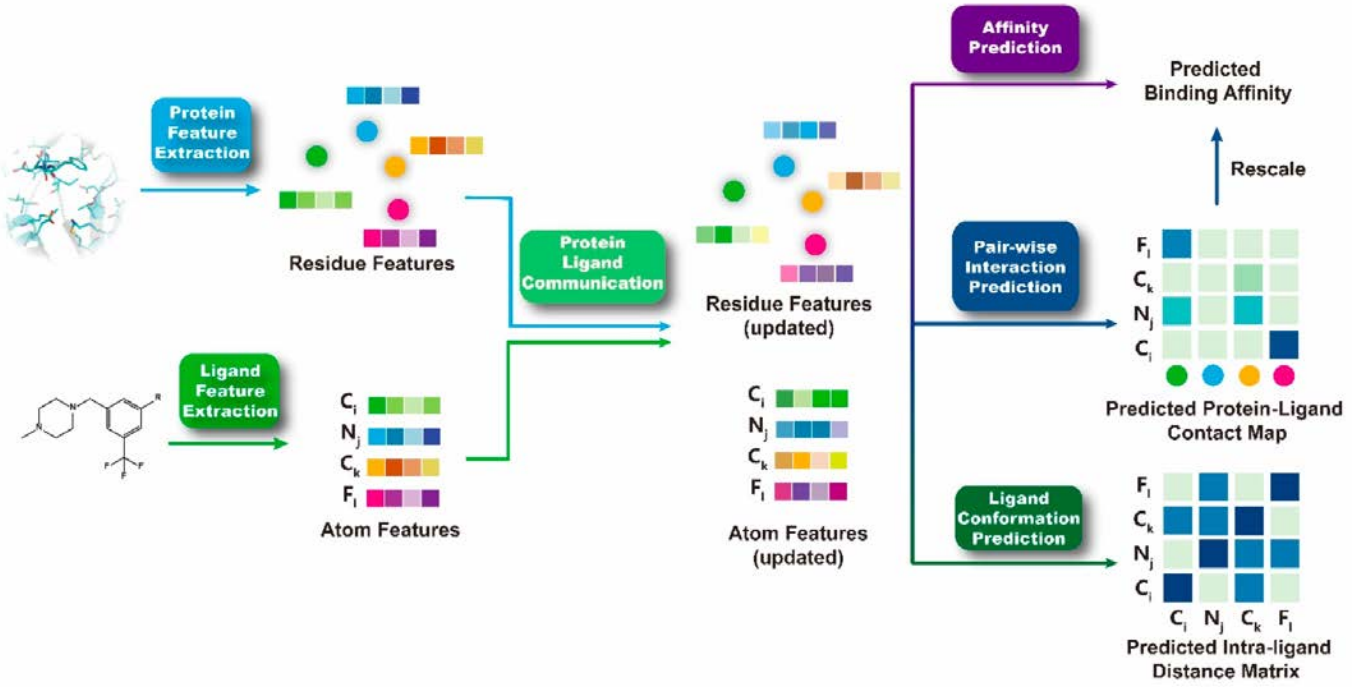
速度慢，效果无法超越商用软件Glide

OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. ACS omega, 2019. (NTU)

基于深度学习的药物虚拟筛选

改进方案二：不依赖于对接构象，直接预测靶点和分子的结合能

- 速度上有所提升，不需要对接。但是依然需要大量复杂神经网络前向计算
- 效果依然不及商用软件Glide

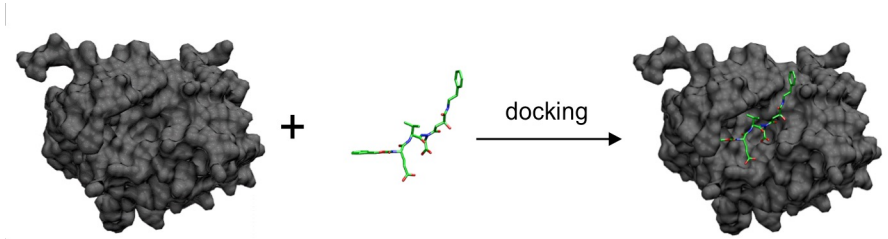


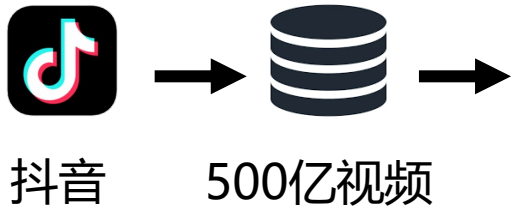
Model	AUC ROC		$EF^{0.5\%}$	$EF^{1\%}$	$EF^{5\%}$
	mean±std	median	mean	mean	mean
PLANET	0.761±0.123	0.770	10.234	8.832	5.402
$\Delta_{Vina}RF_{20}$ [a]	0.697±0.102	0.695	9.516	7.995	4.382
RFscore-V4 [a]	0.652±0.108	0.655	4.902	4.521	2.981
NNscore2.0 [a]	0.683±0.098	0.683	4.162	4.022	3.123
Pafnucy [a]	0.631±0.124	0.639	4.241	3.861	2.767
OnionNet [a]	0.597±0.102	0.606	2.840	2.840	2.205
Glide SP [a]	0.767±0.116	0.783	19.389	16.182	7.231
Vina [b]	0.716	N.A.	9.139	7.321	4.444

PLANET: A Multi-objective Graph Neural Network Model for Protein–Ligand Binding Affinity Prediction. JCIM, 2023. (Fudan)

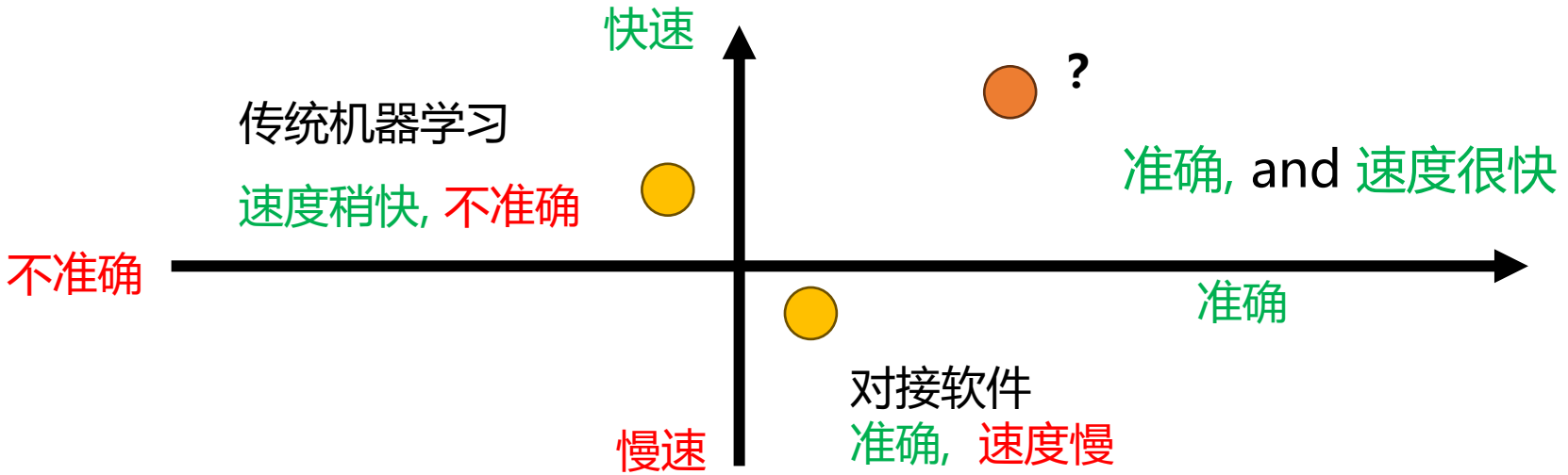
更快更准的虚拟筛选方法

对接软件：预测分子和蛋白结合后的构象





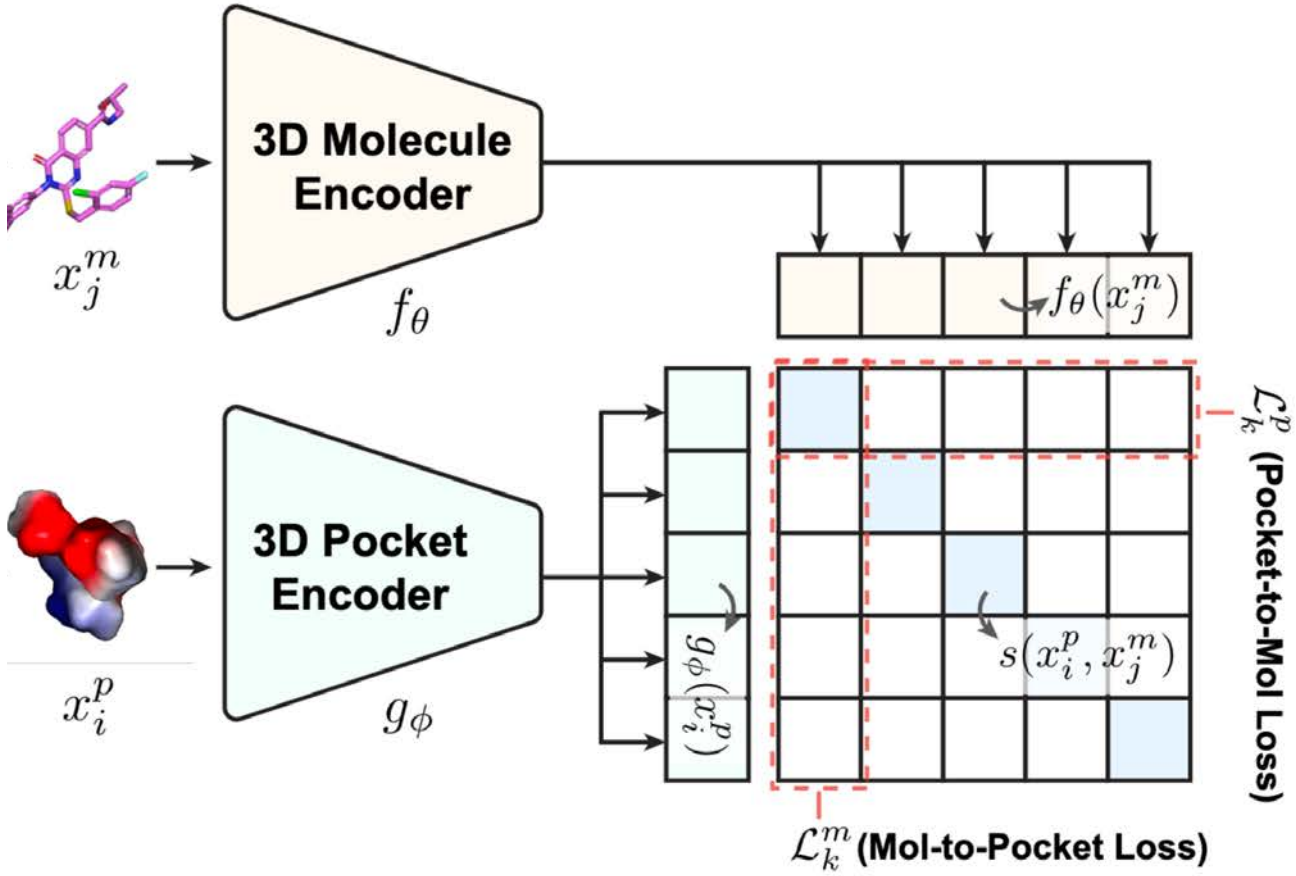
是否可以设计算法快速为蛋白质**推荐**小分子?



基于对比学习的虚拟筛选方法DrugCLIP

对比学习框架

通过构建batch内负例，最大化正样本的相似性，最小化负样本的相似性



配体侧in-batch优化：

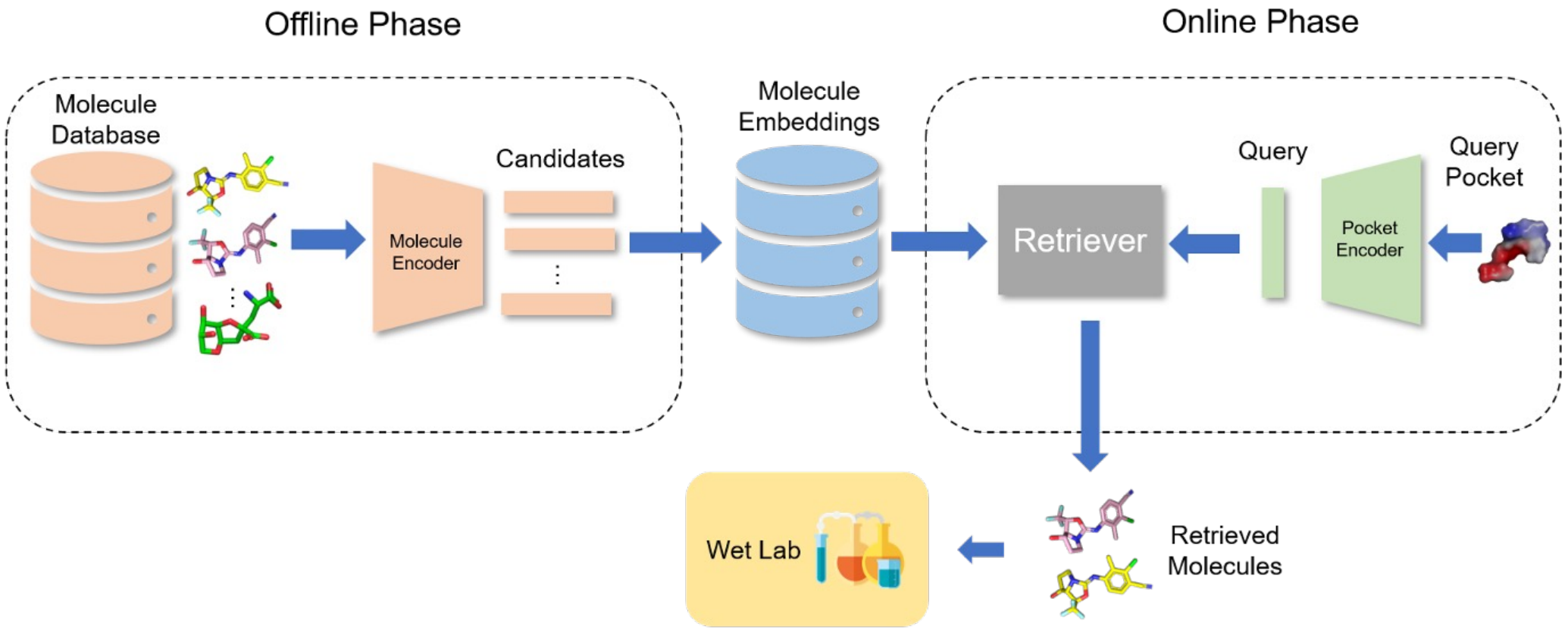
$$\mathcal{L}_k^p(x_k^p, \{x_i^m\}_{i=1}^N) = -\frac{1}{N}\log\frac{\exp(s(x_k^p, x_k^m)/\tau)}{\sum_i \exp(s(x_k^p, x_i^m)/\tau)}$$

口袋侧in-batch优化：

$$\mathcal{L}_k^m(x_k^m, \{x_k^p\}_{i=1}^N) = -\frac{1}{N}\log\frac{\exp(s(x_k^p, x_k^m)/\tau)}{\sum_i \exp(s(x_i^p, x_k^m)/\tau)}$$

DrugCLIP: Contrastive Protein-Molecule Representation Learning for Virtual Screening. NeurIPS, 2023 (Tsinghua AIR)

基于DrugCLIP稠密检索的虚拟筛选流程



离线分子特征提取

- 通过与训练分子编码器对分子库进行预编码
- 提取的分子特征可复用

在线口袋快速检索

- 对于用户查询口袋编码后执行向量化检索
- 只需点乘计算，速度极快

DrugCLIP虚拟筛选实验结果

DrugCLIP 在基准数据集 DUD-E 和 LIT-PCBA 上取得了卓越的表现，

超过了传统对接方法和深度学习方法

	AUROC (%)	BEDROC (%)	EF 0.5%	EF 1%	EF 5%
Glide-SP [11]	76.70	40.70	19.39	16.18	7.23
Vina [39]	71.60	-	9.13	7.32	4.44
NN-score [6]	68.30	12.20	4.16	4.02	3.12
RFscore [1]	65.21	12.41	4.90	4.52	2.98
Pafnucy [36]	63.11	16.50	4.24	3.86	3.76
OnionNet [50]	59.71	8.62	2.84	2.84	2.20
Planet [47]	71.60	-	10.23	8.83	5.40
DrugCLIP_{ZS}	**80.93**	**50.52**	**38.07**	**31.89**	**10.66**

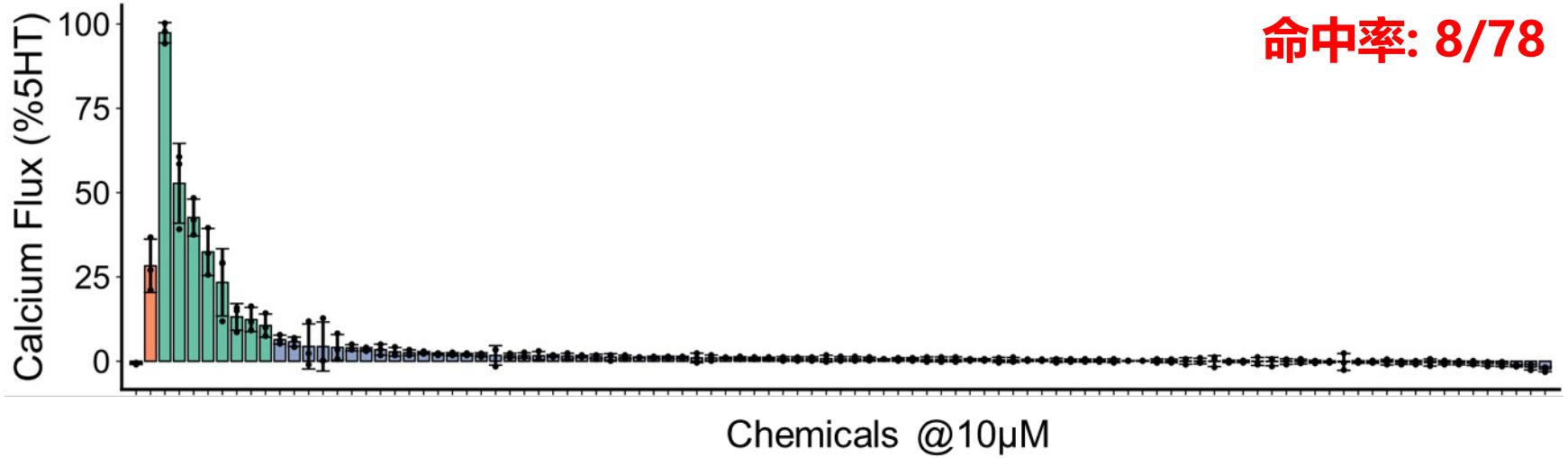
DUD-E数据集虚拟筛选结果

	AUROC (%)	BEDROC (%)	EF 0.5%	EF 1%	EF 5%
Surflex [35]	51.47	-	-	2.50	-
Glide-SP [11]	53.15	4.00	3.17	3.41	2.01
Planet [47]	57.31	-	4.64	3.87	2.43
Gnina [26]	**60.93**	5.40	-	4.63	-
DeepDTA [29]	56.27	2.53	-	1.47	-
BigBind [2]	60.80	-	-	3.82	-
DrugCLIP	57.17	**6.23**	**8.56**	**5.51**	**2.27**

LIT-PCBA数据集虚拟筛选结果

DrugCLIP湿实验结果：$5HT_{2A}R$激动剂

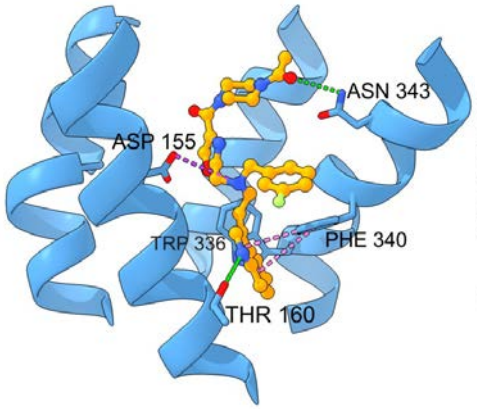
使用钙流试验测试化合物的激动剂活性

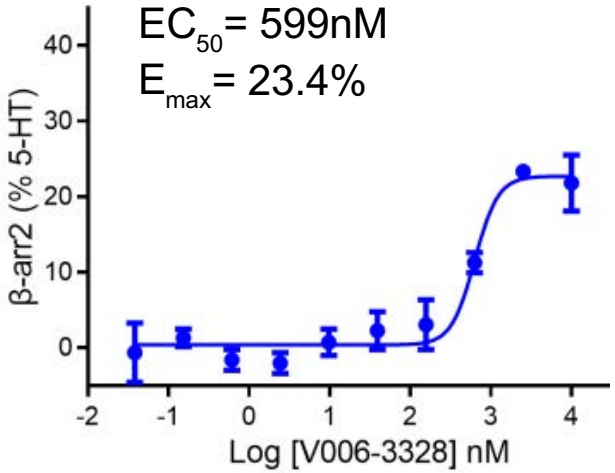


命中率: 8/78

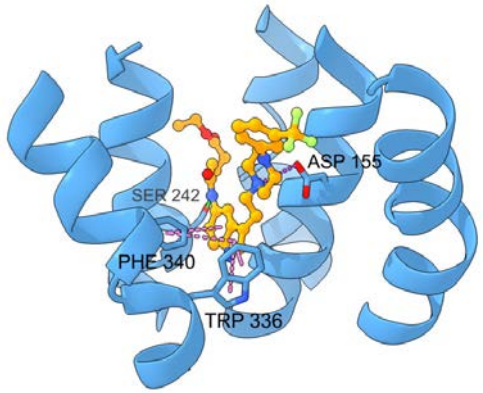
PDB: 7WC8

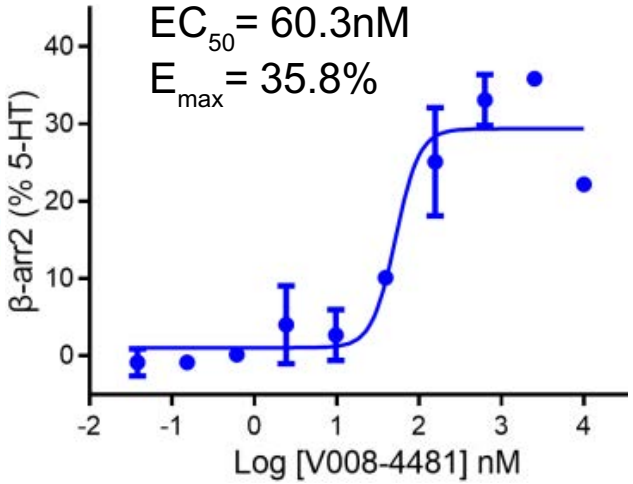
复合物结构对接预测，同位素标记配体竞争性结合试验与β-arrestin2结合试验





K_i = 3.51 μM

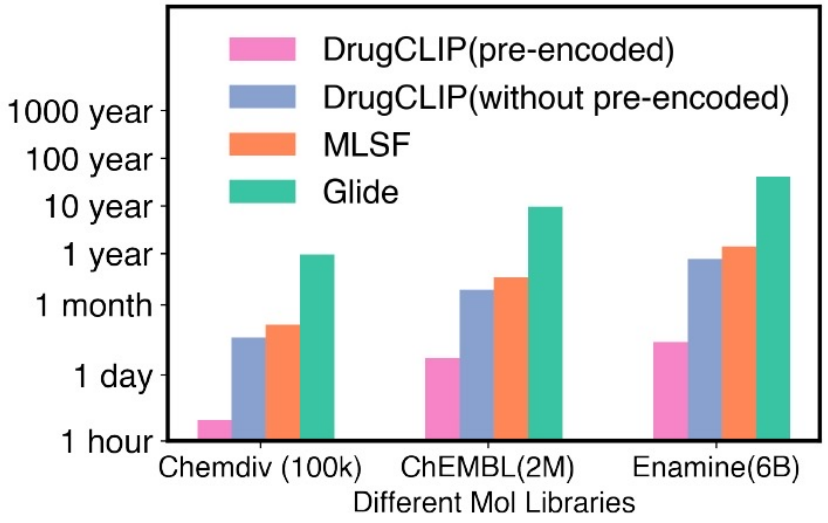




K_i = 21.0 nM

DrugCLIP虚拟筛选速度优势

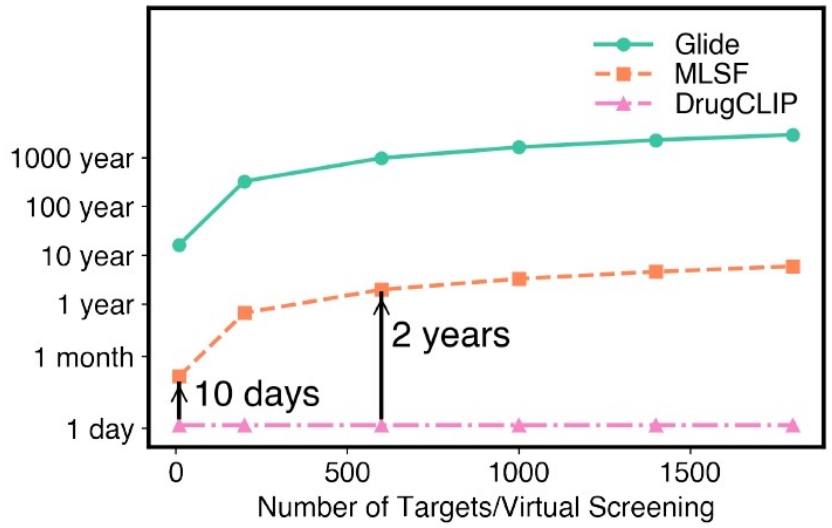
DrugCLIP 极大的减少了筛选所需速度，尤其是不同靶点多次筛选的情况下



在不同化合物库的测试结果

100 Years for Glide and

10 days for DrugClip



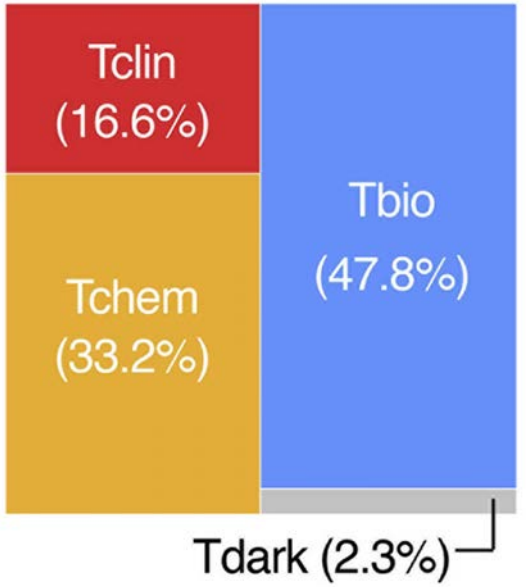
多次筛选下的时间

Fast to perform Multiple targets

DrugCLIP: Contrastive Protein-Molecule Representation Learning for Virtual Screening. NeurIPS, 2023 (Tsinghua AIR)

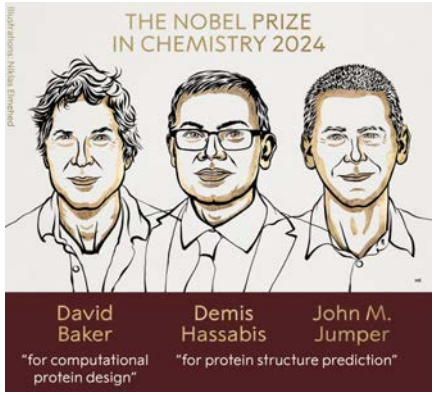
基于DrugCLIP的人类全基因组级别筛选

- 依托于CRISPR和AlphaFold2, 我们可以挖掘可能治疗疾病的靶点，并预测其结构
- 大部分靶点（>50%）缺少已知的结合的化合物



CRISPR:挖掘所有与疾病相关的基因

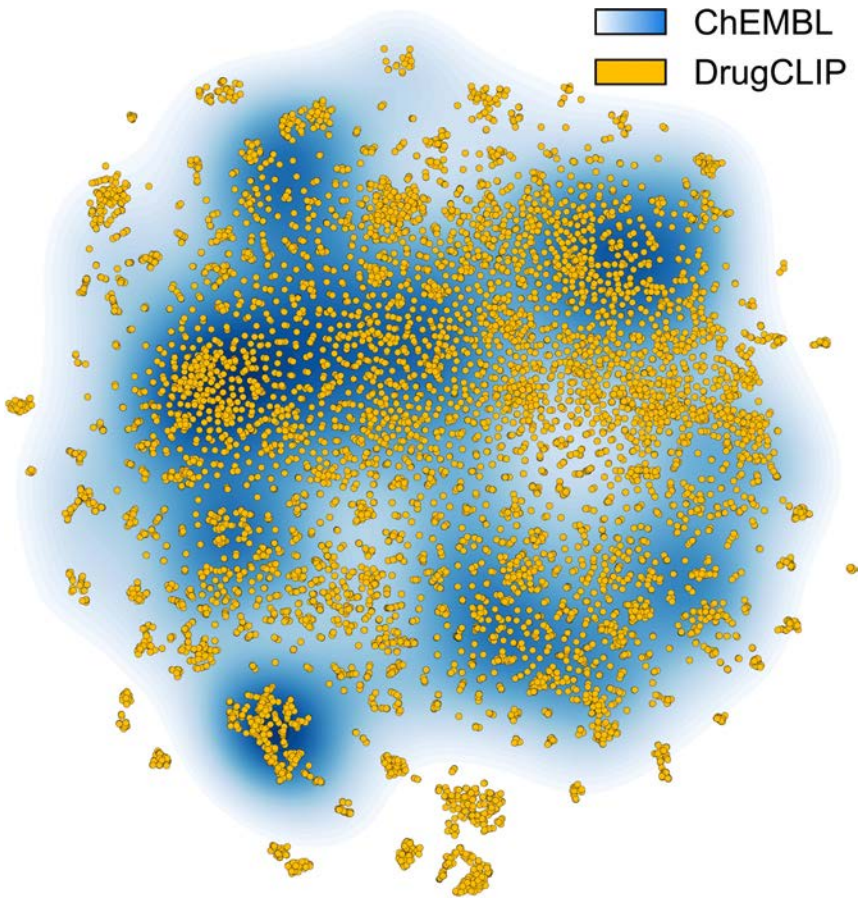
AlphaFold 2:预测/折叠所有蛋白质靶点结构



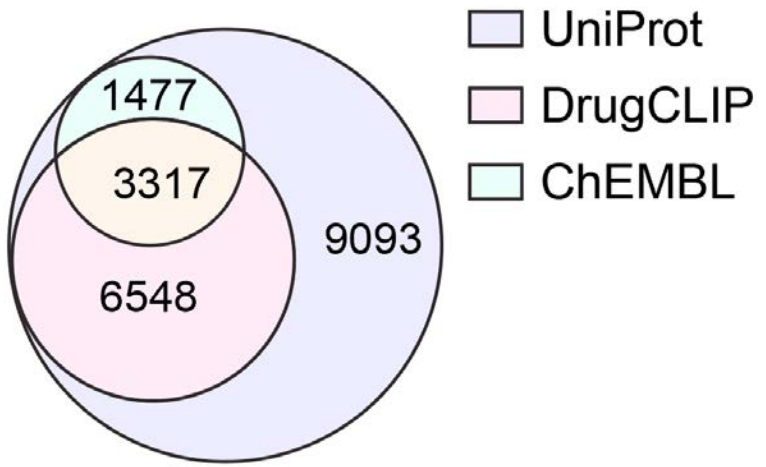
是否能够给每个靶点都找到对应的药物?

基于DrugCLIP的人类全基因组级别筛选

利用DrugCLIP的快速筛选能力，
实现了首个**全基因组范围**的虚拟筛选



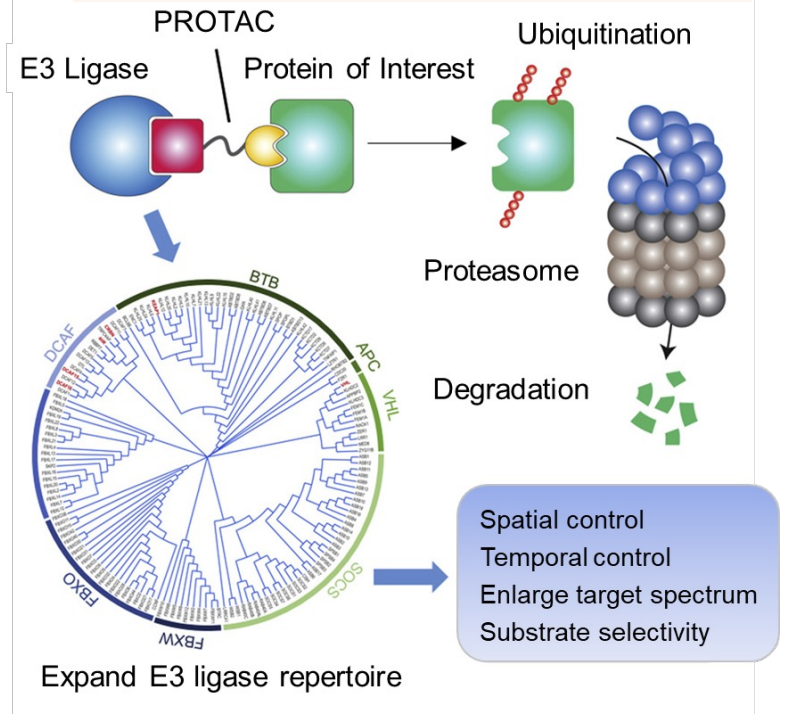
- ~10,000 人类蛋白
- 5亿分子
- > 200万 结合构象



DrugCLIP湿实验结果: TRIP12结合配体

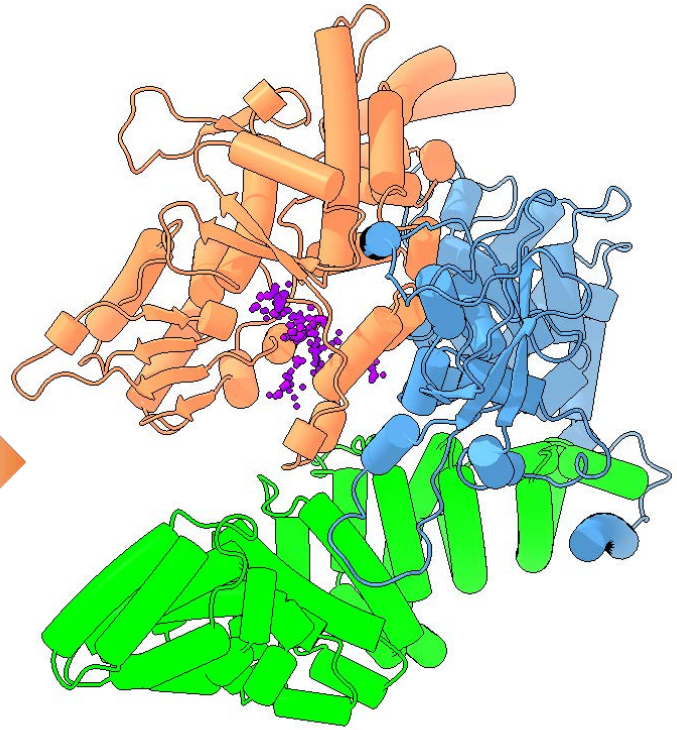
E3泛素化酶配体是开发新型靶向蛋白质降解药物的重要基础，**算法基于AlphaFold2预测的结构筛选出首个结合TRIP12的小分子配体**，筛选准确率**超过17.5%**，解决了**多数重要靶点缺少实验结构的困难**。

超过600个E3泛素化酶
仅不足20个被开发利用



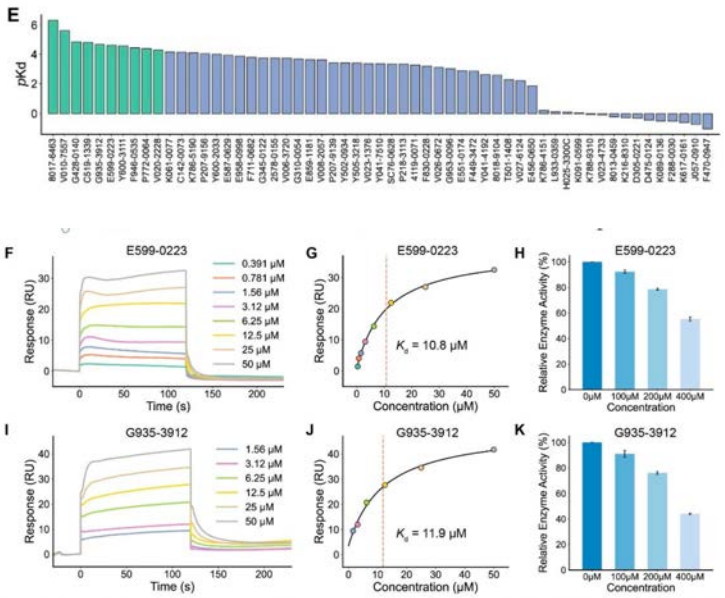
Kannt, A. *et al., Cell Chem. Bio.*, 2021

难点



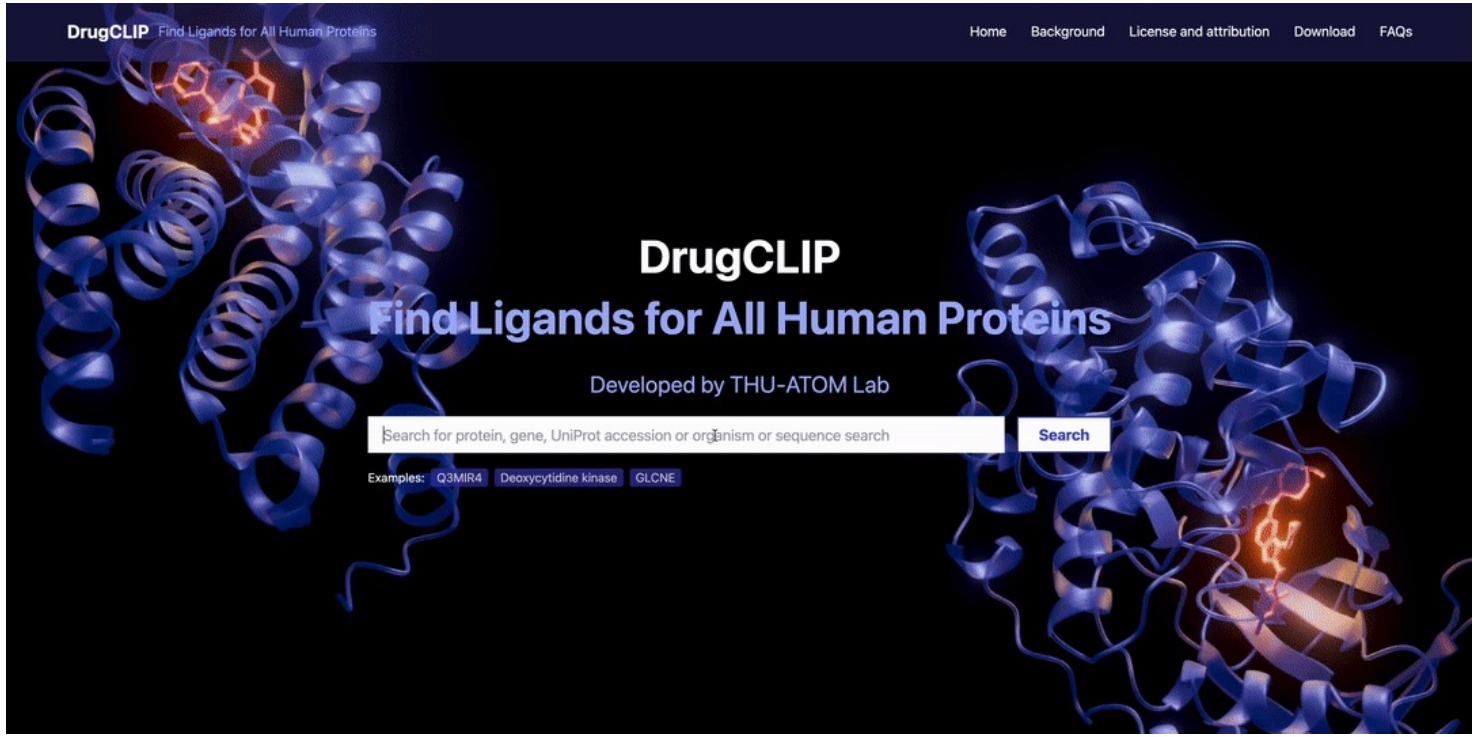
只能利用AlphaFold2预测的结构进行口袋定位与虚拟筛选

与清华刘磊团队合作



2个小分子活性在10um，通过荧光泛素化实验确认了剂量对活性的依赖性抑制， 在最高浓度下未显示出脱靶抑制作用

基于DrugCLIP的人类全基因组级别筛选

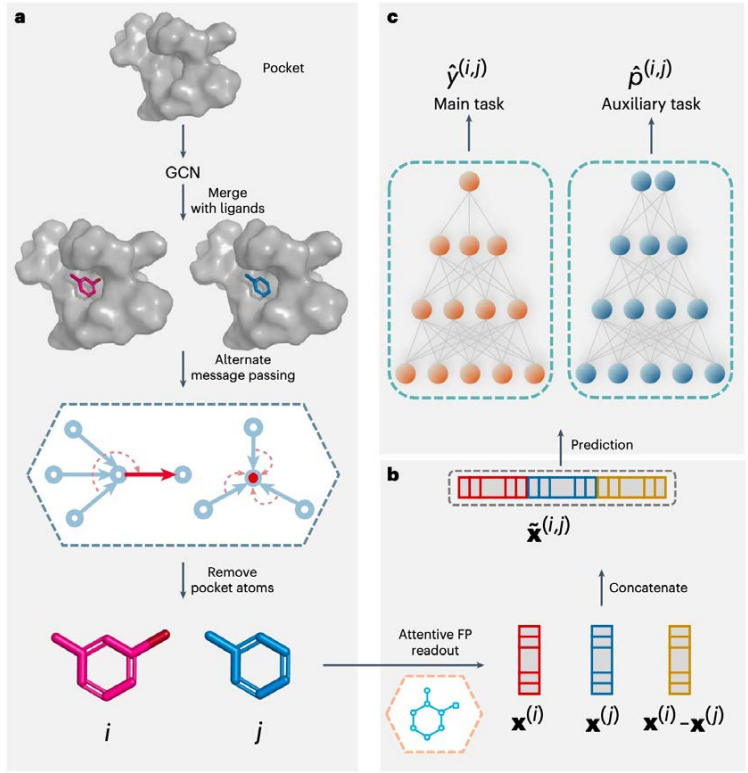


https://drug-the-whole-genome.yanyanlan.com

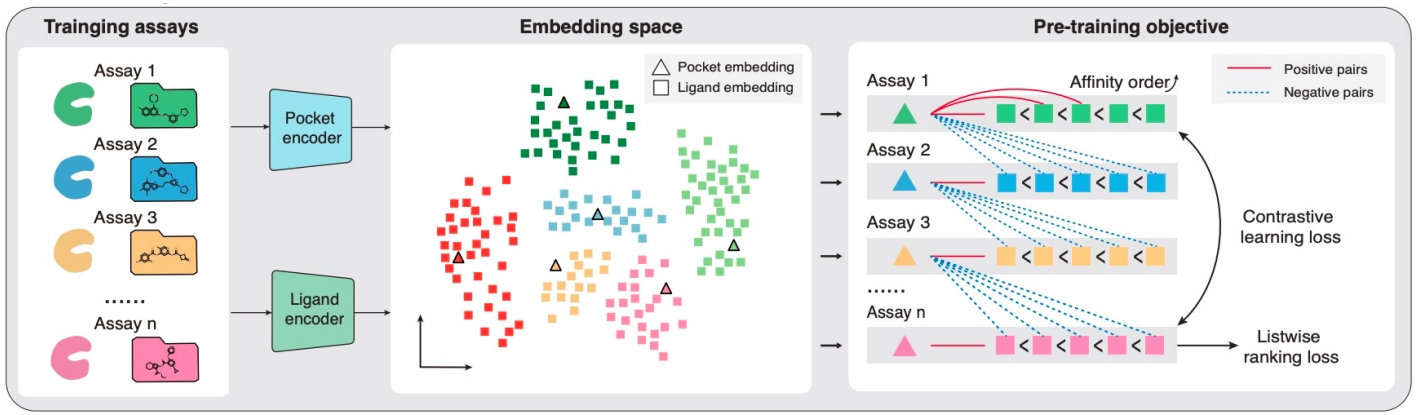
1天内进行了超过 10 万亿次评分计算（8块GPU）

虚拟筛选发展趋势

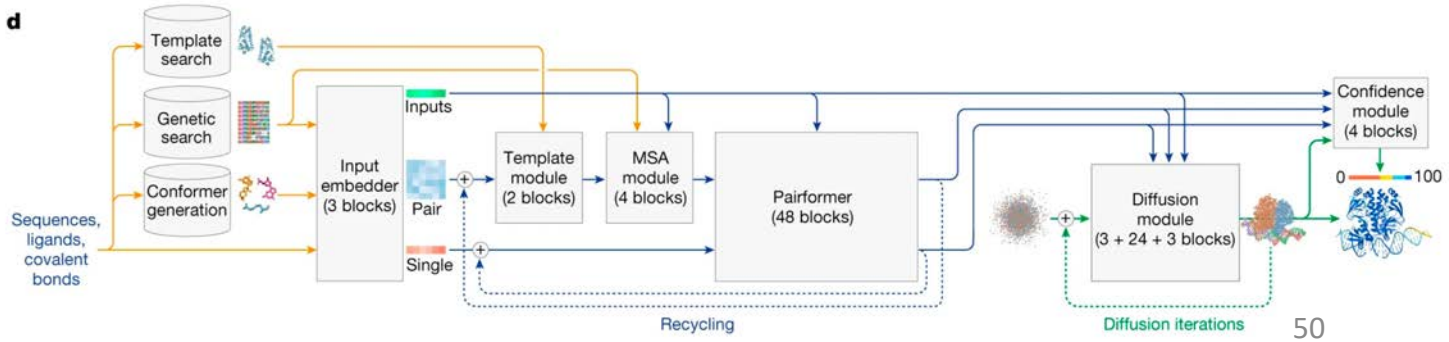
从大规模虚筛到精排



虚筛与精排的统一建模



利用AlphaFold3进行筛选排序



小分子精排

10^9 小分子

↓ **虚筛**

10^4 小分子(top 0.01%)

↓ **过滤**

1000小分子

↓ **精排**

100 Hits

虚筛

从大量候选分子（百万-亿）中召回一定数量（前1%）分子

- 对速度要求极高
- 对准确率有一定要求
- 测试集: DUD-E, LIT-PCBA

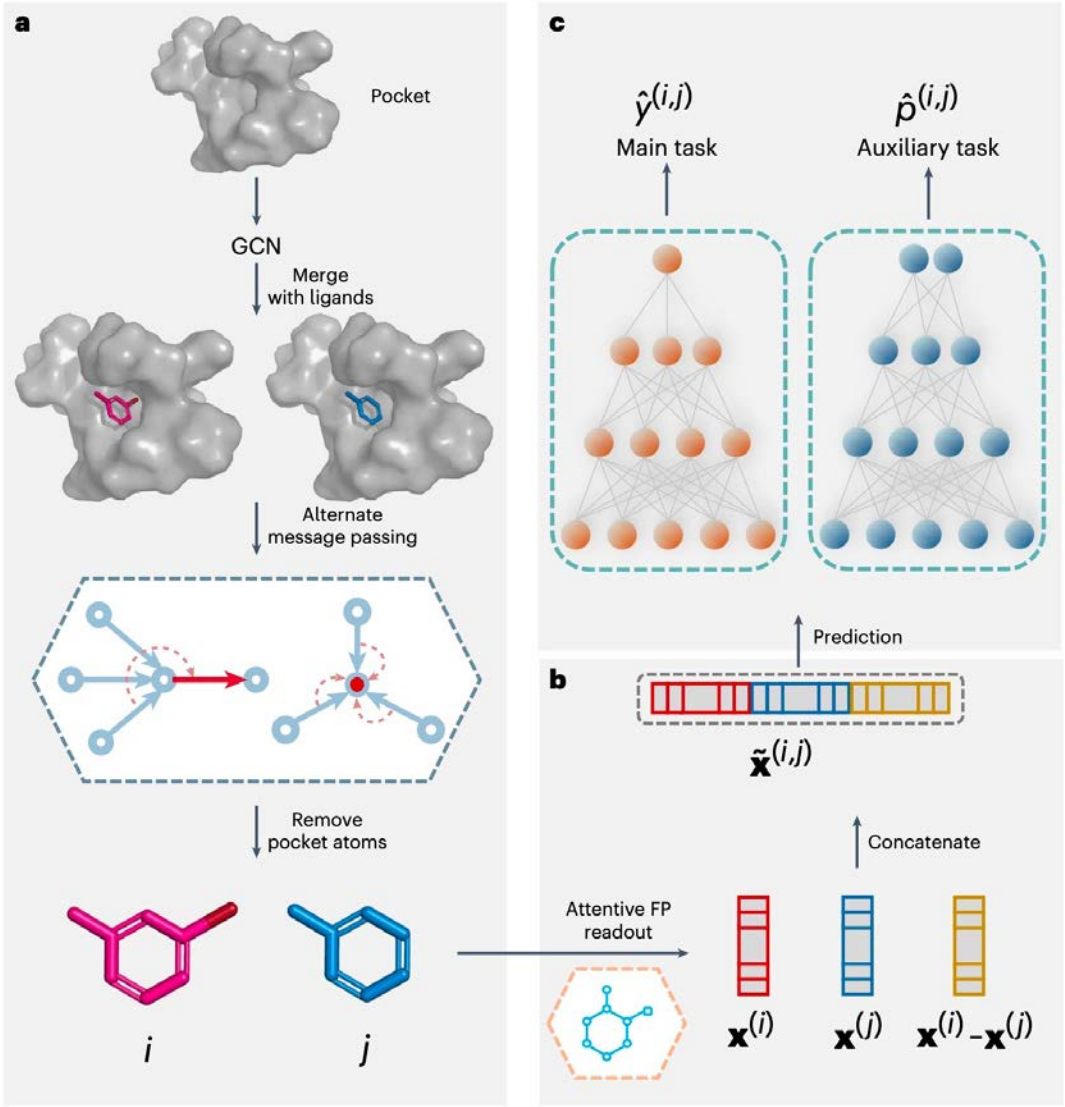
精排

从几百到几千的小范围化合物中挑选中最好的几个，进行湿实验验证

- 速度要求一般
- 对准确率要求极高
- 测试集：相对结合能预测

发展趋势: 精细化排序-相对结合能

- **问题：**传统基于相对结合自由能进行小分子精排的方法（如FEP+）计算量大，8卡GPU一天只能完成10余个小分子排序。
- **主要思路：**在编码后的embedding层面相减，显式建模差异信息（$\Delta x(i,j)$）
- **双任务学习：**
  - 相对排序任务（Pairwise Ranking）：判断分子 i 和 j 哪个结合更强
  - 结合能差值预测（$\Delta\Delta G$）：输出相对结合自由能

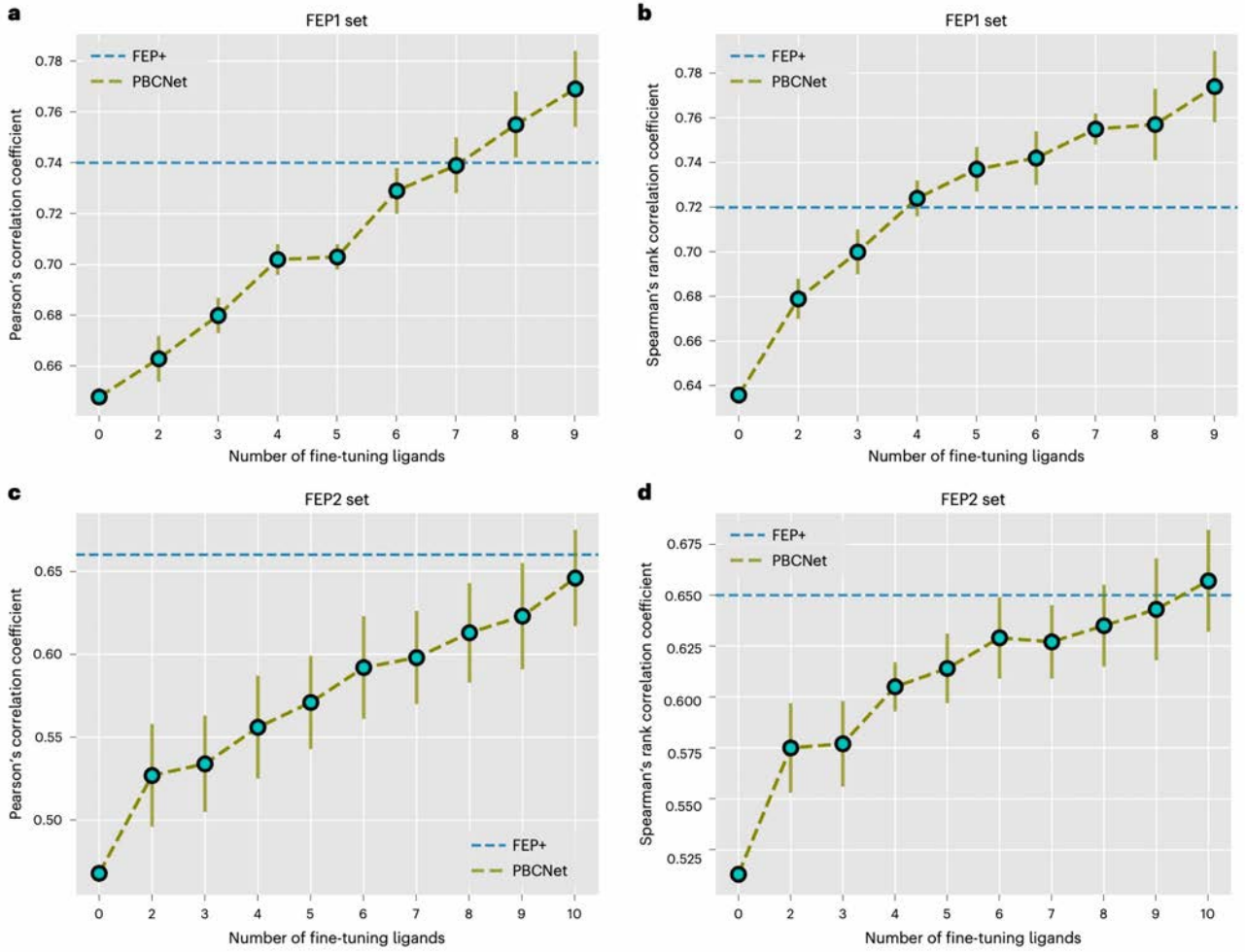


Computing the relative binding affinity of ligands based on a pairwise binding comparison network. Nature Computational Science, 2023.(Shanghai Institute of Materia Medica)

发展趋势: 精细化排序-相对结合能

Table 2 The performance of PBCNet with zero-shot learning on the FEP2 set [a]

		CDK8	c-Met	Eg5	HIF-2α	PFKFB3	SHP-2	SYK	TNKS2	Average
No. of compounds		33	24	28	42	40	26	44	27	33
FEP+	R	0.62	0.90	0.71	0.61	0.79	0.71	0.50	0.40	**0.66**
	ρ	0.74	0.88	0.72	0.59	0.79	0.78	0.42	0.41	**0.67**
	$RMSE_{pw}$ ($kcal \cdot mol^{-1}$)	2.09	1.43	1.23	1.60	1.78	1.39	1.61	2.20	1.67
Glide SP	R	0.00	0.00	0.00	0.40	0.47	0.44	0.24	0.37	0.24
	ρ	0.13	0.13	-0.08	0.42	0.51	0.44	0.21	0.32	0.26
	$RMSE_{pw}$ ($kcal \cdot mol^{-1}$)	2.49	3.01	1.90	1.51	1.57	1.52	1.27	1.35	1.83
MM-GB/SA	R	0.77	0.60	0.14	0.54	0.50	0.60	0.00	0.26	0.43
	ρ	0.82	0.64	0.10	0.48	0.54	0.50	-0.12	0.22	0.40
	$RMSE_{pw}$ ($kcal \cdot mol^{-1}$)	7.03	5.96	10.09	11.69	6.99	8.76	15.81	7.90	9.28
PIGNet	R	0.40	0.52	-0.08	-0.03	0.44	0.61	0.09	-0.17	0.22
	ρ	0.52	0.44	-0.15	0.29	0.48	0.45	0.10	-0.07	0.26
	$RMSE_{pw}$ ($kcal \cdot mol^{-1}$)	1.94	2.22	2.08	2.19	2.52	3.72	1.59	3.58	2.48
PBCNet [b]	R	$0.55^{0.58}_{0.50}$	$0.70^{0.81}_{0.60}$	$0.64^{0.69}_{0.52}$	$0.19^{0.25}_{0.11}$	$0.43^{0.51}_{0.35}$	$0.40^{0.42}_{0.36}$	$0.47^{0.56}_{0.29}$	$0.36^{0.37}_{0.31}$	$0.47^{0.52}_{0.38}$
	ρ	$0.63^{0.71}_{0.56}$	$0.76^{0.80}_{0.71}$	$0.58^{0.60}_{0.54}$	$0.30^{0.36}_{0.25}$	$0.47^{0.52}_{0.40}$	$0.55^{0.60}_{0.47}$	$0.48^{0.58}_{0.29}$	$0.32^{0.36}_{0.25}$	$0.51^{0.56}_{0.43}$
	$RMSE_{pw}$ ($kcal \cdot mol^{-1}$)	1.61	1.88	1.11	1.57	1.40	1.57	1.08	1.71	**1.49**
	$RMSE_{pw}$ (pIC_{50})	1.18	1.38	0.82	1.16	1.03	1.15	0.79	1.26	1.10



在zero-shot情况下超越其他方法，只低于FEP+

随着训练使用配体小分子数量增多，可以超过FEP+方法

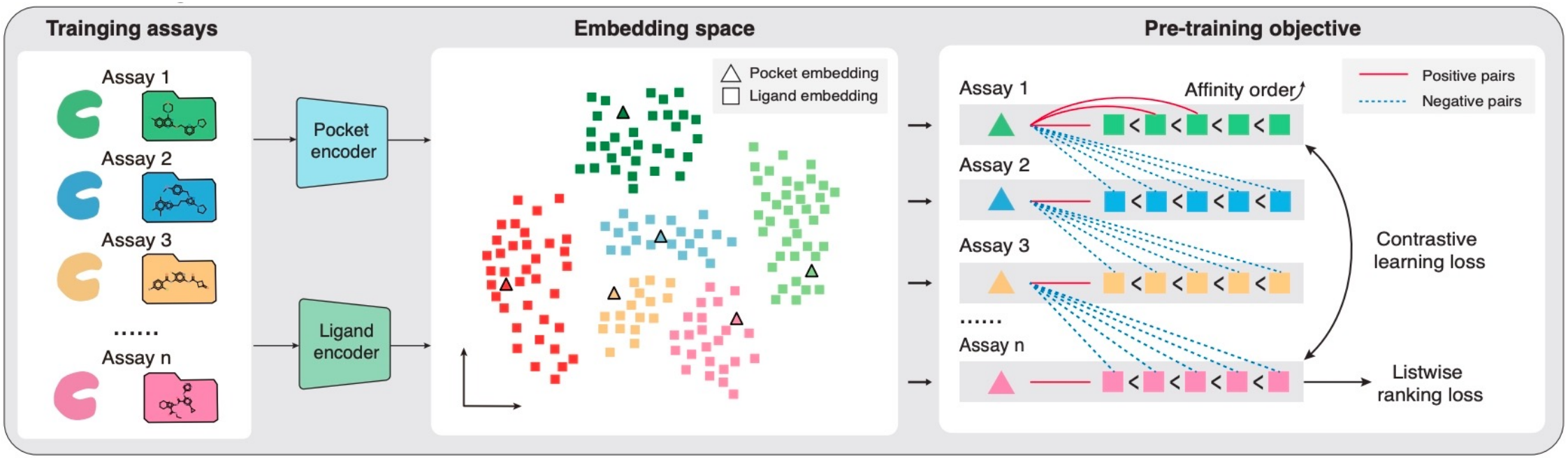
Computing the relative binding affinity of ligands based on a pairwise binding comparison network. Nature Computational Science, 2023.(Shanghai Institute of Materia Medica)

发展趋势：虚筛和精排的统一建模

Assay之间的Contrastive Loss：建模虚筛能力

Assay内部的Ranking Loss：建模精排能力

相辅相成，各自能力都显著提升

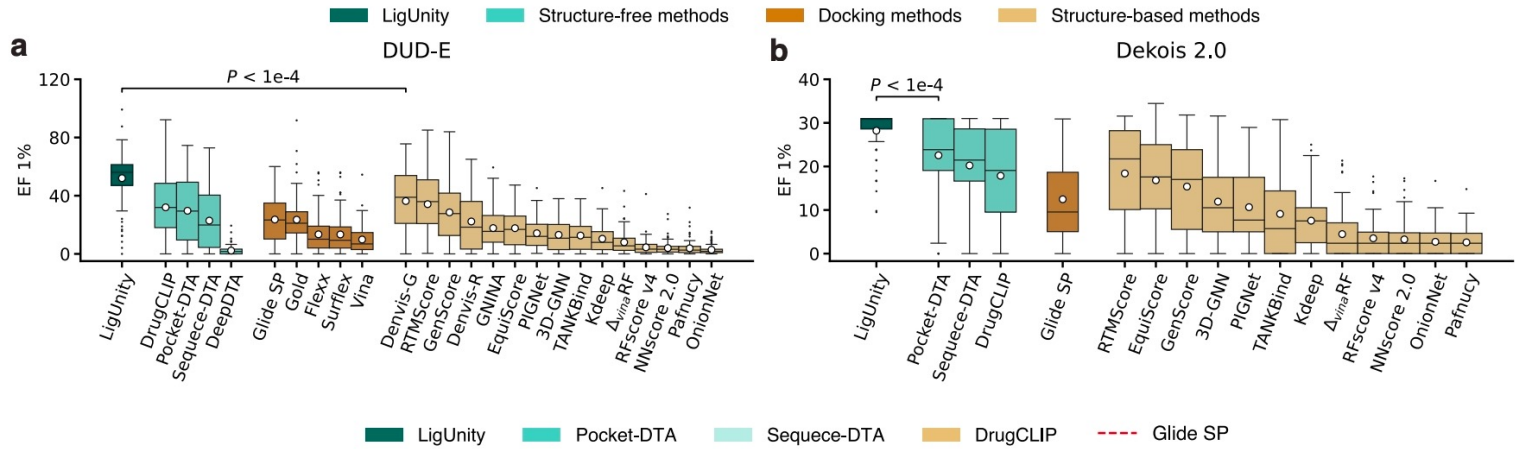


A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization. bioRxiv, 2025.(University of Washington)

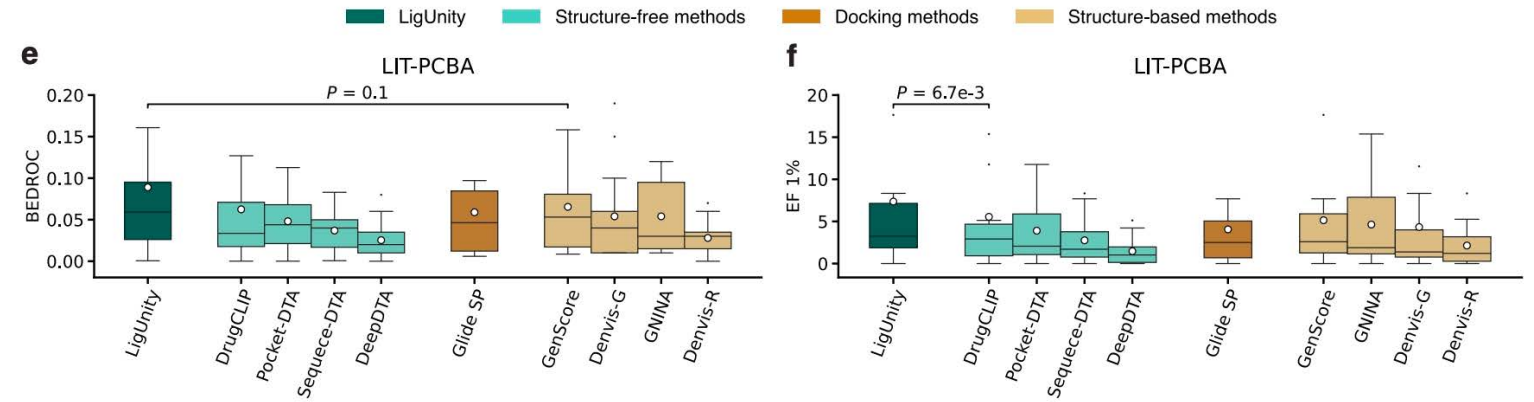
发展趋势: 虚筛和精排的统一建模

LigUnity在虚拟筛选上性能超过只进行对比学习优化的DrugCLIP

在DUD-E测试集（虚筛）的结果



在LIT-PCBA测试集（虚筛）的结果

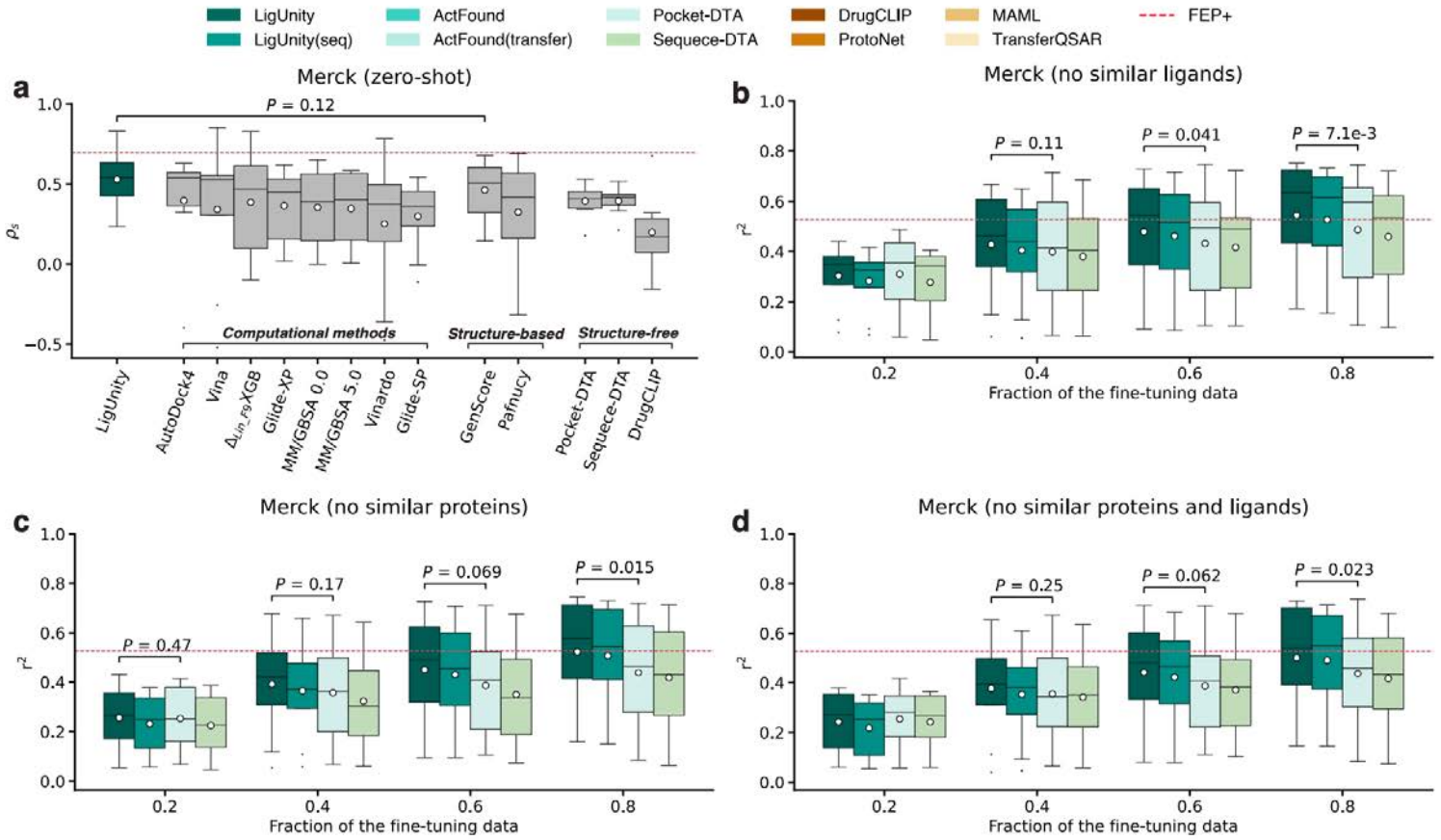


A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization. bioRxiv, 2025.(University of Washington)

发展趋势: 虚筛和精排的统一建模

在Merk数据集（精排）上的测试结果

- Zero-shot接近FEP+的效果
- 去除一定阈值内相似蛋白和分子的finetune的效果超过FEP+

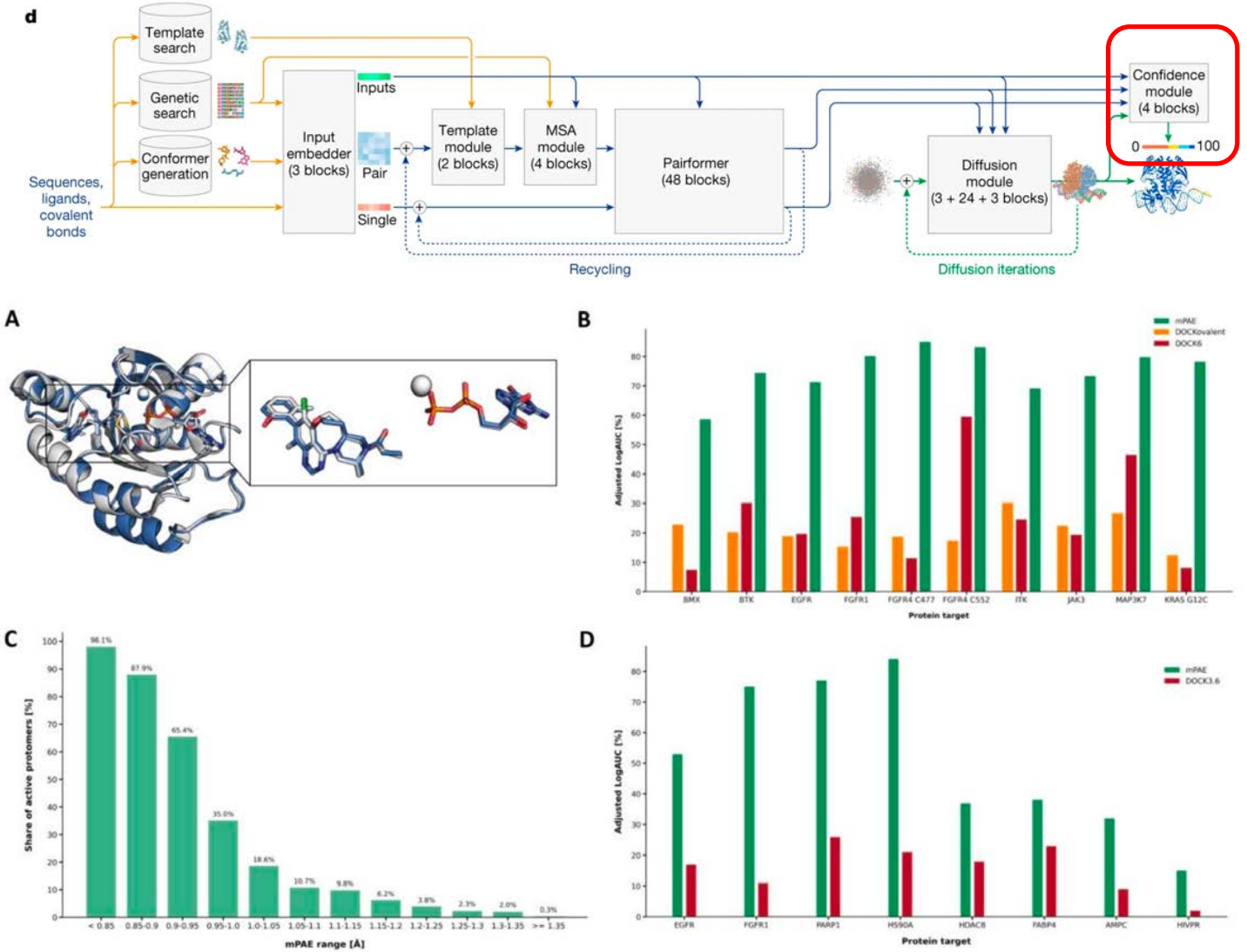


A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization. bioRxiv, 2025.(University of Washington)

发展趋势: 基于AF3的虚拟筛选

- 利用AF3对候选化合物和蛋白质一起预测构象
- 用AF3预测时的“Confidence”作为分子打分
- 在虚筛上取得了远超传统对接方法的效果

思考：如何更好的利用AF3的能力进行虚筛？使用置信分数作为排序依据是否足够？

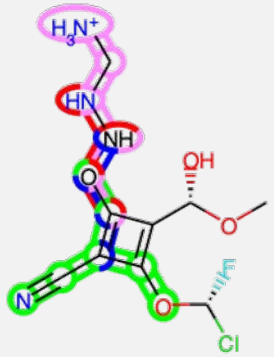


State-of-the-art covalent virtual screening with AlphaFold3. bioRxiv , 2025. (The Weizmann Institute of Science)

生成式药物设计：问题和任务描述

无条件生成

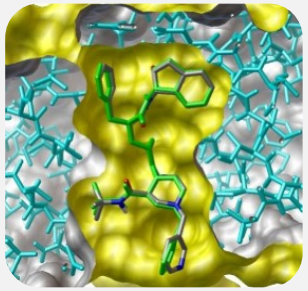
问题：从头生成合法的新分子、探索化学空间



- 合法性
- 新颖性
- 分布相似性

条件生成

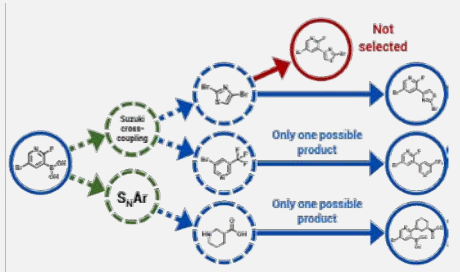
问题：给定条件生成满足条件的、合法的分子



- 蛋白亲和力
- ADMET性质(吸收、代谢、毒性等)

反应空间中的分子生成

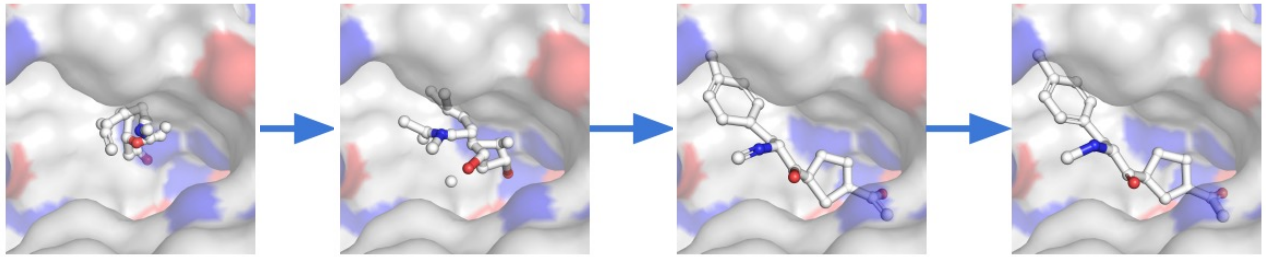
问题：生成新分子的同时设计合成路径



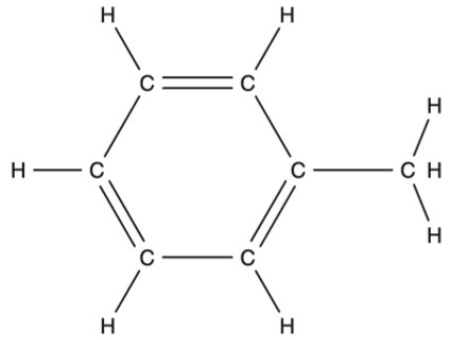
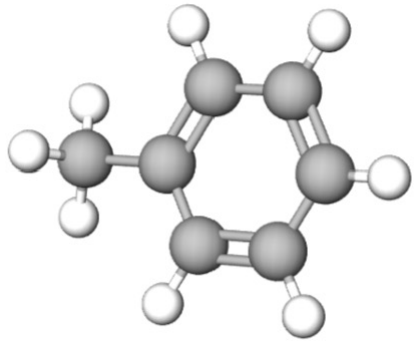
- 产率
- 可合成性

例： 基于结构的药物设计(SBDD)

- 条件：蛋白质口袋结构
- 目标：生成高亲和力分子



现有分子生成方法的分类

生成1D分子序列 ⟷	生成2D分子图	生成3D分子结构
• SMILES string: Cc1ccccc1 • SELFIES string: [C][C][=C][C][=C][C][=C][Ring1][=Branch1]		
优点: • 简洁、易于编码 • 数据充足 **缺点**: • 容易生成无效分子 • 表示质量不足	**优点**: • 更自然地表达分子拓扑结构 **缺点**: • 忽略空间结构 • GNN表达能力有限	**优点**: • 可建模构象、空间化学性质 • 对接 SBDD / ADMET 任务最相关 **缺点**: • 稳态构象数据稀缺 • 等变性建模复杂

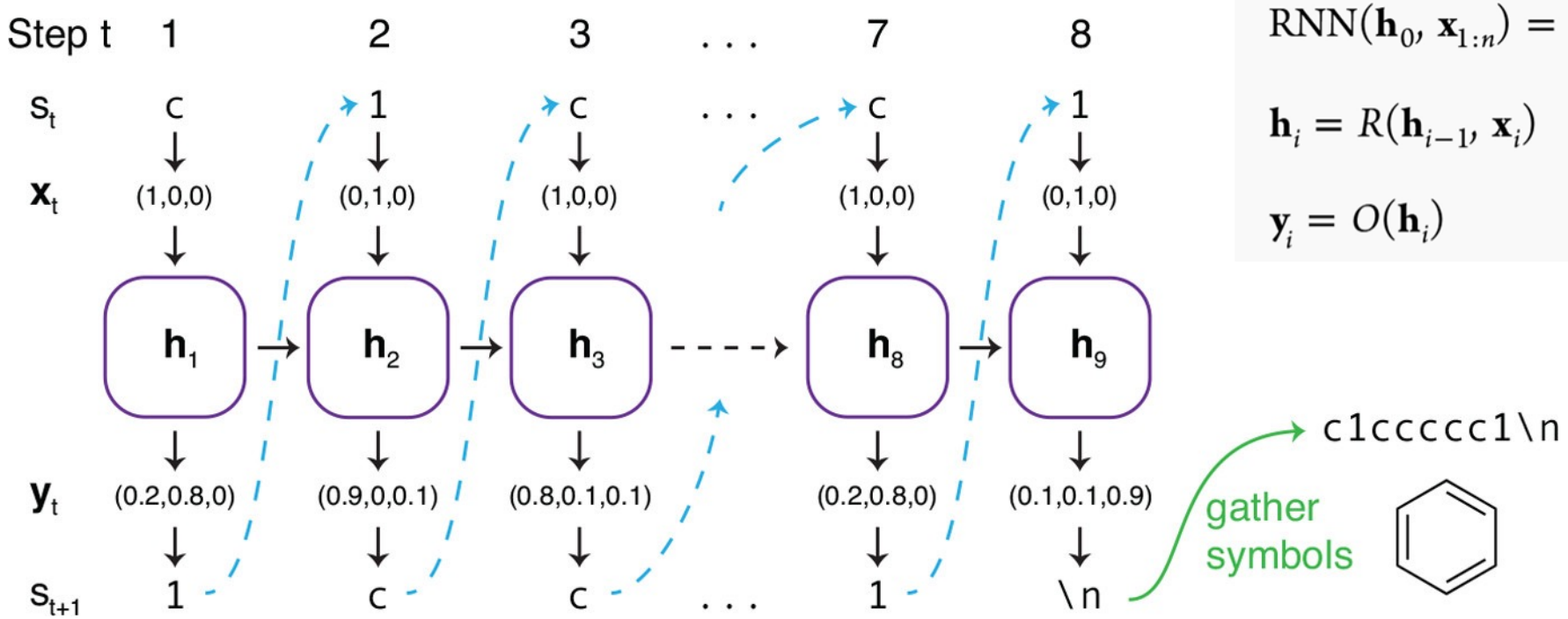
A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. Briefings in Bioinformatics, 2024.(Yale)

基于自回归语言模型的分子序列生成

自回归语言模型

$$P_\theta(S) = P_\theta(s_1) \cdot \prod_{t=2}^{T} P_\theta(s_t | s_{t-1}, \ldots, s_1)$$

循环神经网络



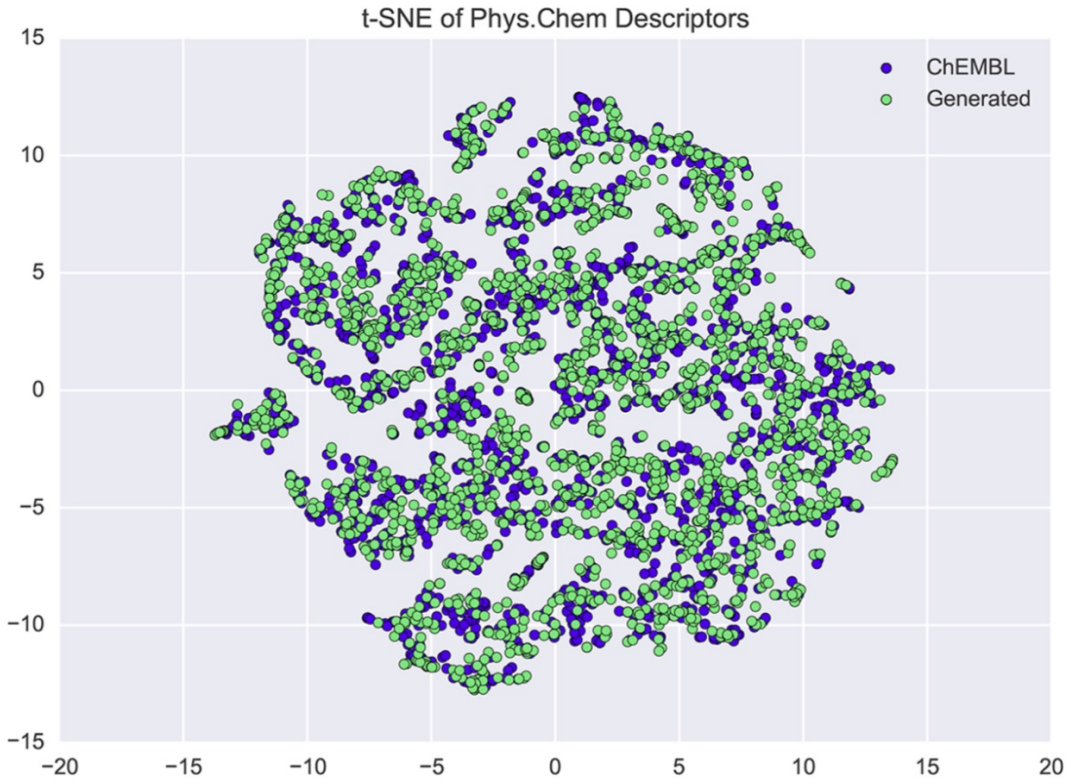
$$\mathrm{RNN}(\mathbf{h}_0, \mathbf{x}_{1:n}) = \mathbf{h}_{1:n}, \mathbf{y}_{1:n}$$

$$\mathbf{h}_i = R(\mathbf{h}_{i-1}, \mathbf{x}_i)$$

$$\mathbf{y}_i = O(\mathbf{h}_i)$$

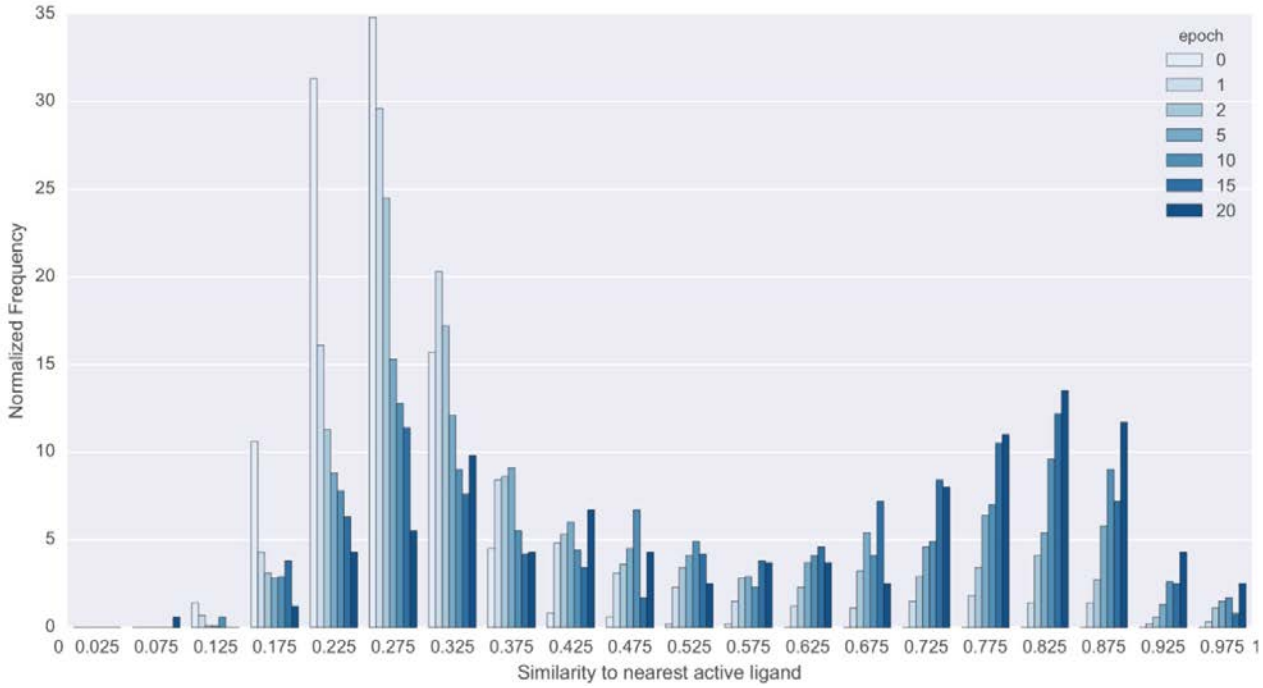
Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science, 2018. (Westfälische Wilhelms-Universität Münster)

基于自回归语言模型的分子序列生成



无条件生成小分子，与真实小分子分布相似

Targeting the 5-HT 2A Receptor

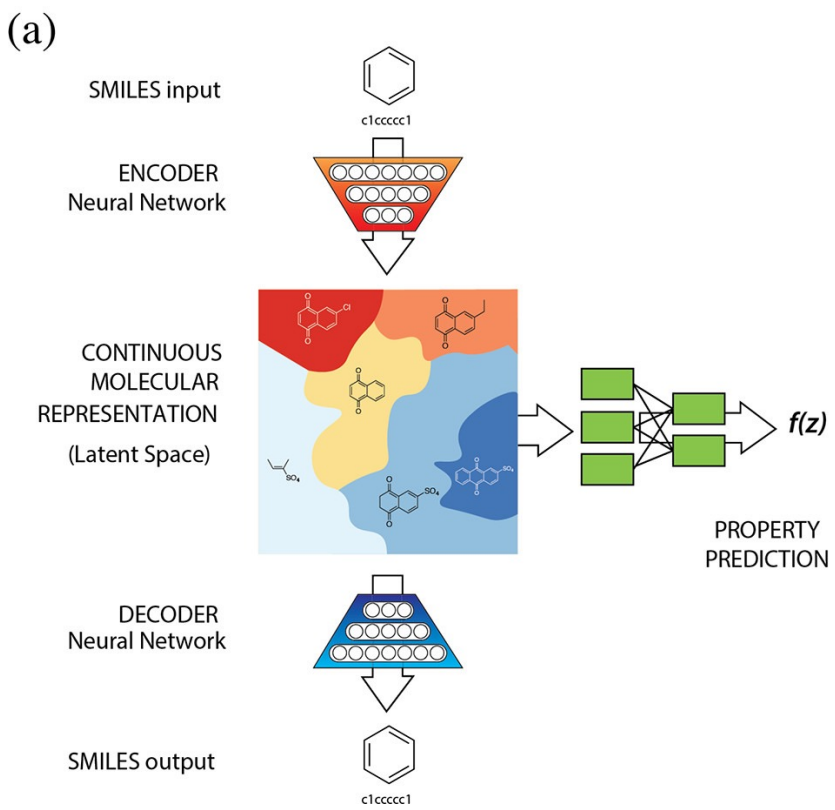


在活性分子集上微调模型，能够生成与活性分子更接近的分子

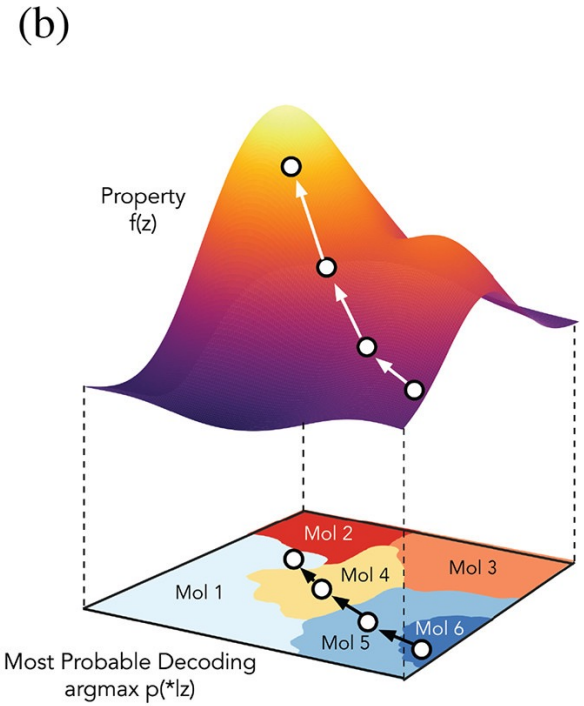
Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science, 2018. (Westfälische Wilhelms-Universität Münster)

基于VAE的分子序列生成

核心思路：将离散的分子序列通过VAE转化为与属性关联的连续分子表示。



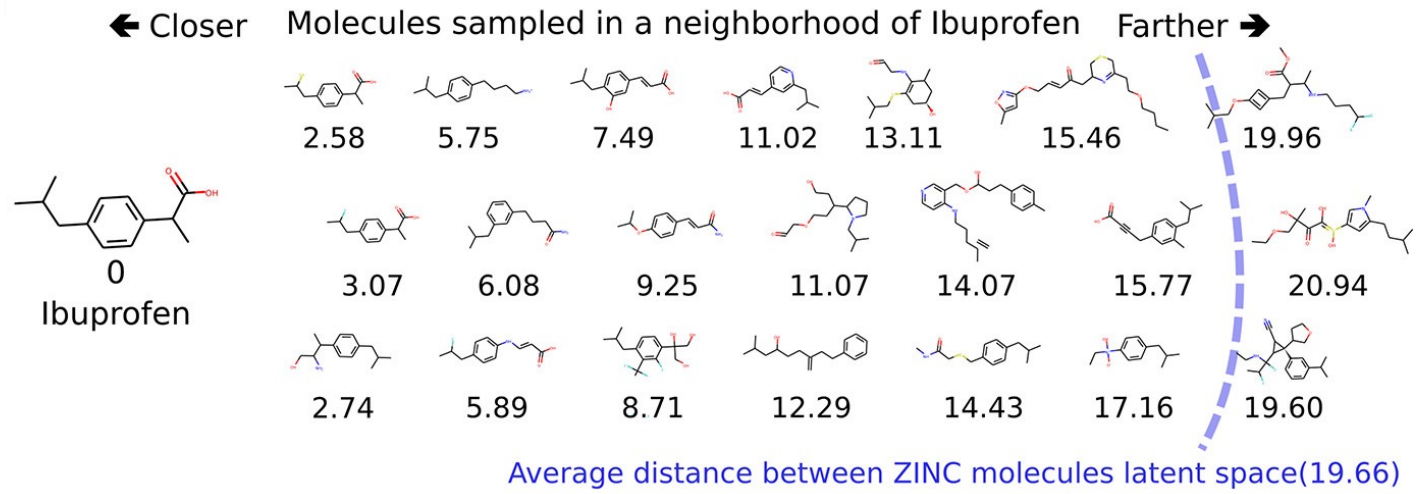
用于分子设计的自动编码器示意图，包含联合属性预测模型



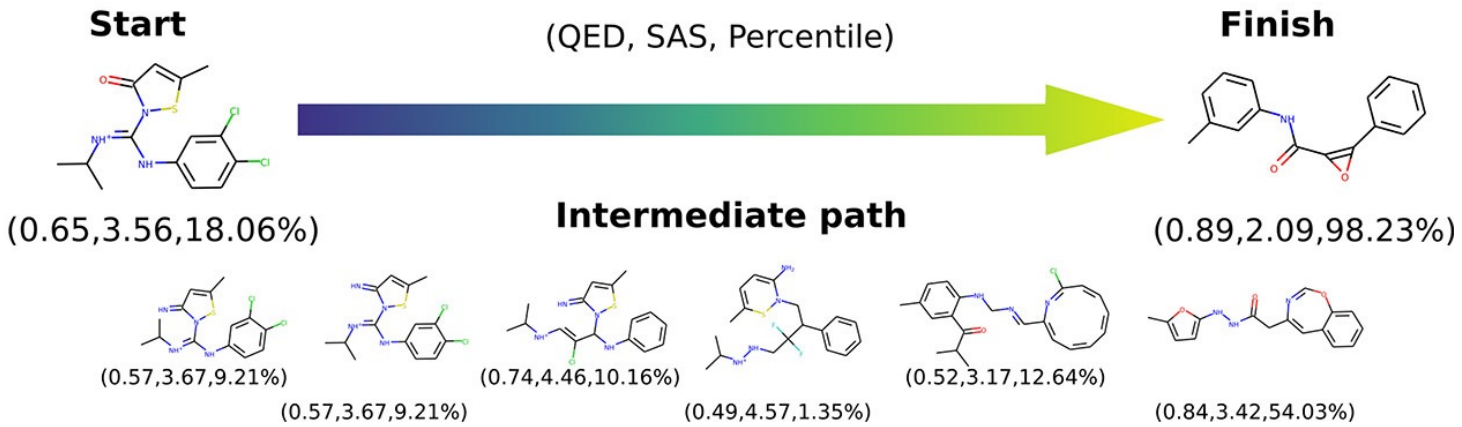
以预期属性值为目标的隐空间优化与序列生成

Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, 2018. (Harvard University)

基于VAE的分子序列生成



探索连续空间上目标分子邻近表示的分子序列生成



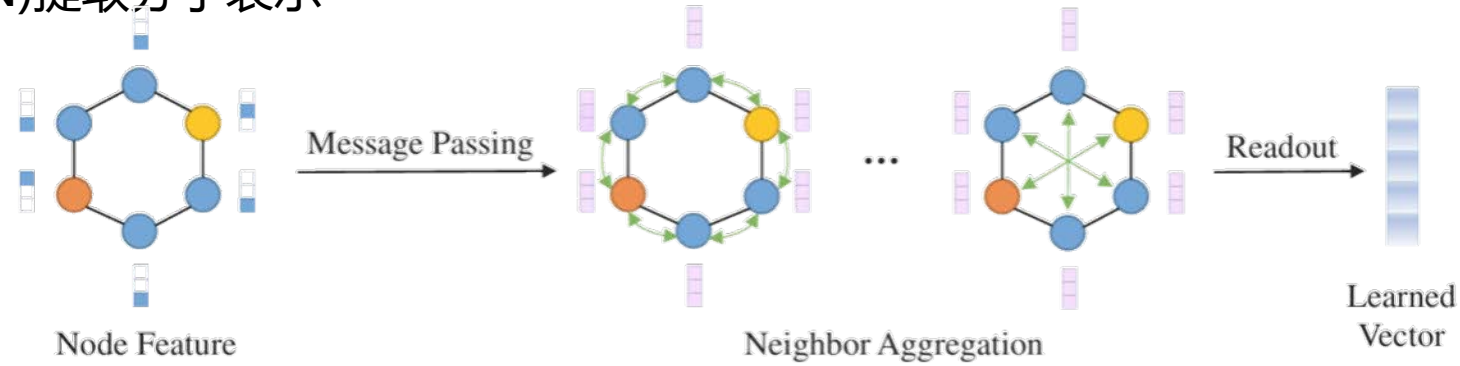
给定分子性质条件的分子序列生成

Automatic chemical design using a data-driven continuous representation of molecules. ACS central science, 2018. (Harvard University)

2D分子图生成方法基本思路

数据：原子类型+ 化学键类型

常用图神经网络(GNN)提取分子表示



分子图具有层次结构：分子 → 子结构 → 原子

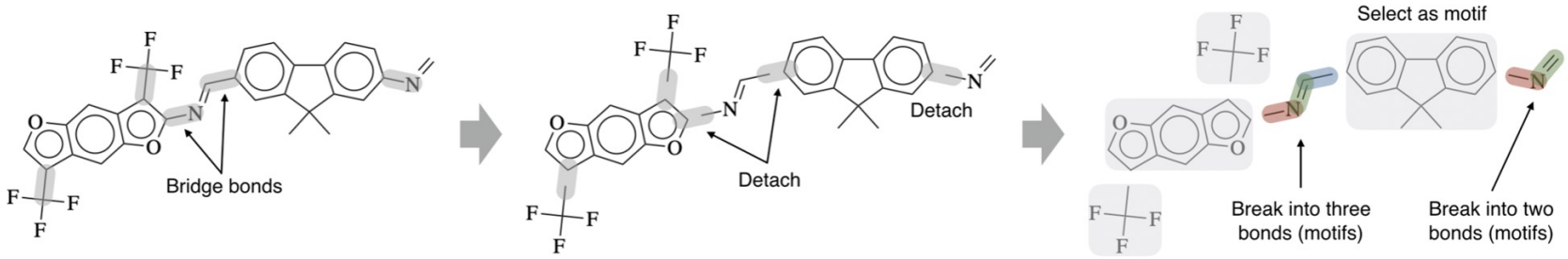


Figure 6. Illustration of motif extraction procedure.

常用生成算法：自回归、变分自编码器、扩散模型……

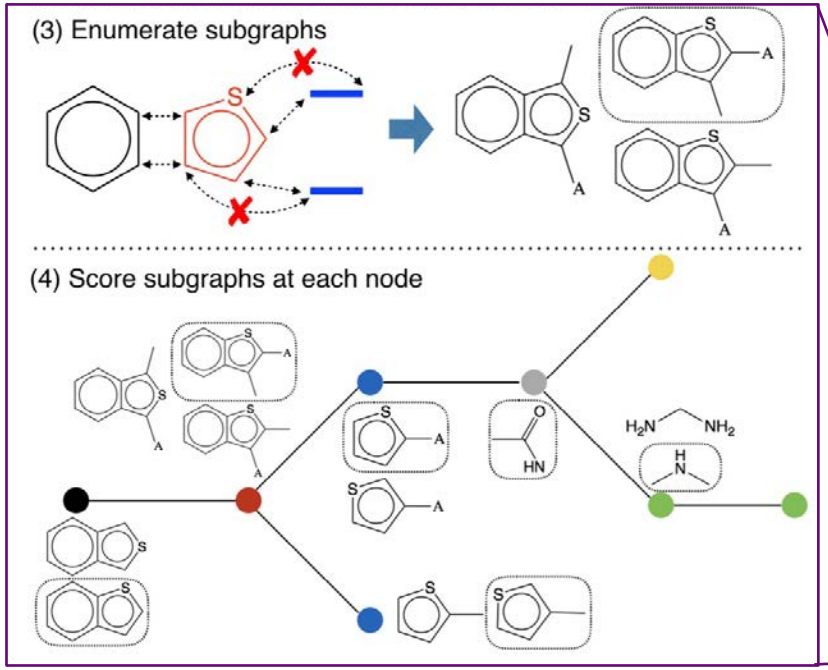
层次化的2D分子图生成

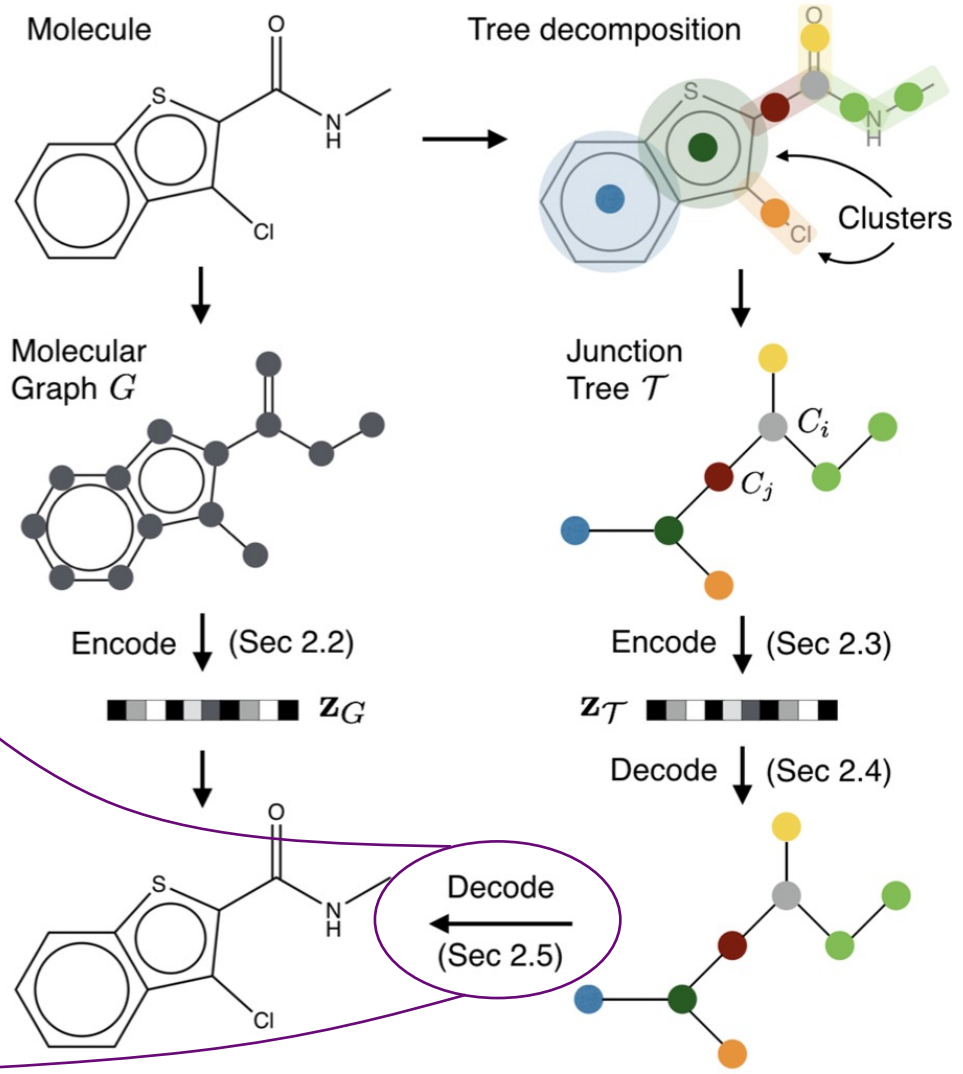
先生成子结构级别的图，再生成原子级别的图

a. 将分子图分解为连接树 ,每个节点代表一个子结构。

b. 提取树和图的潜空间表示。

c. 根据树的潜空间表示重建连接树，然后根据图的潜空间表示将树中的节点重新组装回完整的分子图。





Junction Tree Variational Autoencoder for Molecular Graph Generation. ICML, 2018. (MIT)

层次化的2D分子图生成

无条件生成

Table 1. Reconstruction accuracy and prior validity results. Baseline results are copied from Kusner et al. (2017); Dai et al. (2018); Simonovsky & Komodakis (2018); Li et al. (2018).

Method	Reconstruction	Validity
CVAE	44.6%	0.7%
GVAE	53.7%	7.2%
SD-VAE	76.2%	43.5%
GraphVAE	-	13.5%
Atom-by-Atom LSTM	-	89.2%
JT-VAE	**76.7%**	**100.0%**

生成中使用合法子结构可保证有效性

学习潜空间的性质预测器，将参考分子的潜空间表示作为起点，进行梯度上升

条件生成

方法：贝叶斯优化或潜空间梯度上升

优化目标：$y(m) = logP(m) - SA(m) - cycle(m)$ where $cycle(m)$ counts the number of rings that have more than six atoms.

Table 2. Best molecule property scores found by each method. Baseline results are from Kusner et al. (2017); Dai et al. (2018).

Method	1st	2nd	3rd
CVAE	1.98	1.42	1.19
GVAE	2.94	2.89	2.80
SD-VAE	4.04	3.50	2.96
JT-VAE	**5.30**	**4.93**	**4.49**

约束优化：最大化性质值，且分子相似度满足阈值。

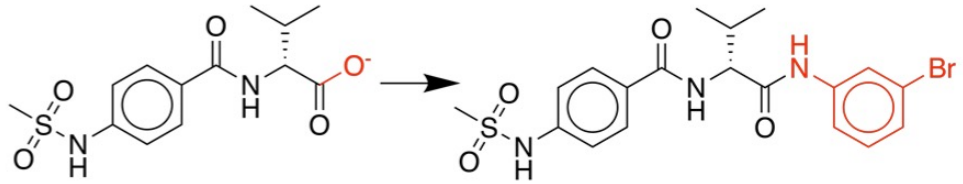
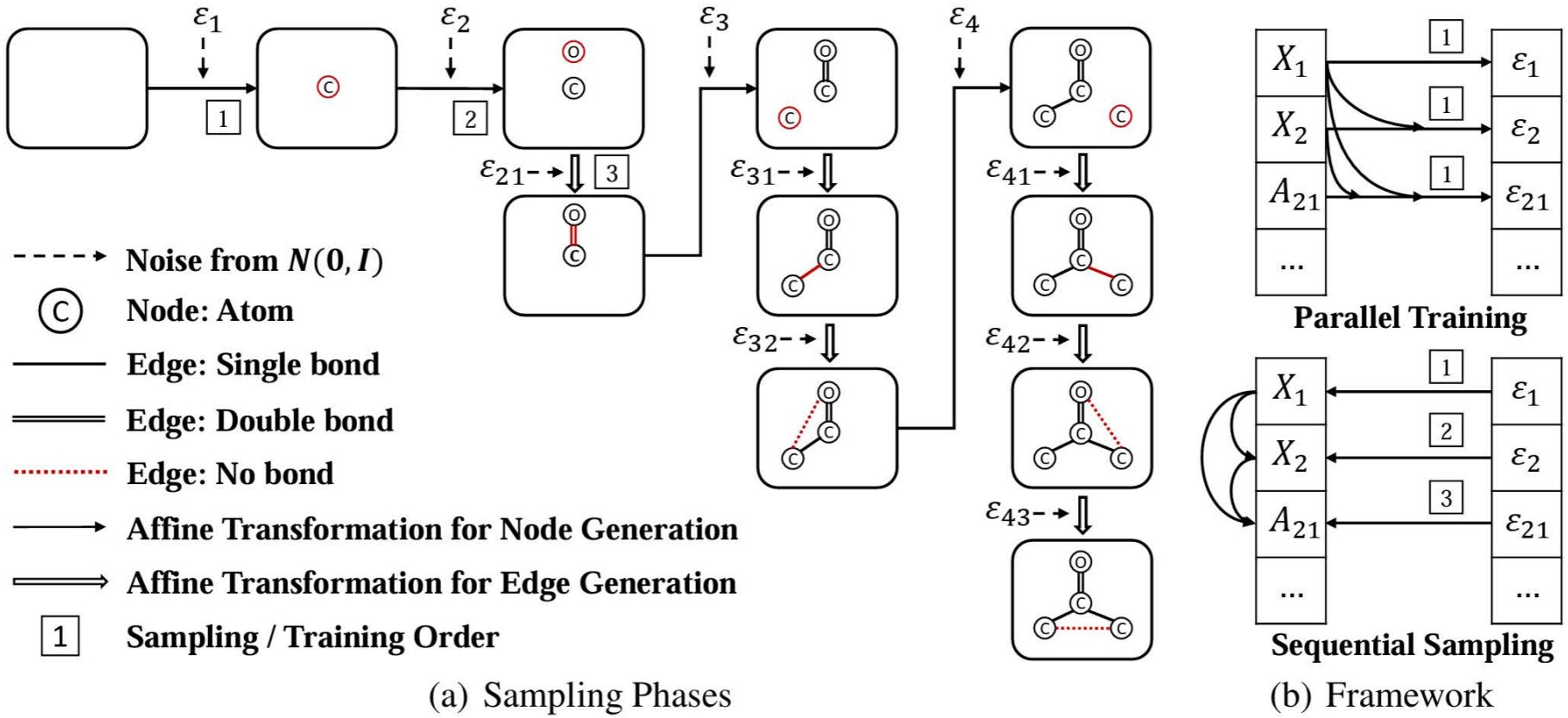


Figure 8. A molecule modification that yields an improvement of 4.0 with molecular similarity 0.617 (modified part is in red).

Junction Tree Variational Autoencoder for Molecular Graph Generation. ICML, 2018. (MIT)

基于自回归流模型的2D分子图生成



(a) Sampling Phases (b) Framework

核心思想：将分子生成问题形式化为一个顺序决策过程

自回归流： 特殊的正规化流模型，Jacobi矩阵是三角阵

采样过程：自回归生成，允许生成步骤中利用化学知识并进行价态检查，提升分子合法性。

GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation. ICLR 2020. (Peking University)

基于自回归流模型的2D分子图生成

无条件生成

Table 2: Comparison of different models on density modeling and generation. *Reconstruction* is only evaluated on latent variable models. *Validity w/o check* is only evaluated on models with valency constraints. Result with † is obtained by running GCPN's open-source code. Results with ‡ are taken from Popova et al. (2019).

Method	Validity	Validity w/o check	Uniqueness	Novelty	Reconstruction
JT-VAE	100%	—	100%‡	100%‡	76.7%
GCPN	100%	20%†	99.97%‡	100%‡	—
MRNN	100%	65%	99.89%	100%	—
GraphNVP	42.60%	—	94.80%	100%	100%
GraphAF	100%	68%	99.10%	100%	100%

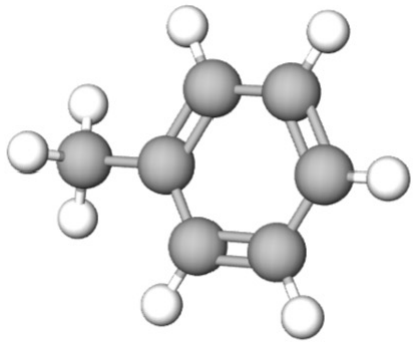
Table 6: Comparison of results on constrained property optimization.

δ	JT-VAE			GCPN			GraphAF		
	Improvement	Similarity	Success	Improvement	Similarity	Success	Improvement	Similarity	Success
0.0	1.91 ± 2.04	0.28 ± 0.15	97.5%	4.20 ± 1.28	0.32 ± 0.12	100%	$\mathbf{13.13 \pm 6.89}$	0.29 ± 0.15	100%
0.2	1.68 ± 1.85	0.33 ± 0.13	97.1%	4.12 ± 1.19	0.34 ± 0.11	100%	$\mathbf{11.90 \pm 6.86}$	0.33 ± 0.12	100%
0.4	0.84 ± 1.45	0.51 ± 0.10	83.6%	2.49 ± 1.30	0.47 ± 0.08	100%	$\mathbf{8.21 \pm 6.51}$	0.49 ± 0.09	99.88%
0.6	0.21 ± 0.71	0.69 ± 0.06	46.4%	0.79 ± 0.63	0.68 ± 0.08	100%	$\mathbf{4.98 \pm 6.49}$	0.66 ± 0.05	96.88%

条件生成：使用强化学习 Proximal Policy Optimization 策略，以目标性质以及化学合法性作为奖励函数。约束优化： 将参考分子的片段作为起点开始自回归生成

3D分子生成方法基本思路

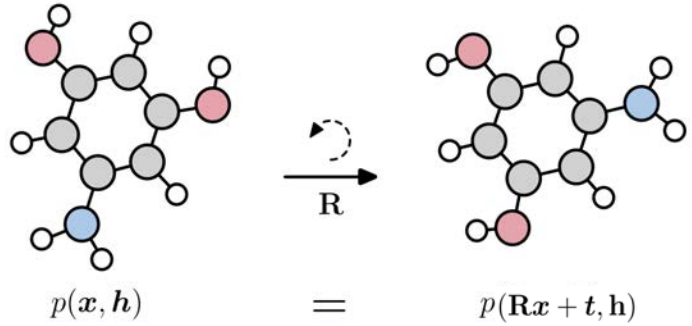
数据：



原子类型h(离散) + 原子位置x(连续)

H	-0.21463	0.97837	0.33136
C	-0.38325	0.66317	-0.70334
C	-1.57552	0.03829	-1.05450
H	-2.34514	-0.13834	-0.29630
C	-1.78983	-0.36233	-2.36935
H	-2.72799	-0.85413	-2.64566
C	-0.81200	-0.13809	-3.33310
H	-0.98066	-0.45335	-4.36774
C	0.38026	0.48673	-2.98192
H	1.14976	0.66307	-3.74025
C	0.59460	0.88737	-1.66708
H	1.53276	1.37906	-1.39070

网络结构：等变图神经网络，保证建模的分布满足平移旋转不变性

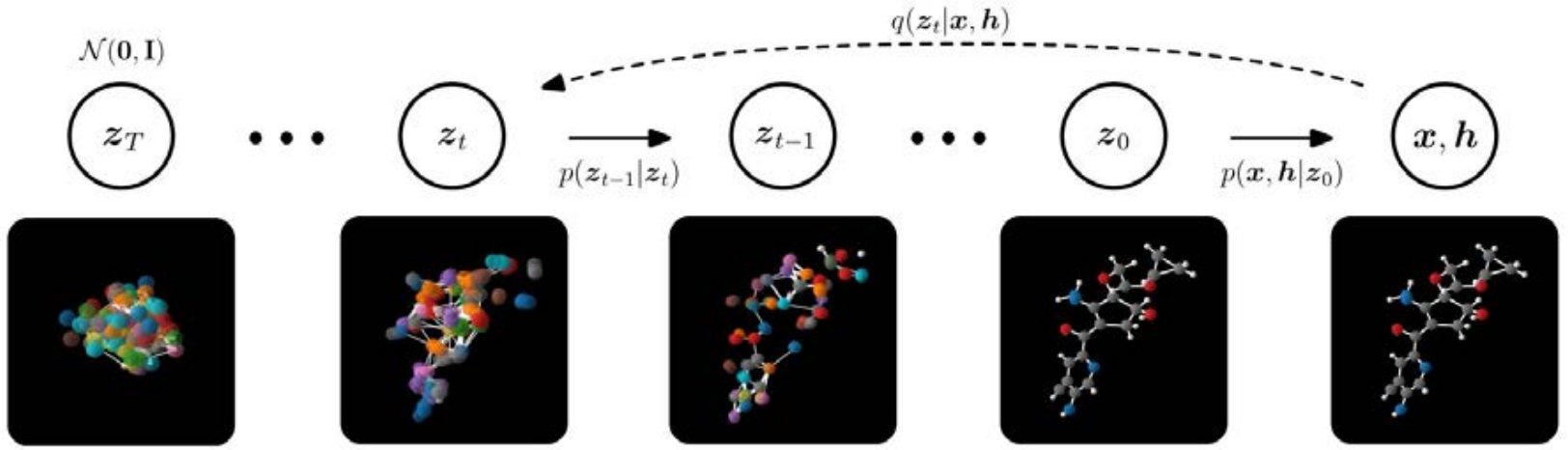


对原子位置做平移旋转，生成概率不变

挑战：引入等变性的分子归纳偏置统一离散与连续模态的分子数据

生成算法：规范化流(Normalizing Flow)、扩散模型(Diffusion Model)、流匹配(Flow Matching)、贝叶斯流网络(Bayesian Flow Networks)、自回归(Autoregressive)等……

基于等变扩散模型的3D分子生成方法EDM



离散数据处理：One Hot表示

Algorithm 1 Optimizing EDM

Input: Data point $\boldsymbol{x}$, neural network ϕ
Sample $t \sim \mathcal{U}(0,\ldots,T)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
Subtract center of gravity from $\boldsymbol{\epsilon}^{(x)}$ in $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}^{(x)}, \boldsymbol{\epsilon}^{(h)}]$
Compute $\boldsymbol{z}_t = \alpha_t[\boldsymbol{x},\boldsymbol{h}] + \sigma_t\epsilon$ ◁ 加噪过程
Minimize $||\boldsymbol{\epsilon} - \phi(\boldsymbol{z}_t, t)||^2$ ◁ 神经网络拟合噪声

Algorithm 2 Sampling from EDM

Sample $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
for t in $T, T-1, \ldots, 1$ where $s = t-1$ **do**
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
  Subtract center of gravity from $\boldsymbol{\epsilon}^{(x)}$ in $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}^{(x)}, \boldsymbol{\epsilon}^{(h)}]$
  $\boldsymbol{z}_s = \frac{1}{\alpha_{t|s}}\boldsymbol{z}_t - \frac{\sigma^2_{t|s}}{\alpha_{t|s}\sigma_t} \cdot \phi(\boldsymbol{z}_t, t) + \sigma_{t\to s} \cdot \boldsymbol{\epsilon}$
end for ◁ 去噪生成过程
Sample $\boldsymbol{x}, \boldsymbol{h} \sim p(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{z}_0)$

保证等变性：1. 在质心为0的子空间采样标准高斯噪声；2. 使用等变网络（如EGNN）拟合噪声

Equivariant Diffusion for Molecule Generation in 3D. ICML 2022. (University of Amsterdam)

基于等变扩散模型的3D分子生成方法EDM

无条件生成：

Table 1. Neg. log-likelihood $-\log p(\mathbf{x}, \mathbf{h}, M)$, atom stability and molecule stability with standard deviations across 3 runs on QM9, each drawing 10000 samples from the model.

# Metrics	NLL	Atom stable (%)	Mol stable (%)
E-NF	-59.7	85.0	4.9
G-Schnet	N.A	95.7	68.1
GDM	-94.7	97.0	63.2
GDM-aug	-92.5	97.6	71.6
EDM (ours)	**-110.7**$\pm$1.5	**98.7**$\pm$0.1	**82.0**$\pm$0.4
Data		99.0	95.2

等变生成优于 ▷ 非等变方法

Table 2. Validity and uniqueness over 10000 molecules with standard deviation across 3 runs. Results marked (*) are not directly comparable, as they do not use 3D coordinates to derive bonds. H: model hydrogens explicitly

Method	H	Valid (%)	Valid and Unique (%)
Graph VAE (*)		55.7	42.3
GTVAE (*)		74.6	16.8
Set2GraphVAE (*)		59.9$\pm$1.7	56.2$\pm$1.4
EDM (ours)		**97.5**$\pm$0.2	**94.3**$\pm$0.2
E-NF	✓	40.2	39.4
G-Schnet	✓	85.5	80.3
GDM-aug	✓	90.4	89.5
EDM (ours)	✓	**91.9**$\pm$0.5	**90.7**$\pm$0.6
Data	✓	97.7	97.7

◁ 合法性优于2D方法

条件生成: 条件为分子极化率

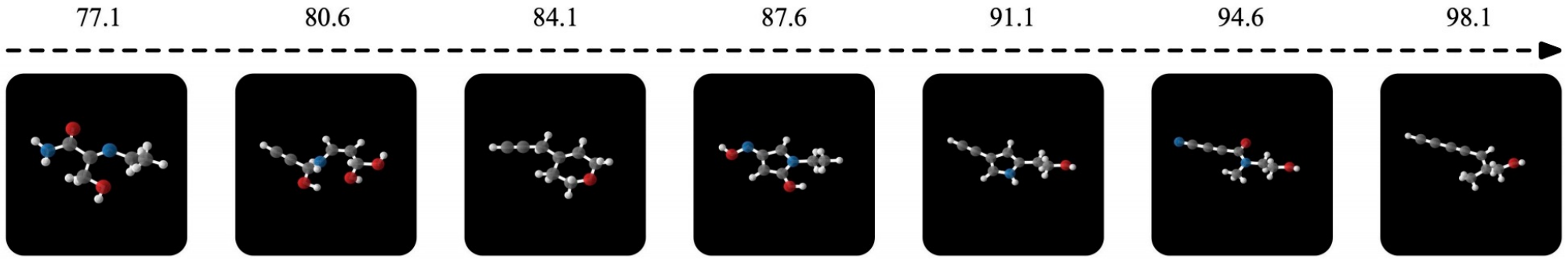
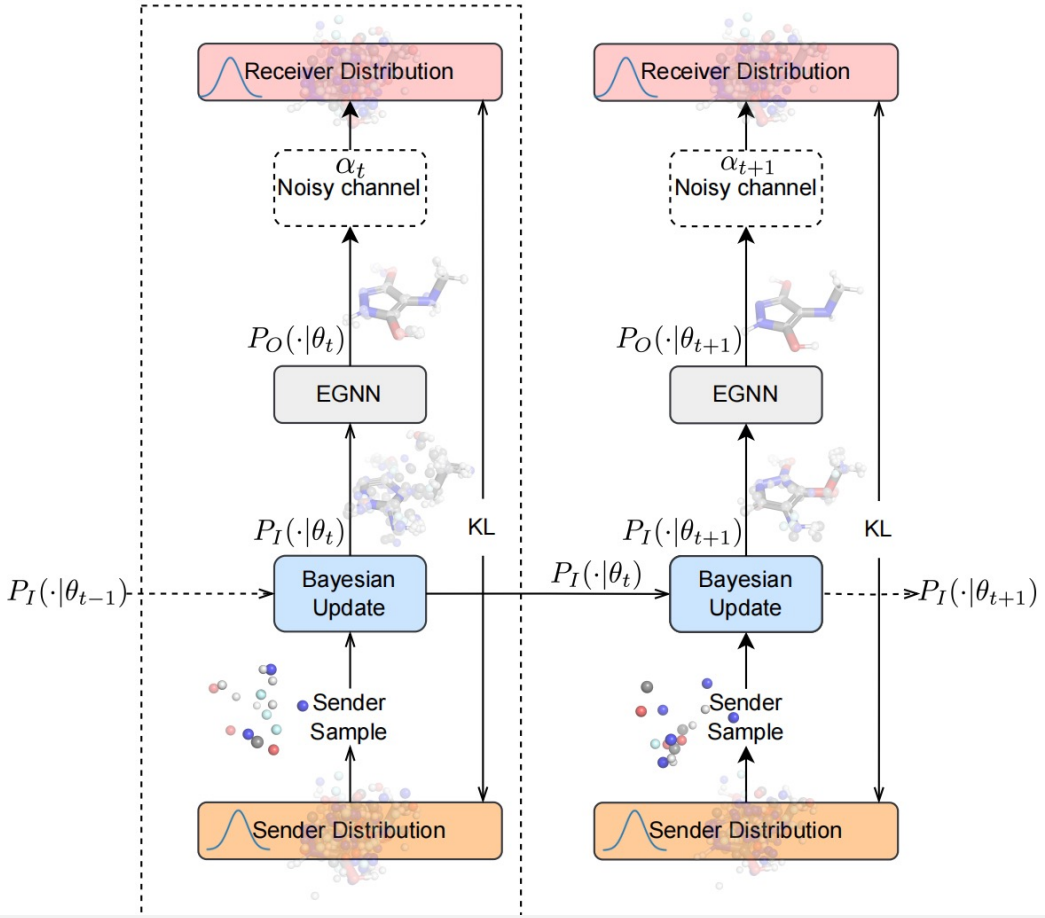


Figure 4. Generated molecules by our Conditional EDM when interpolating among different Polarizability α values with the same reparametrization noise ϵ. Each α value is provided on top of each image.

基于几何贝叶斯流网络的3D分子生成GeoBFN

1. BFN在参数空间进行贝叶斯推断，统一了离散和连续数据的生成算法

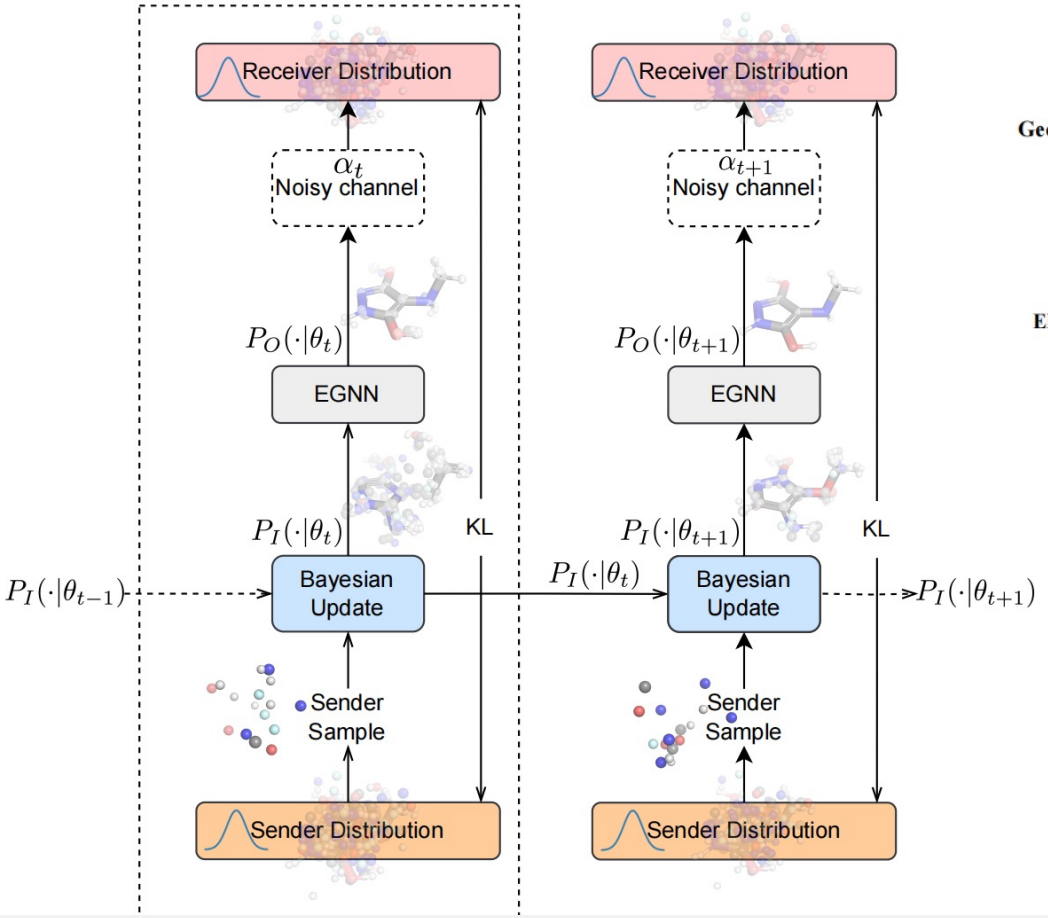


- 将预测的数据分布参数与真实参数的差异作为损失函数，优化神经网络
- 神经网络根据贝叶斯估计结果预测数据分布参数
- 根据当前的噪声数据、上一时刻对分布参数的估计，对分布参数进行贝叶斯估计
- 对数据加噪，噪声大小与时间有关
- 假设数据各维度服从高斯/类别分布

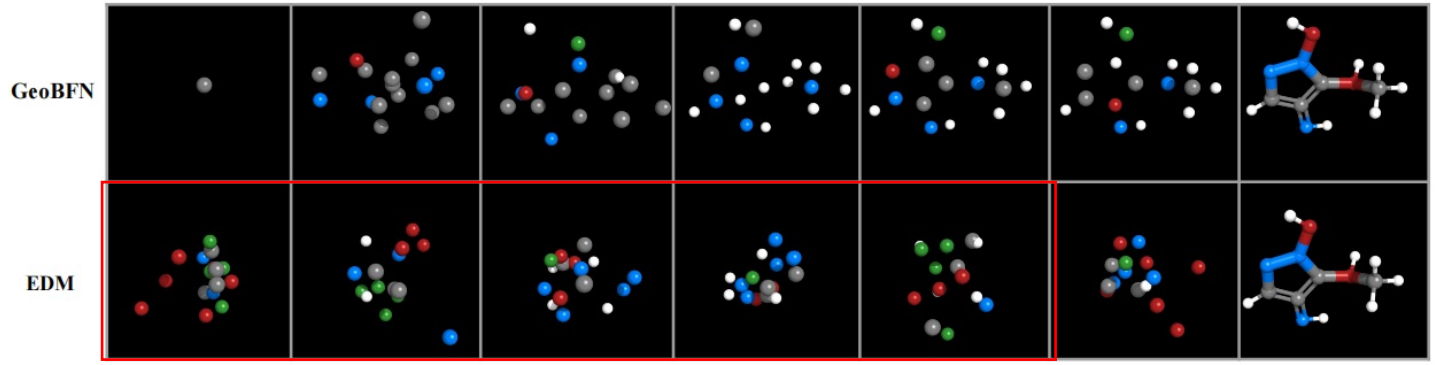
Unified Generative Modeling of 3D Molecules via Bayesian Flow Networks. ICLR 2024. (AIR)

基于几何贝叶斯流网络的3D分子生成GeoBFN

1. BFN在参数空间进行贝叶斯推断，统一了离散和连续数据的生成算法



2. BFN引入的噪声方差远低于扩散模型，适应原子位置的噪声敏感性



△ 扩散模型EDM生成过程中存在大量随机的分子结构，

GeoBFN生成过程更平滑

Unified Generative Modeling of 3D Molecules via Bayesian Flow Networks. ICLR 2024. (AIR)

基于几何贝叶斯流网络的3D分子生成GeoBFN

GeoBFN在无条件生成和条件生成任务上均超过基于扩散模型的EDM

无条件生成

Table 1: Results of atom stability, molecule stability, validity, validity×uniqueness (V×U), and novelty. A higher number indicates a better generation quality. The results marked with an asterisk were obtained from our own tests. And GeoBFN$_k$ denote the results of sampling the molecules with a specific number of steps k

# Metrics	QM9					DRUG	
	Atom Sta (%)	Mol Sta (%)	Valid (%)	V×U (%)	Novelty (%)	Atom Sta (%)	Valid (%)
Data	99.0	95.2	97.7	97.7	-	86.5	99.9
ENF	85.0	4.9	40.2	39.4	-	-	-
G-Schnet	95.7	68.1	85.5	80.3	-	-	-
GDM-AUG	97.6	71.6	90.4	89.5	74.6	77.7	91.8
EDM	98.7	82.0	91.9	90.7	58.0	81.3	92.6
EDM-Bridge	98.8	84.6	92.0	90.7	-	82.4	92.8
GEOLDM	98.9 ± 0.1	89.4 ± 0.5	93.8 ± 0.4	92.7 ± 0.5	57.0	84.4	**99.3**
GeoBFN $_{50}$	98.28 ± 0.1	85.11 ± 0.5	92.27 ± 0.4	90.72 ± 0.3	72.9	75.11	91.66
GeoBFN $_{100}$	98.64 ± 0.1	87.21 ± 0.3	93.03 ± 0.3	91.53 ± 0.3	70.3	78.89	93.05
GeoBFN $_{500}$	98.78 ± 0.8	88.42 ± 0.2	93.35 ± 0.2	91.78 ± 0.2	67.7	81.39	93.47
GeoBFN $_{1k}$	**99.08** ± 0.06	**90.87** ± 0.2	**95.31** ± 0.1	**92.96** ± 0.1	66.4	**85.60**	92.08
GeoBFN $_{2k}$	**99.31** ± 0.03	**93.32** ± 0.1	**96.88** ± 0.1	**92.41** ± 0.1	65.3	**86.17**	91.66

条件生成

针对不同的条件，需要将条件加入去噪网络的输入从头训练。使用预训练的性质预测其来评判条件生成效果。

Table 2: Mean Absolute Error for molecular property prediction with 500 sampling steps. A lower number indicates a better controllable generation result.

Property Units	α Bohr^3	$\Delta\varepsilon$ meV	$\varepsilon_{\text{HOMO}}$ meV	$\varepsilon_{\text{LUMO}}$ meV	μ D	C_v $\frac{\text{cal}}{\text{mol}}$K
QM9*	0.10	64	39	36	0.043	0.040
Random*	9.01	1470	645	1457	1.616	6.857
N_{atoms}	3.86	866	426	813	1.053	1.971
EDM	2.76	655	356	584	1.111	1.101
GEOLDM	2.37	587	340	522	1.108	1.025
GeoBFN	**2.34**	**577**	**328**	**516**	**0.998**	**0.949**

基于直线扩散模型的3D分子生成方法SLDM

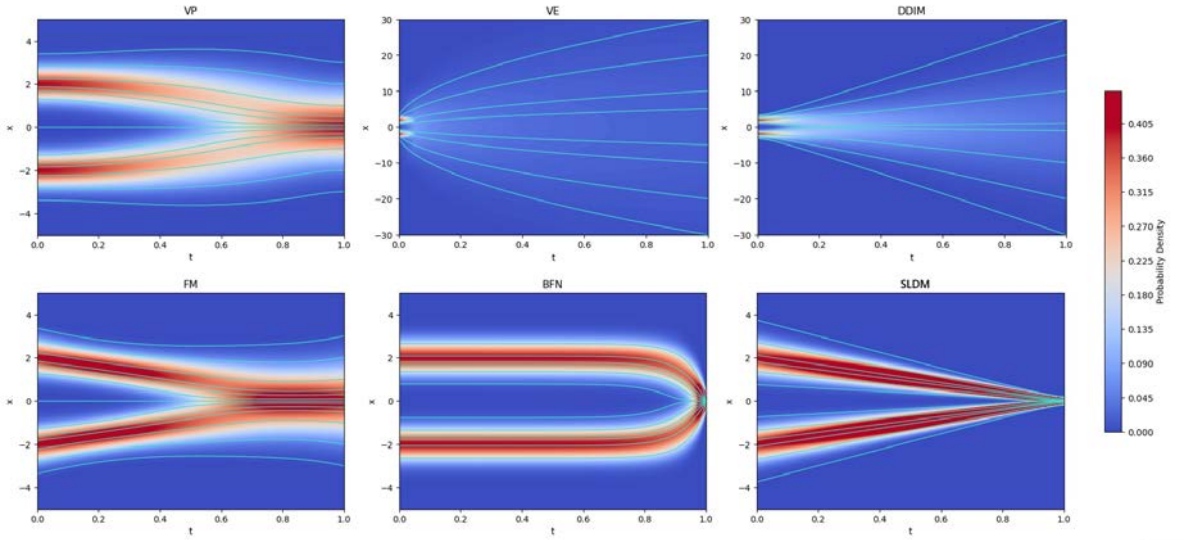
1. 提出直线轨迹扩散生成，加速分子生成效率

SLDM: $x_t = (1 - \mathrm{t})\mathrm{x}_0 + \sigma\epsilon, \epsilon \sim N(0, I), t \in [0, 1].$

Diffusion采样过程可视为ODE离散化

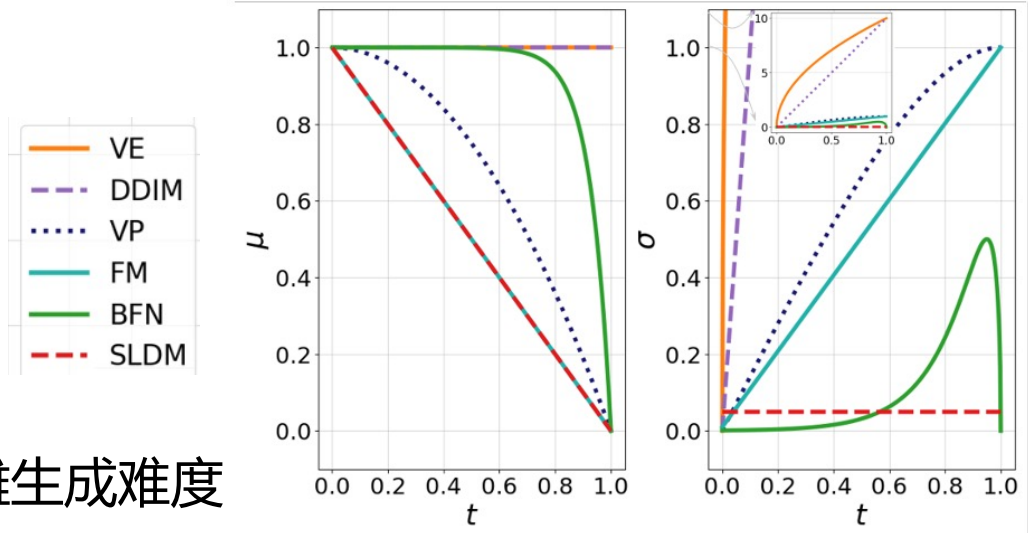
$$\mathrm{d}\boldsymbol{x}_t = \left[\frac{\dot{\mu}(t)}{\mu(t)}\boldsymbol{x}_t - \left(\sigma(t)\dot{\sigma}(t) - \sigma(t)^2\frac{\dot{\mu}(t)}{\mu(t)}\right)\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t.$$

ODE轨迹越直，截断误差越小

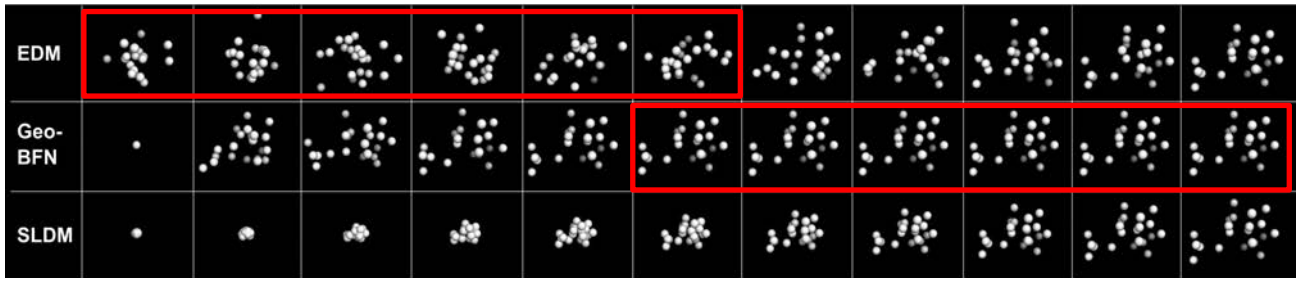


2. 适合分子的生成过程

（1）加噪过程的引入的噪声方差更小



（2）均摊生成难度

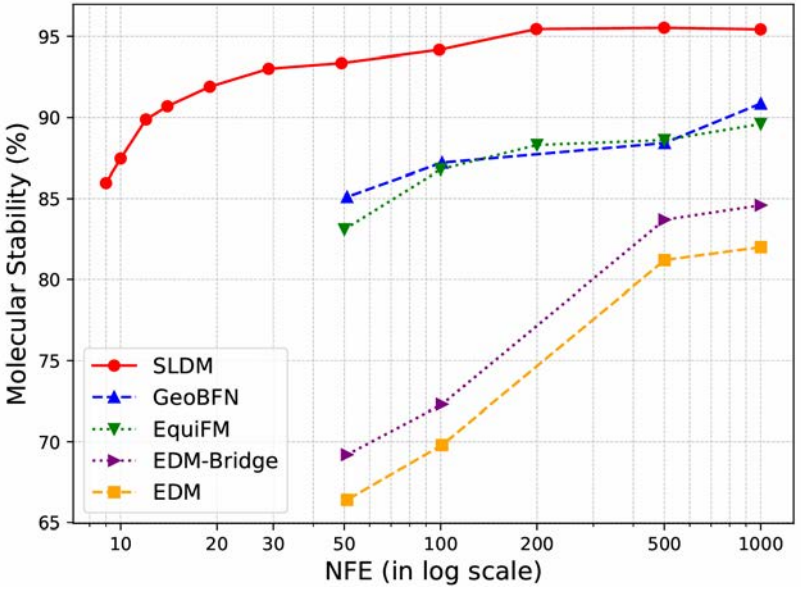


Straight-Line Diffusion Model for Efficient 3D Molecular Generation. Arxiv, 2025. (THU AIR)

基于直线扩散模型的3D分子生成方法SLDM

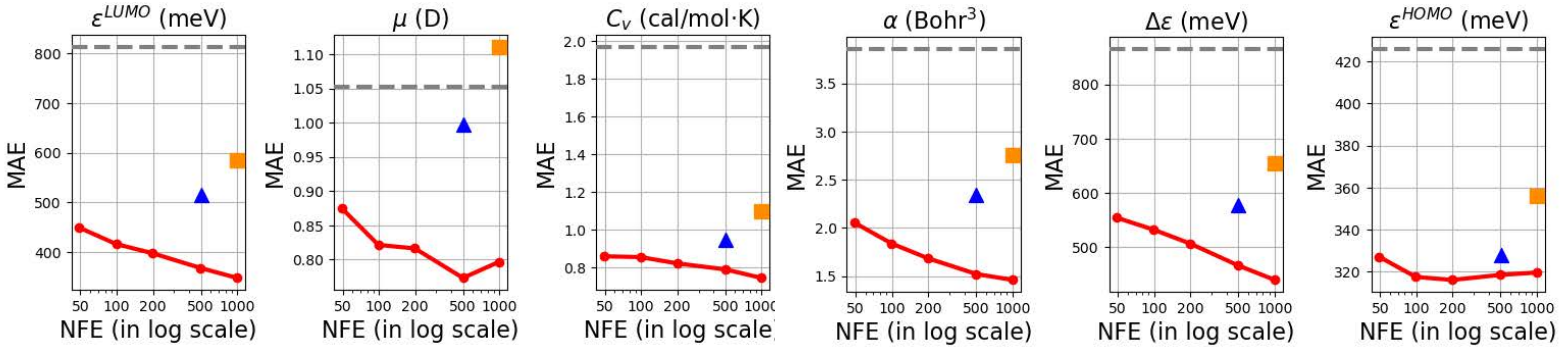
无条件生成：达到新SOTA，大幅加速 1000步 → 10~15步

	QM9				GEOM-Drugs	
#Metrics	Atom sta(%)	Mol sta(%)	Valid(%)	V*U(%)	Atom sta(%)	Valid(%)
Data	99.0	95.2	97.7	97.7	86.5	99.9
E-NF	85.0	4.9	40.2	39.4	-	-
G-Schnet	95.7	68.1	85.5	80.3	-	-
EDM (T=1000)	98.7	82.0	91.9	90.7	81.3	92.6
GDM (T=1000)	97.6	71.6	90.4	89.5	77.7	91.8
EDM-Bridge (T=1000)	98.8	84.6	92.0	90.7	82.4	92.8
GeoLDM (T=1000)	98.9	89.4	93.8	92.7	84.4	99.3
EquiFM (T=200)	98.9	88.3	94.7	93.5	84.1	98.9
GeoBFN (T=1000)	99.08	90.87	95.31	92.96	85.60	92.08
SLDM (T=1000)	**99.43**	**95.42**	**97.07**	90.42	88.30	**99.95**
SLDM (T=50)	99.30	93.37	96.24	**93.63**	**89.03**	99.57



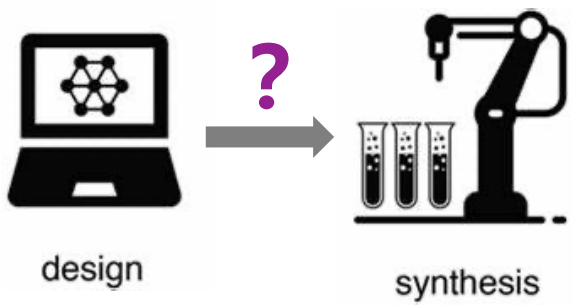
条件生成：达到新SOTA，加速 1000步 → 50步

Property Units	α Bohr^3	$\Delta\epsilon$ meV	ϵ^{HOMO} meV	ϵ^{LUMO} meV	μ D	C_v $\frac{\text{cal}}{\text{mol K}}$
Data	0.10	64	39	36	0.043	0.040
Random	9.01	1470	645	1457	1.616	6.857
N_{atoms}	3.86	866	426	813	1.053	1.971
EDM	2.76	655	356	584	1.111	1.101
GeoLDM	2.37	587	340	522	1.108	1.025
GeoBFN	2.34	577	328	516	0.998	0.949
SLDM	**1.46**	**440**	**320**	**348**	**0.797**	**0.745**



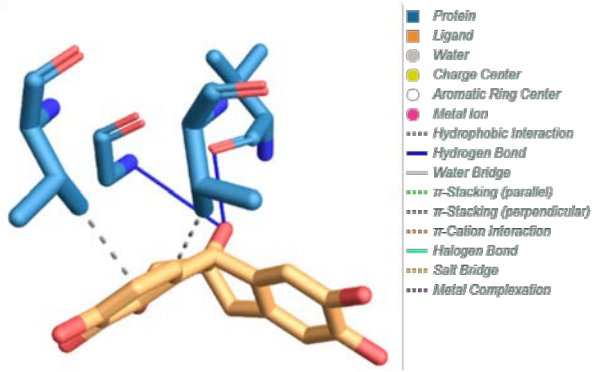
挑战性问题

可合成性问题



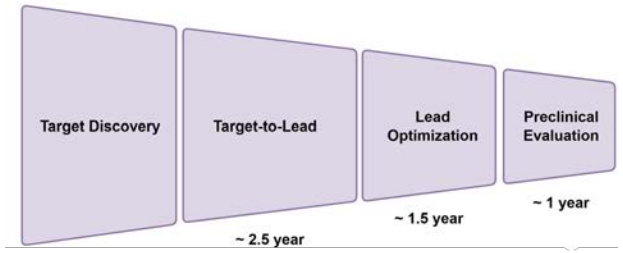
模型生成的分子，如何获取对应的合成路径?

结合特异性问题



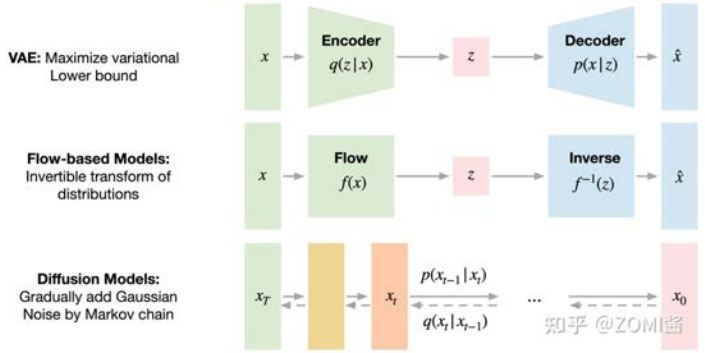
如何提升生成分子与靶点的结合特异性?

成药性问题



分子生成如何综合考虑药物研发各个阶段的约束?

方法论问题

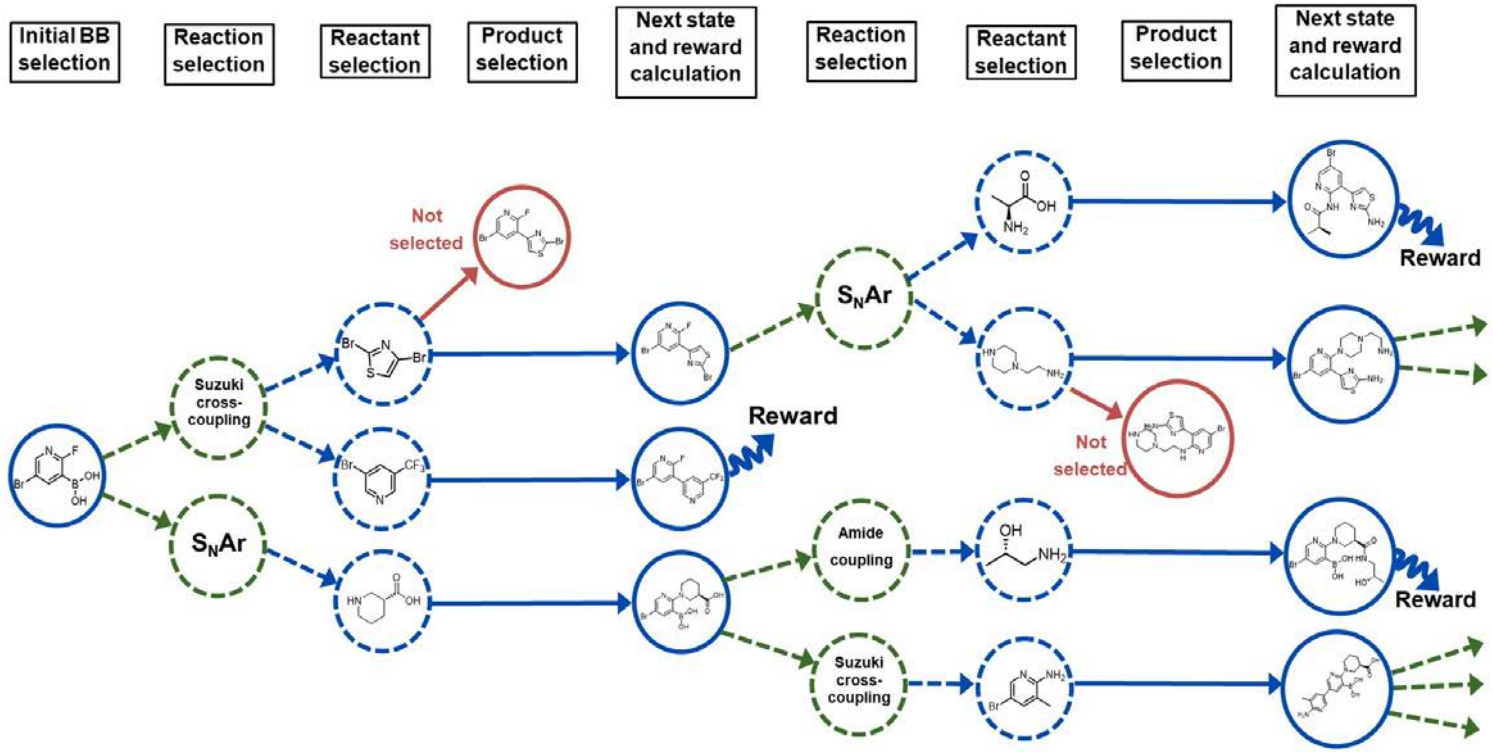


不同生成方法各有缺陷，能否探索新的生成方法?

可合成性问题

核心思想：将分子生成过程建模为一个在化学反应空间中的 Markov 决策过程。

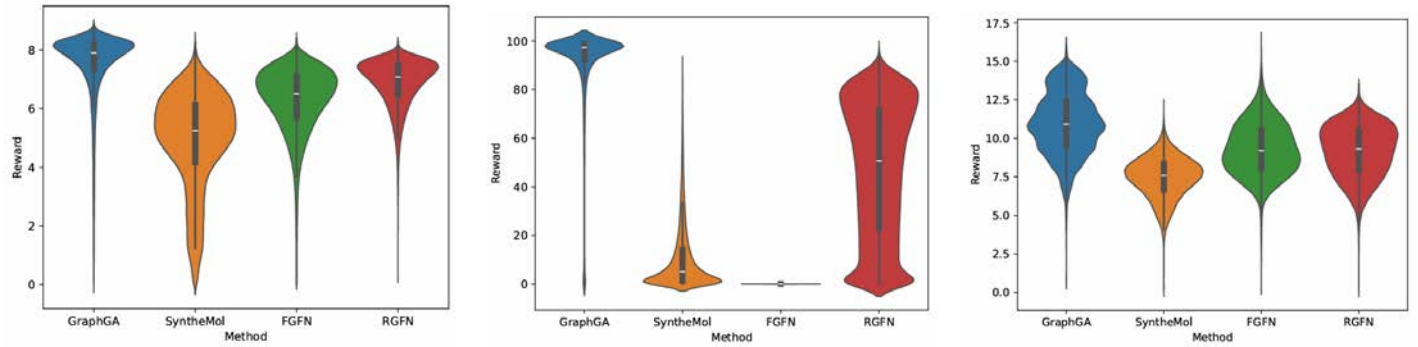
训练过程：从反应物集合选择反应物，从高产化学反应中选择反应，使用 RDKit 模拟反应结果。重复进行直到选择停止操作。使用GFlowNet学习策略，利用奖励函数对获得的分子进行评估，优化模型效果。



RGFN: Synthesizable Molecular Generation Using GFlowNets. Neurips, 2024. (Mila)

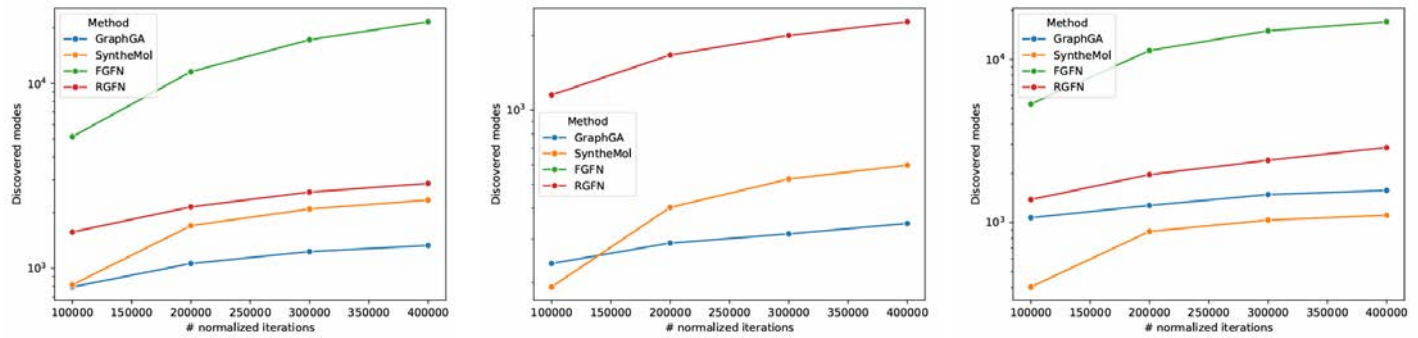
可合成性问题

在保证可合成的前提下，达到和一般方法接近的靶点亲和力以及多样性。



(a) sEH (proxy) (b) senolytics (proxy) (c) ClpP (docking)

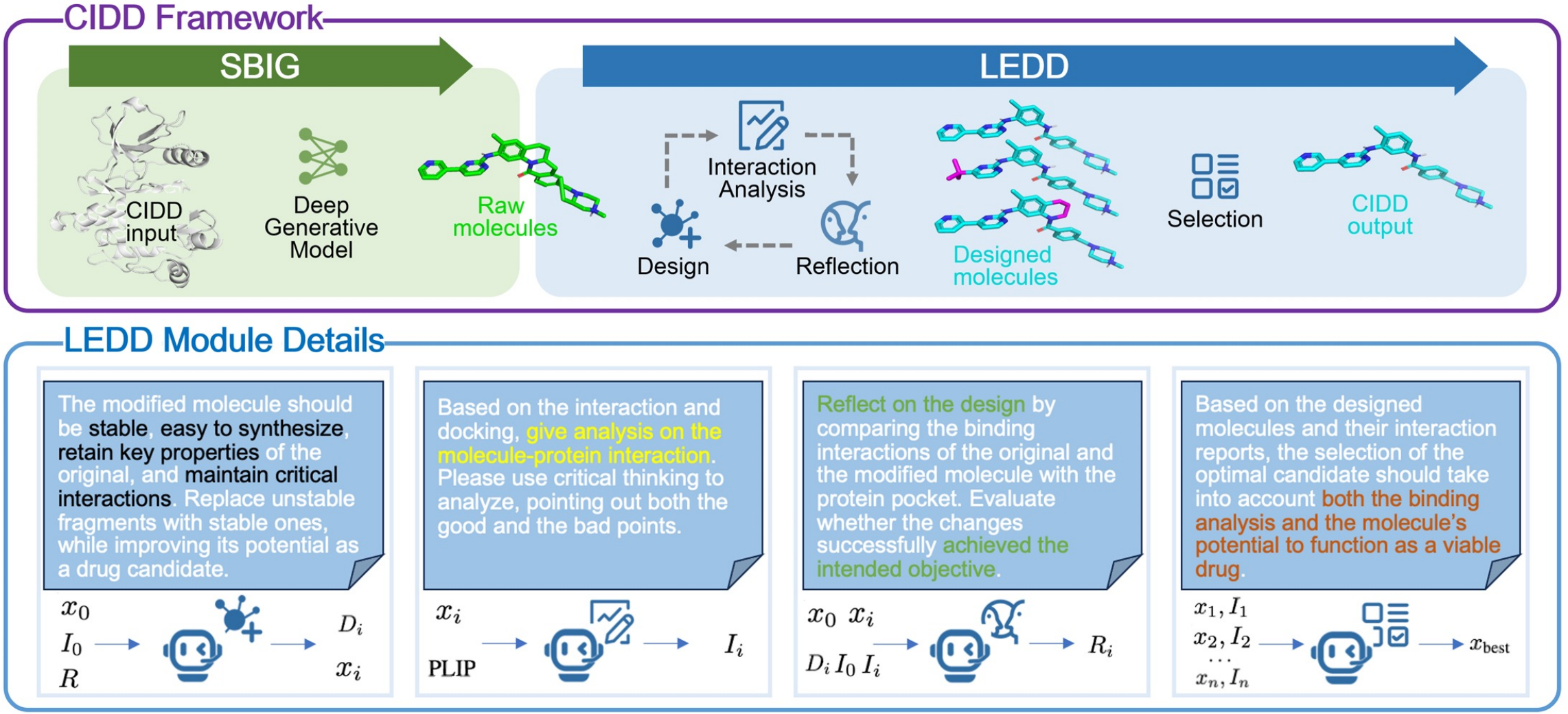
Figure 3: Distributions of rewards across different tasks.



(a) sEH (proxy) (b) senolytics (proxy) (c) ClpP (docking)

Figure 4: Number of discovered modes as a function of normalized iterations. Log scale used.

成药性问题：大模型指导的3D分子生成方法



Pushing the boundaries of Structure-Based Drug Design through Collaboration with Large Language Models. Arxiv, 2025. (AIR)

成药性问题：大模型指导的3D分子生成方法

CrossDocked2020

Method	Vina ↓	QED ↑	SA score ↑	MRR ↑	AUR ↓	Success Ratio ↑	QikProp ↑	Diversity ↑
AR	-6.737	0.507	0.635	56.67%	34.72%	3.28%	18.66%	0.836
Pocket2Mol	-7.246	**0.573**	**0.758**	67.88%	20.14%	14.60%	29.58%	0.866
TargetDiff	-7.452	0.474	0.579	37.81%	43.40%	3.04%	27.63%	**0.890**
DecompDiff	-8.260	0.444	0.609	62.60%	21.76%	15.72%	29.04%	0.877
MolCRAFT	-7.783	0.503	0.685	58.47%	25.59%	13.72%	22.37%	0.870
CIDD	**-9.019**	0.525	0.694	**76.54%**	**11.44%**	**37.94%**	**37.54%**	0.870

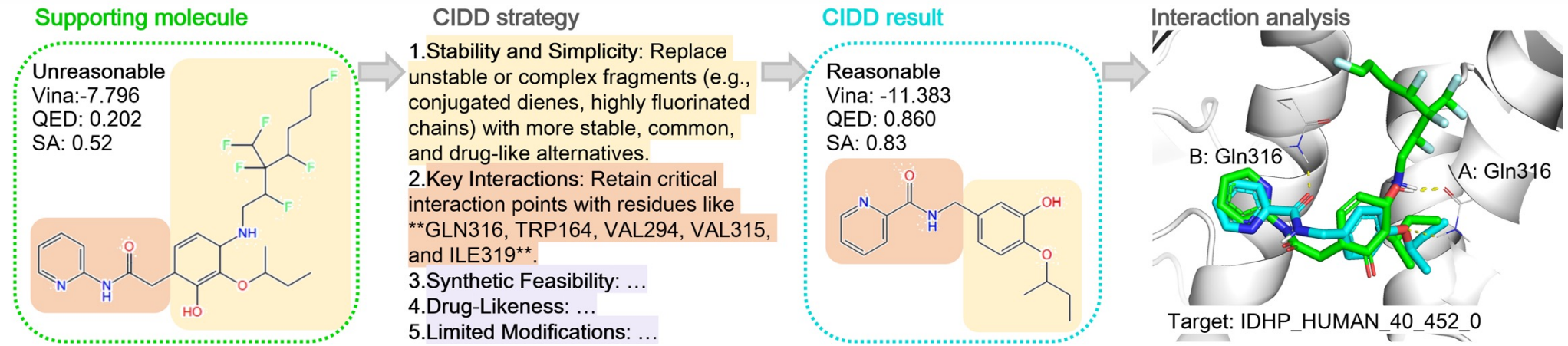
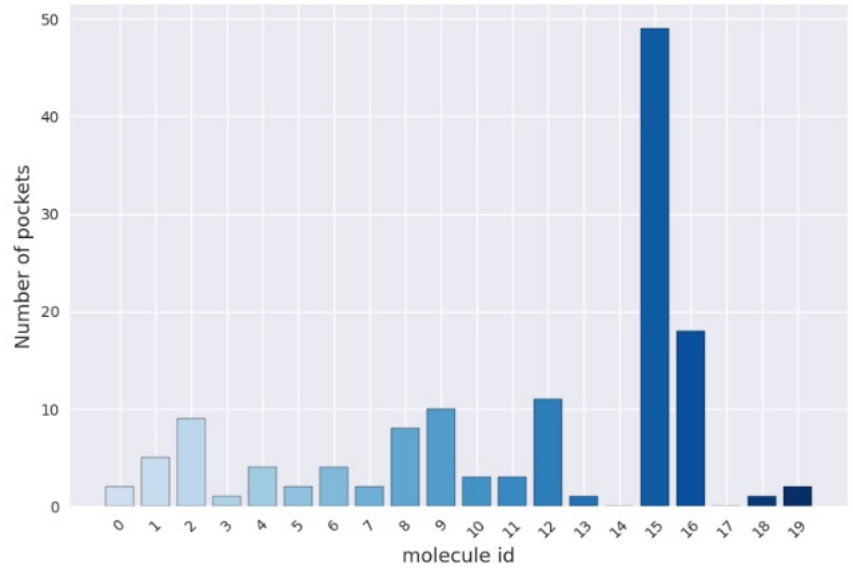


Figure 4. **Generation Case and Corresponding Design Strategy Produced by CIDD.**

结合特异性问题

特异性在药物设计中至关重要；非特异性分子可能与非预期靶标结合，从而可能导致不良反应并降低疗效。



(b) The number of pockets which surpass true target on docking score

现有模型生成分子的特异性较差

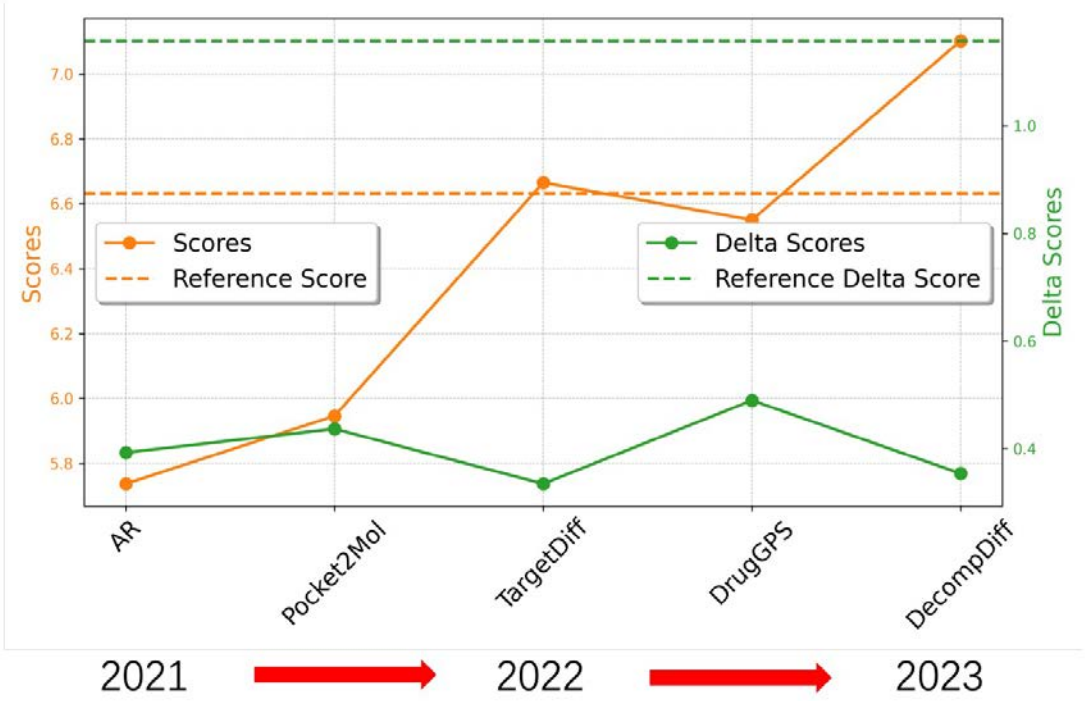
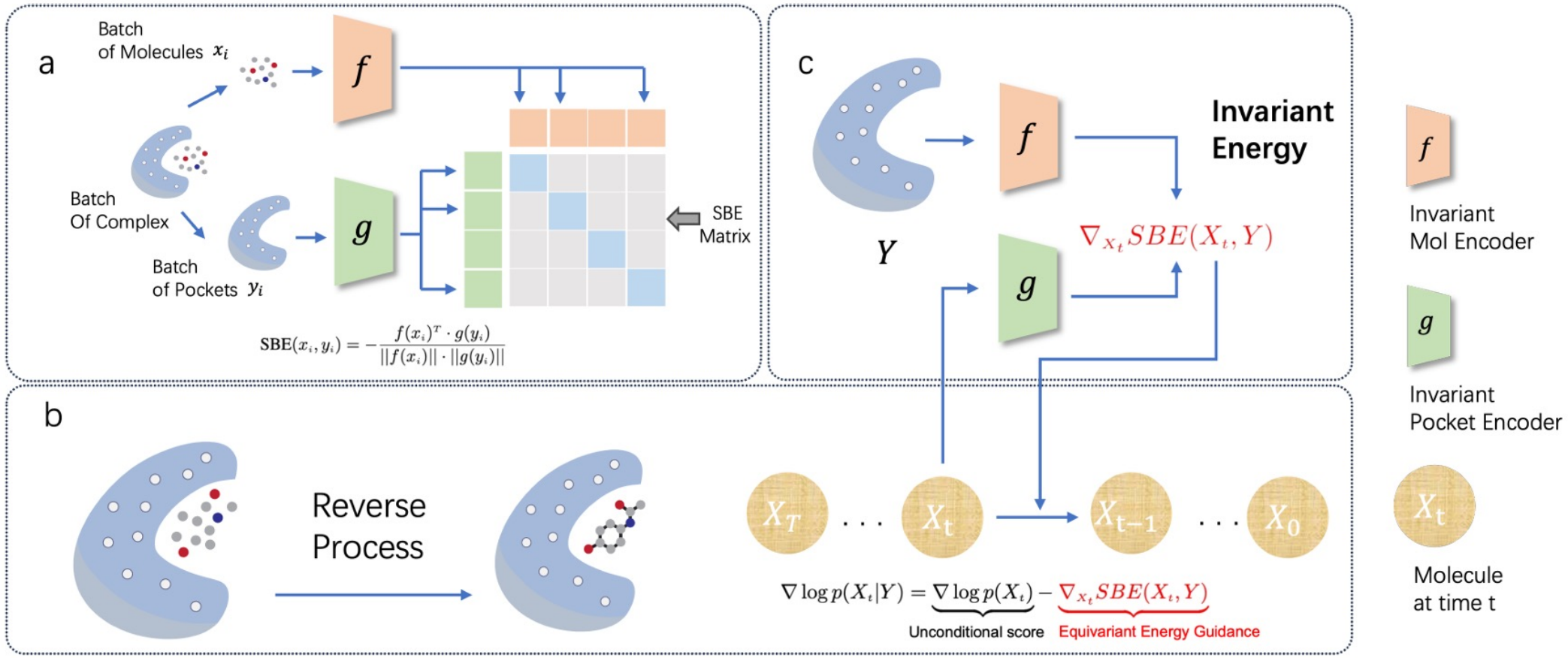


Figure 3. The evolution of absolute docking scores and delta scores obtained by various methods, organized chronologically.

Rethinking Specificity in SBDD: Leveraging Delta Score and Energy-Guided Diffusion. ICML, 2024. (AIR)

能量引导的扩散模型

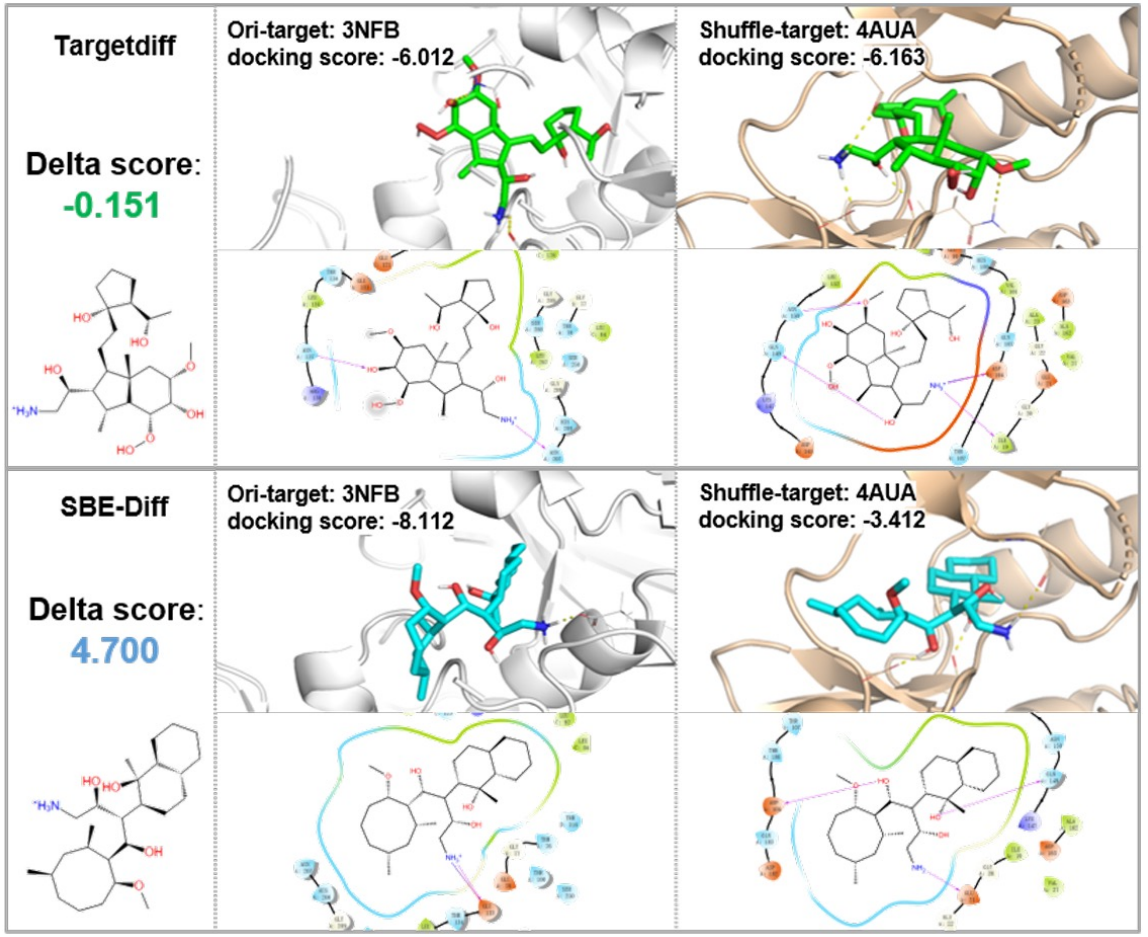


能量引导扩散生成，能量函数基于对比学习训练，刻画特异性结合亲和力。

生成分子的特异性效果

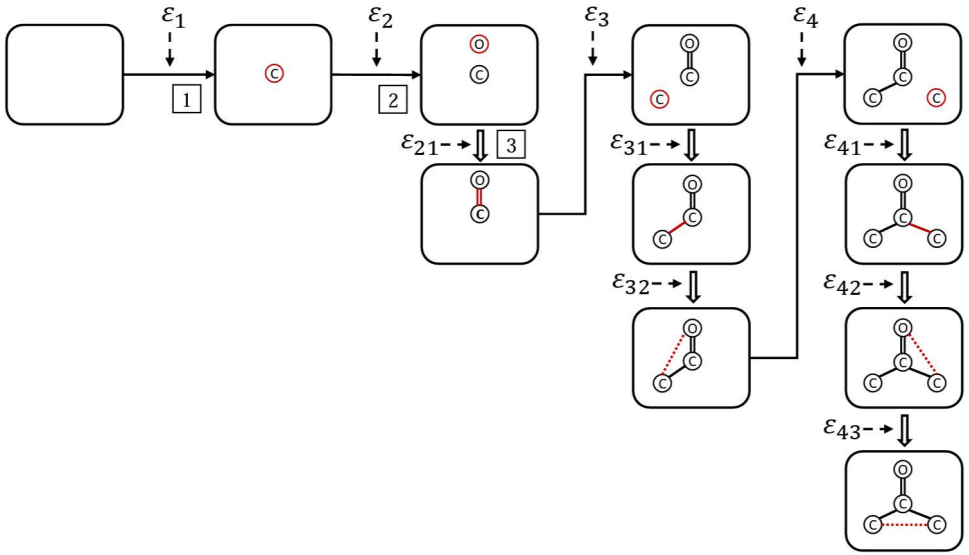
Table 2. Experimental Results for different methods. We show the absolute docking score, delta score, and ratio of the generated molecule better than the reference ligand. For each metric, the best result is **bold** and the second best result is underlined.

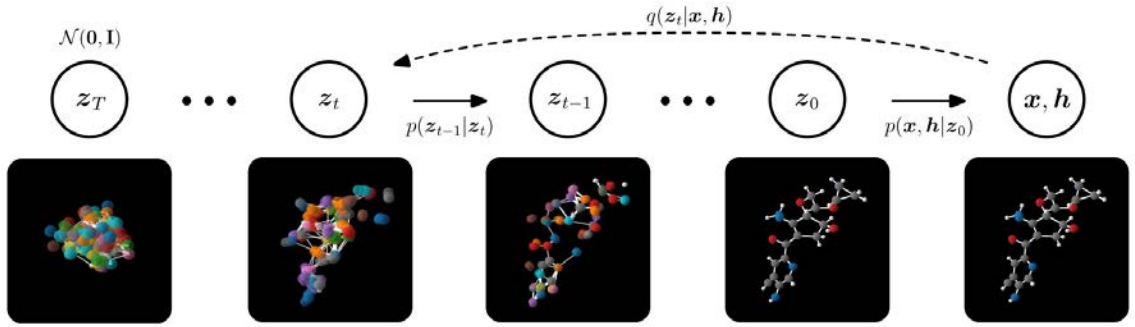
	Absolute ↑	Delta ↑		Ratio ↑	
Methods	mean	mean	median	mean	median
trainset	5.727	0.044	-0.073	0.159	0.053
AR	5.737	0.393	0.200	0.150	0.039
Pocket2Mol	5.946	0.437	0.106	0.170	0.050
DrugGPS	6.554	0.490	**0.387**	0.235	0.134
TargetDiff	6.665	0.335	0.102	0.259	0.129
DecompDiff	**7.102**	0.354	0.220	0.274	0.133
SBE-Diff	6.815	**0.552**	0.250	**0.300**	**0.216**
Reference	6.632	1.158	1.029	-	-



Rethinking Specificity in SBDD: Leveraging Delta Score and Energy-Guided Diffusion. ICML, 2024. (AIR)

方法论：Diffusion v.s. Autoregressive





Autoregressive生成

适合1D序列生成

- 长序列生成易累积误差；
- 分子图生成需额外设计生成顺序。

Diffusion生成

- 数值精度高，适合3D空间坐标生成
- 多模态（连续 + 离散）生成存在挑战
- 生成过程中原子数固定，限制生成灵活性

方法论：Diffusion v.s. Autoregressive

多层次分子生成模型，不同层次使用不同生成方法，可变长且减少误差累积

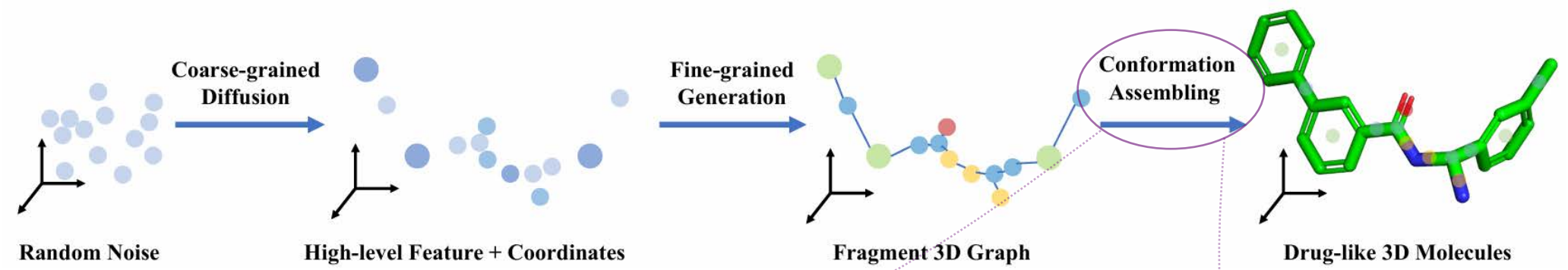


Figure 3. An overview of the hierarchical diffusion model.

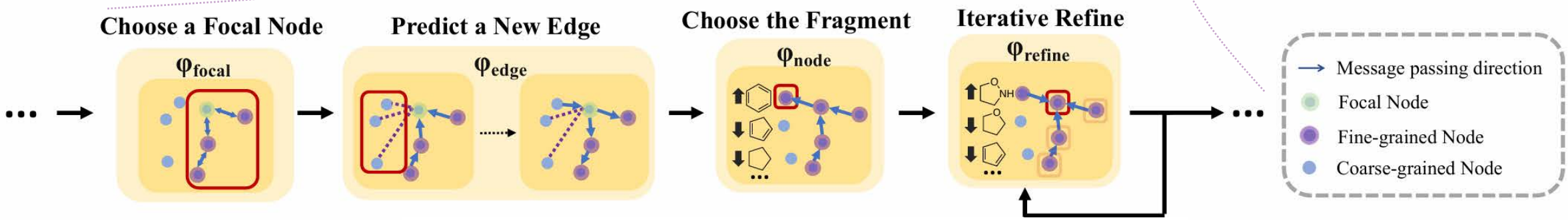
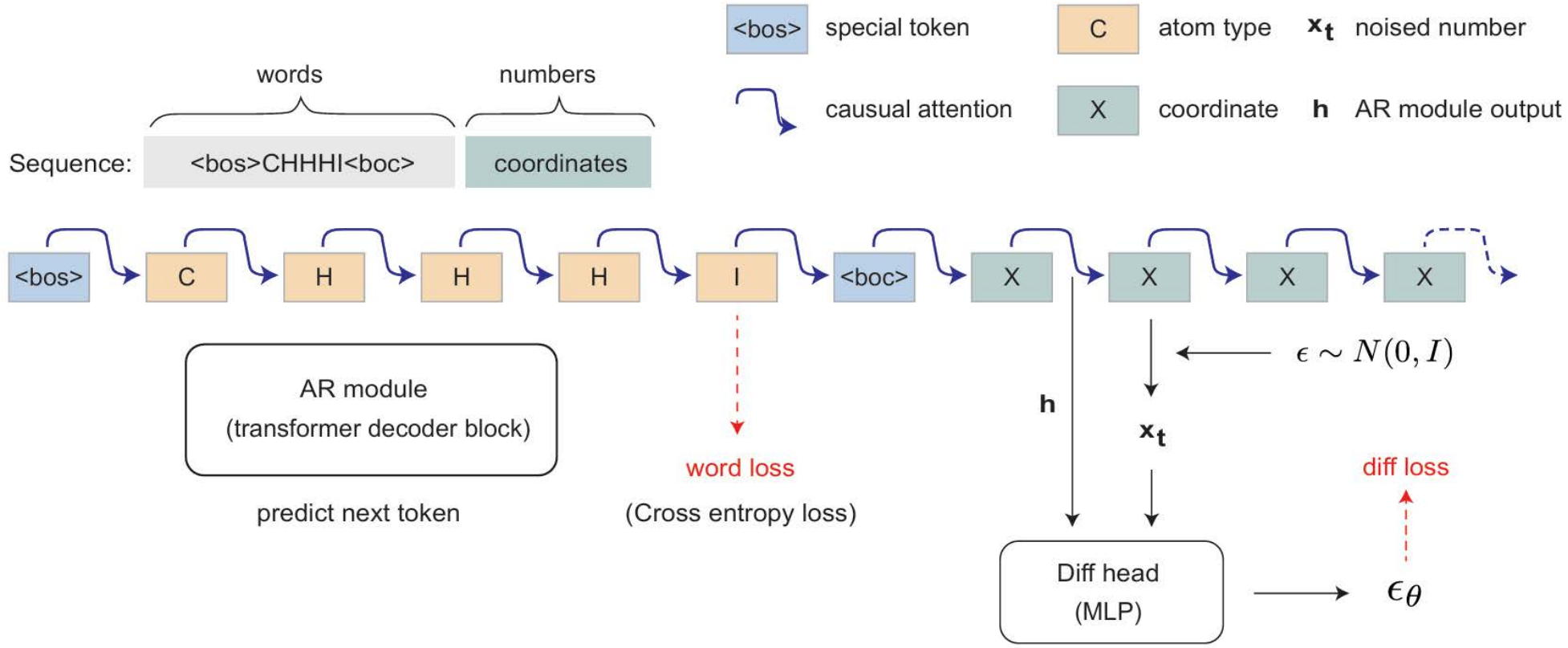


Figure 6. Illustration of the fine-grained atom generation process

Coarse-to-Fine: a Hierarchical Diffusion Model for Molecule Generation in 3D. ICML, 2023. (AIR)

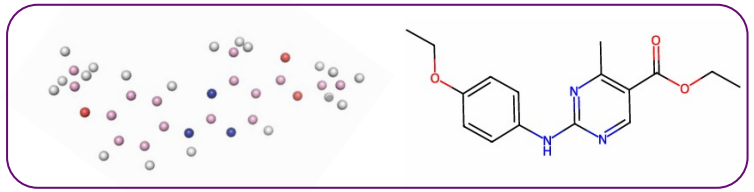
方法论：Diffusion v.s. Autoregressive

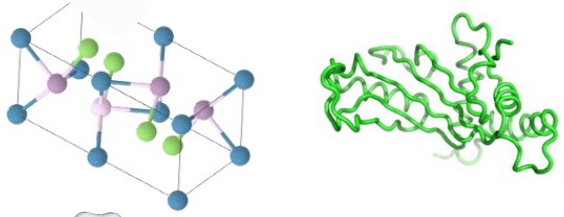
多模态融合生成模型，各模态使用更适合的生成方法

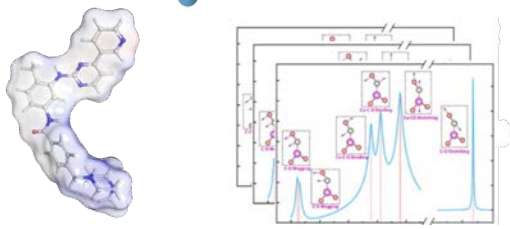


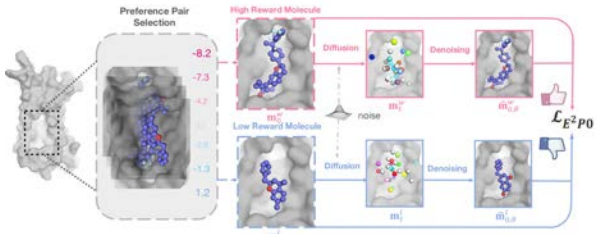
UniGenX: Unified Generation of Sequence and Structure with Autoregressive Diffusion. Arxiv, 2025. (MSRA)

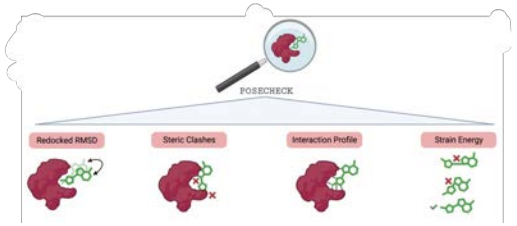
未来发展趋势











- **多模态生成融合**

  统一 2D 化学结构与 3D 构象信息，提升分子合法性

- **通用分子生成范式**

  支持多类型分子体系（小分子、晶体、蛋白等）的统一建模

- **物理知识引入**

  融合电荷密度、分子轨道、光谱等物理量，拓展分子生成模型的科学适用性

- **结合预训练与强化学习**

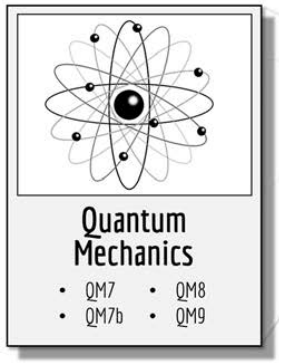
  结合预训练表示与强化学习优化，提升生成模型的泛化与可控性

- **基础数据与评估体系建设**

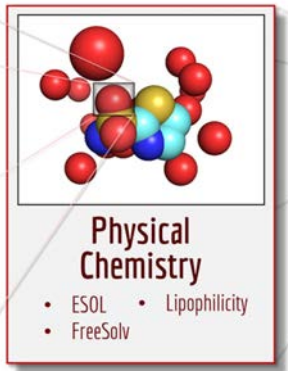
  推动高质量 benchmark 数据集建设，发展多维度、任务相关的评价指标

分子性质预测

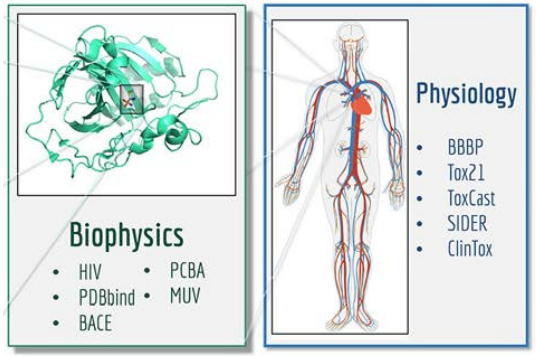
分子性质预测：包括宏观生理性质，物理化学性质以及微观的量子力学性质等不同类型和尺度的性质评测，是评价分子表示学习方法最常用的下游任务。



量子力学性质



物理化学性质



生理性质

微观

宏观

分子基础大模型

海量无标签的分子数据

PubChem 93M

ZINC 35M

OGB OPEN GRAPH BENCHMARK PCQM4Mv2 3M

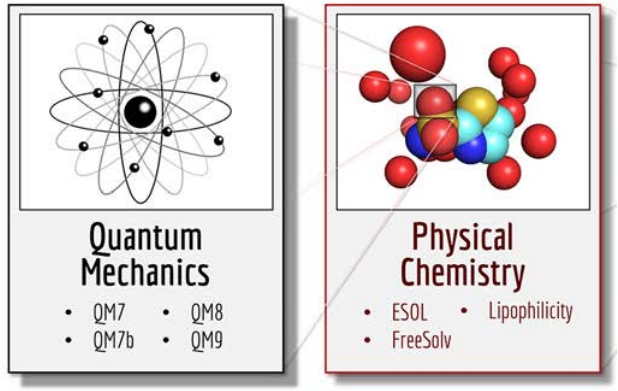
ChEMBL 2.4M

RCSB PDB PROTEIN DATA BANK 0.1M

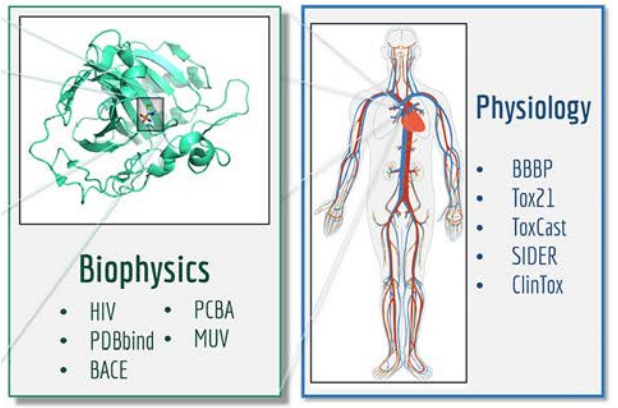
ChemSpider The free chemical database 130M

分子基础大模型

下游少量有标签数据



7K~0.1M 600~4K



5K~0.4M 1K~8K

分子表示学习方法分类

基于序列的方法

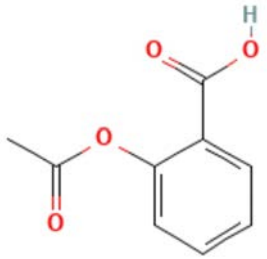
CC(=O)OC1=CC=CC=C1C(=O)O

建模和训练策略借鉴**NLP**，采用**Transformer**结构编码，策略采用**MLM**和生成式(Auto-regressive)

优点：建模简单，预训练数据容易获取，易于scale

缺点：缺乏结构建模，无法用于微观任务

基于2D分子图的方法

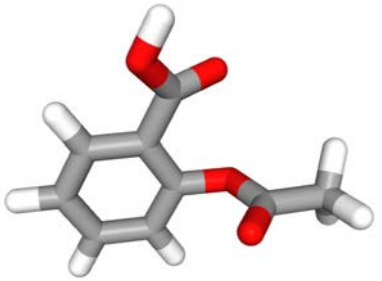


采用GNN网络，策略基于2D图的MLM和生成式预训练

优点：GNN参数量小，结构感知能力强

缺点：不易Scale，无空间结构

基于3D分子结构

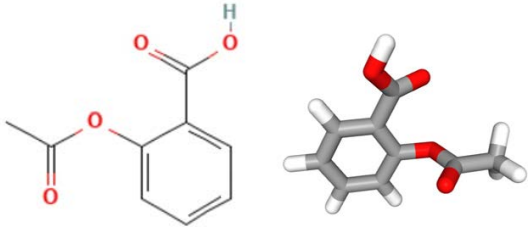


Denosing是3D结构最有效的预训练方法

优点：可以覆盖全类型性质预测任务

缺点：对3D结构准确性有要求，数据量少

多模态融合



2D的共价键拓扑结构可以和3D空间结构产生互补

优点：模态信息互补，可以学习层次化表示

缺点：网络计算量增加

基于序列的分子大模型 ChemBERTa

- **动机**：借鉴NLP建模方法对分子1D序列进行建模，**首个**基于 Transformer 的分子表征方法，以MLM作为预训练目标
- **模型结构**：采用RoBERTa
- **预训练数据**：来自 PubChem 的 77M个SMILES序列
- **下游任务**： MoleculeNet

	BBBP 2,039		ClinTox (CT_TOX) 1,478		HIV 41,127		Tox21 (SR-p53) 7,831	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
ChemBERTa 10M	0.643	0.620	0.733	0.975	0.622	0.119	**0.728**	0.207
D-MPNN	**0.708**	0.697	**0.906**	**0.993**	0.752	0.152	0.688	**0.429**
RF	0.681	0.692	0.693	0.968	**0.780**	**0.383**	0.724	0.335
SVM	0.702	**0.724**	0.833	0.986	0.763	0.364	0.708	0.345

Table 1: Comparison of ChemBERTa pretrained on 10M PubChem compounds and Chemprop baselines on selected MoleculeNet tasks. We report both ROC-AUC and PRC-AUC to give a full picture of performance on class-imbalanced tasks.

简单套用NLP方法用并不能产生SOTA的效果

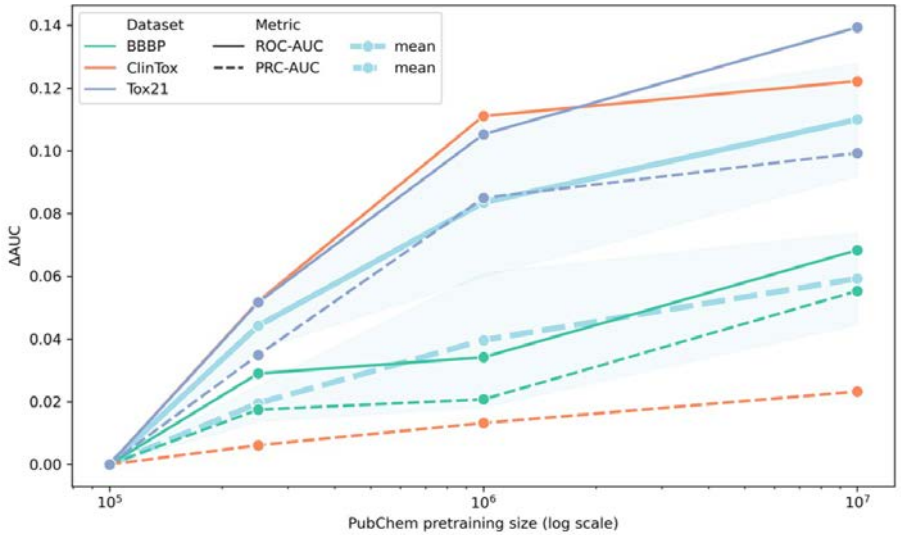


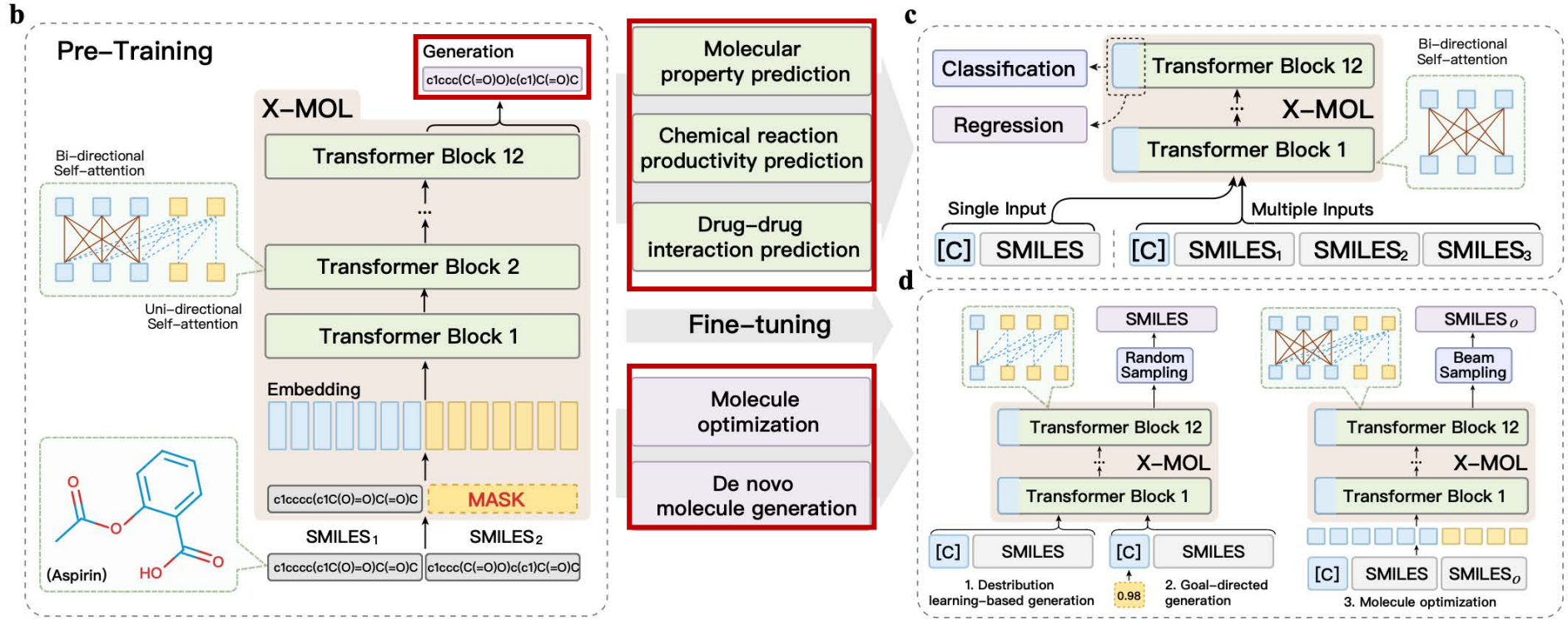
Figure 1: Scaling the pretraining size (100K, 250K, 1M, 10M) produces consistent improvements in downstream task performance on BBBP, ClinTox, and Tox21. Mean ΔAUC across all three tasks with a 68% confidence interval is shown in light blue.

预训练数据集上的scaling law

ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. Preprint, 2020. (DeepChem)

基于序列的分子大模型 X-MOL

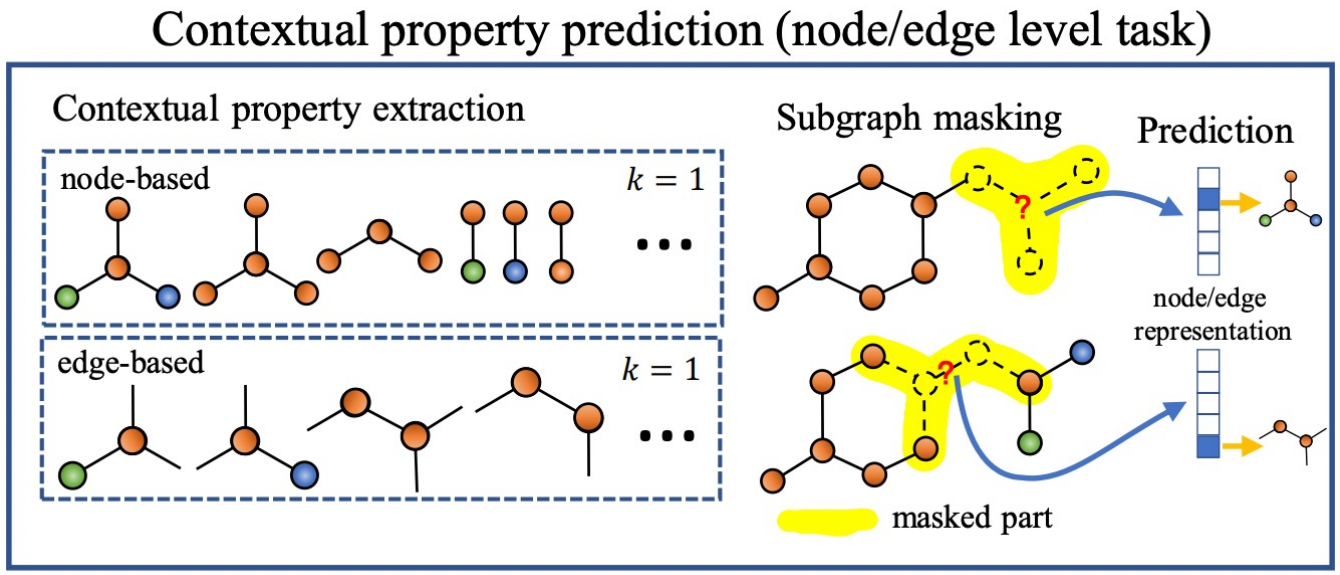
- **动机**：利用序列建模的优势，采用大量的预训练数据进行表示学习（1.1B Molecules）
- **生成式预训练任务**：同一种分子不同的SMILES之间的相互生成(Seq2Seq)
- **优势**：得益于生成式预训练，可同时应用于分子理解（性质预测）和分子生成等下游任务



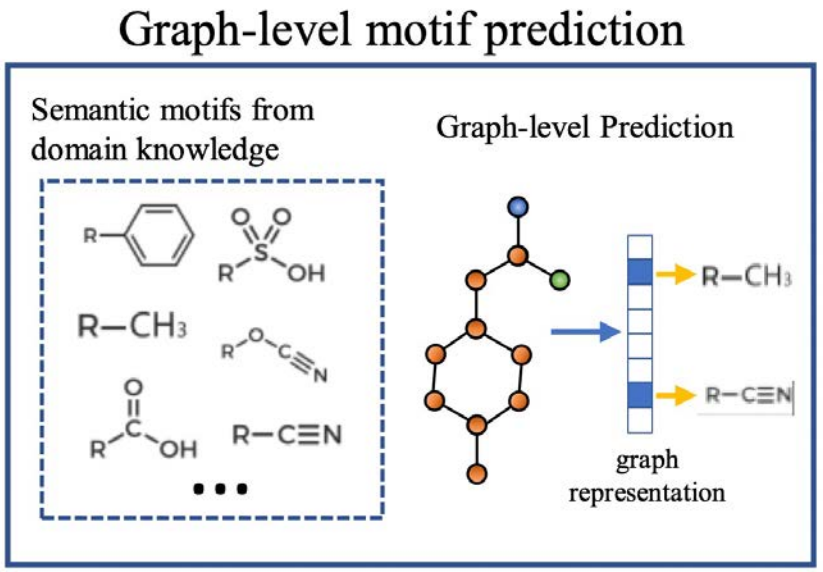
X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. Preprint, 2020. (Baidu)

基于2D分子图的分子大模型 GROVER

- 使用11M的分子数据，在2D Graph上面进行大规模预训练
- 针对分子Graph设计两类自监督学习任务：Graph Masking以及Graph-level motif prediction



在2D图上进行Mask node和edge，利用上下文还原



预测2D分子图是否包含预定义的Motif

Self-Supervised Graph Transformer on Large-Scale Molecular Data. NeurIPS, 2020. (Tencent)

基于2D分子图的分子大模型 GROVER

- 针对Graph设计GNN Transformer架构，在下游任务中表现出良好的scale性能

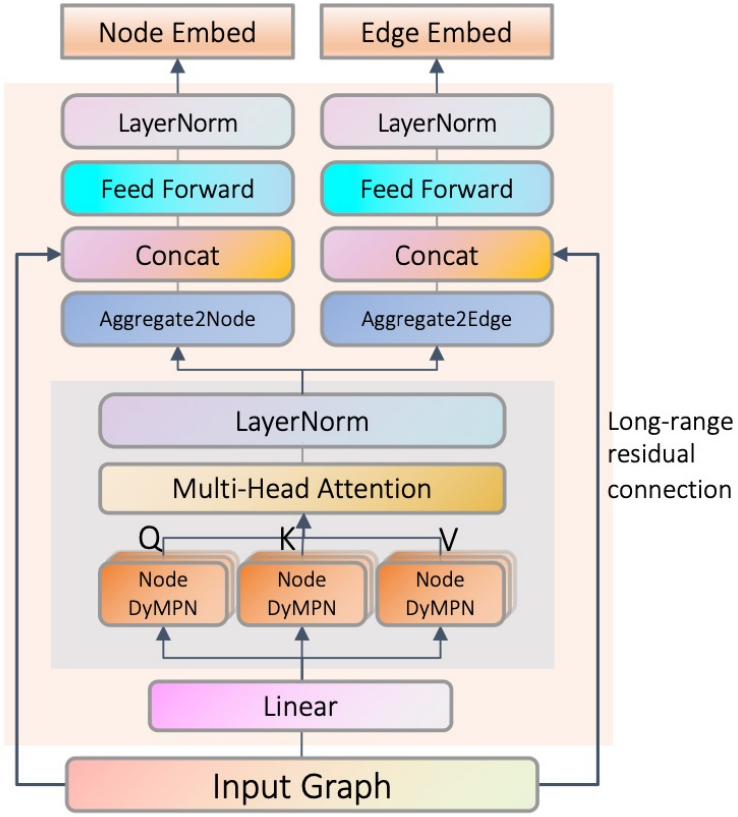


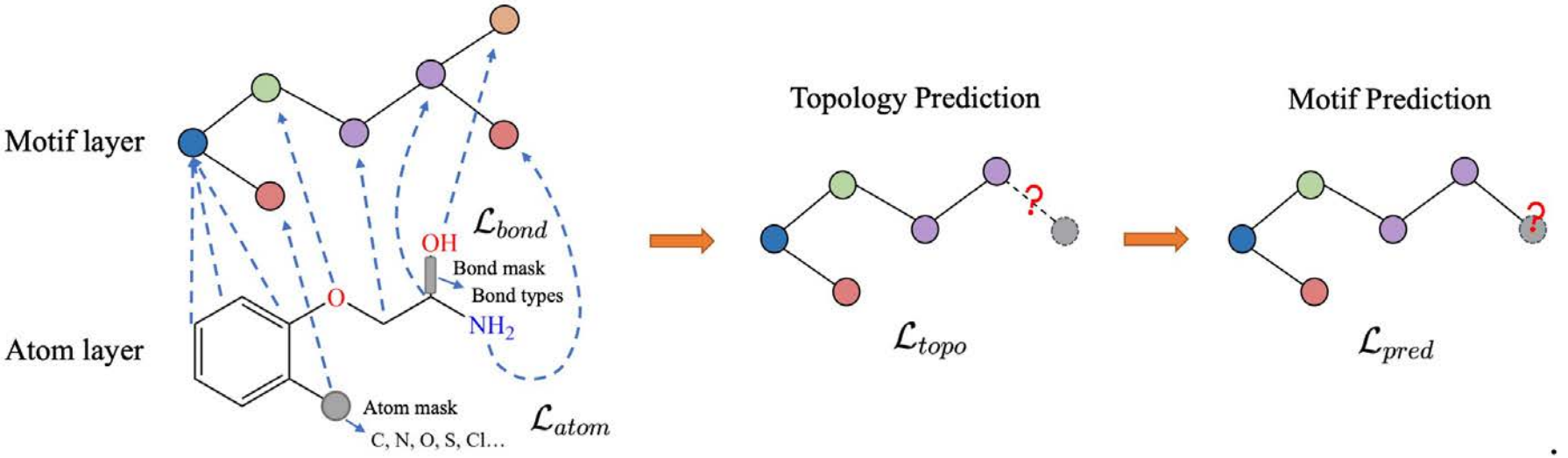
Figure 1: Overview of GTransformer.

Table 1: The performance comparison. The numbers in brackets are the standard deviation. The methods in green are pre-trained methods.

Classification (Higher is better)						
Dataset # Molecules	BBBP 2039	SIDER 1427	ClinTox 1478	BACE 1513	Tox21 7831	ToxCast 8575
TF_Robust [40]	$0.860_{(0.087)}$	$0.607_{(0.033)}$	$0.765_{(0.085)}$	$0.824_{(0.022)}$	$0.698_{(0.012)}$	$0.585_{(0.031)}$
GraphConv [24]	$0.877_{(0.036)}$	$0.593_{(0.035)}$	$0.845_{(0.051)}$	$0.854_{(0.011)}$	$0.772_{(0.041)}$	$0.650_{(0.025)}$
Weave [23]	$0.837_{(0.065)}$	$0.543_{(0.034)}$	$0.823_{(0.023)}$	$0.791_{(0.008)}$	$0.741_{(0.044)}$	$0.678_{(0.024)}$
SchNet [45]	$0.847_{(0.024)}$	$0.545_{(0.038)}$	$0.717_{(0.042)}$	$0.750_{(0.033)}$	$0.767_{(0.025)}$	$0.679_{(0.021)}$
MPNN [13]	$0.913_{(0.041)}$	$0.595_{(0.030)}$	$0.879_{(0.054)}$	$0.815_{(0.044)}$	$0.808_{(0.024)}$	$0.691_{(0.013)}$
DMPNN [63]	$0.919_{(0.030)}$	$0.632_{(0.023)}$	$0.897_{(0.040)}$	$0.852_{(0.053)}$	$0.826_{(0.023)}$	$0.718_{(0.011)}$
MGCN [30]	$0.850_{(0.064)}$	$0.552_{(0.018)}$	$0.634_{(0.042)}$	$0.734_{(0.030)}$	$0.707_{(0.016)}$	$0.663_{(0.009)}$
AttentiveFP [61]	$0.908_{(0.050)}$	$0.605_{(0.060)}$	$0.933_{(0.020)}$	$0.863_{(0.015)}$	$0.807_{(0.020)}$	$0.579_{(0.001)}$
N-GRAM [29]	$0.912_{(0.013)}$	$0.632_{(0.005)}$	$0.855_{(0.037)}$	$0.876_{(0.035)}$	$0.769_{(0.027)}$	-
HU. et.al[18]	$0.915_{(0.040)}$	$0.614_{(0.006)}$	$0.762_{(0.058)}$	$0.851_{(0.027)}$	$0.811_{(0.015)}$	$0.714_{(0.019)}$
$\text{GROVER}_{\text{base}}$	$0.936_{(0.008)}$	$0.656_{(0.006)}$	$0.925_{(0.013)}$	$0.878_{(0.016)}$	$0.819_{(0.020)}$	$0.723_{(0.010)}$
$\text{GROVER}_{\text{large}}$	$\mathbf{0.940}_{(0.019)}$	$\mathbf{0.658}_{(0.023)}$	$\mathbf{0.944}_{(0.021)}$	$\mathbf{0.894}_{(0.028)}$	$\mathbf{0.831}_{(0.025)}$	$\mathbf{0.737}_{(0.010)}$
Regression (Lower is better)						
Dataset # Molecules	FreeSolv 642	ESOL 1128	Lipo 4200	QM7 6830	QM8 21786	
TF_Robust [40]	$4.122_{(0.085)}$	$1.722_{(0.038)}$	$0.909_{(0.060)}$	$120.6_{(9.6)}$	$0.024_{(0.001)}$	
GraphConv [24]	$2.900_{(0.135)}$	$1.068_{(0.050)}$	$0.712_{(0.049)}$	$118.9_{(20.2)}$	$0.021_{(0.001)}$	
Weave [23]	$2.398_{(0.250)}$	$1.158_{(0.055)}$	$0.813_{(0.042)}$	$94.7_{(2.7)}$	$0.022_{(0.001)}$	
SchNet [45]	$3.215_{(0.755)}$	$1.045_{(0.064)}$	$0.909_{(0.098)}$	$74.2_{(6.0)}$	$0.020_{(0.002)}$	
MPNN [13]	$2.185_{(0.952)}$	$1.167_{(0.430)}$	$0.672_{(0.051)}$	$113.0_{(17.2)}$	$0.015_{(0.002)}$	
DMPNN [63]	$2.177_{(0.914)}$	$0.980_{(0.258)}$	$0.653_{(0.046)}$	$105.8_{(13.2)}$	$0.0143_{(0.002)}$	
MGCN [30]	$3.349_{(0.097)}$	$1.266_{(0.147)}$	$1.113_{(0.041)}$	$77.6_{(4.7)}$	$0.022_{(0.002)}$	
AttentiveFP [61]	$2.030_{(0.420)}$	$0.853_{(0.060)}$	$0.650_{(0.030)}$	$126.7_{(4.0)}$	$0.0282_{(0.001)}$	
N-GRAM [29]	$2.512_{(0.190)}$	$1.100_{(0.160)}$	$0.876_{(0.033)}$	$125.6_{(1.5)}$	$0.0320_{(0.003)}$	
$\text{GROVER}_{\text{base}}$	$1.592_{(0.072)}$	$0.888_{(0.116)}$	$0.563_{(0.030)}$	$\mathbf{72.5}_{(5.9)}$	$0.0172_{(0.002)}$	
$\text{GROVER}_{\text{large}}$	$\mathbf{1.544}_{(0.397)}$	$\mathbf{0.831}_{(0.120)}$	$\mathbf{0.560}_{(0.035)}$	$72.6_{(3.8)}$	$\mathbf{0.0125}_{(0.002)}$	

基于2D分子图的分子大模型 MGSSL

MGSSL 是一种生成式预训练方法，以分子的 motif 为单位将分子划分为以 motif 为节点的树，通过深度或广度优先遍历确定节点顺序，按自回归方式生成树，生成目标涵盖节点类型和链接关系。



拆分为motif树　　预测连接关系　　预测节点

Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. NeurIPS, 2021. (Tencent)

基于2D分子图的分子大模型 MGSSL

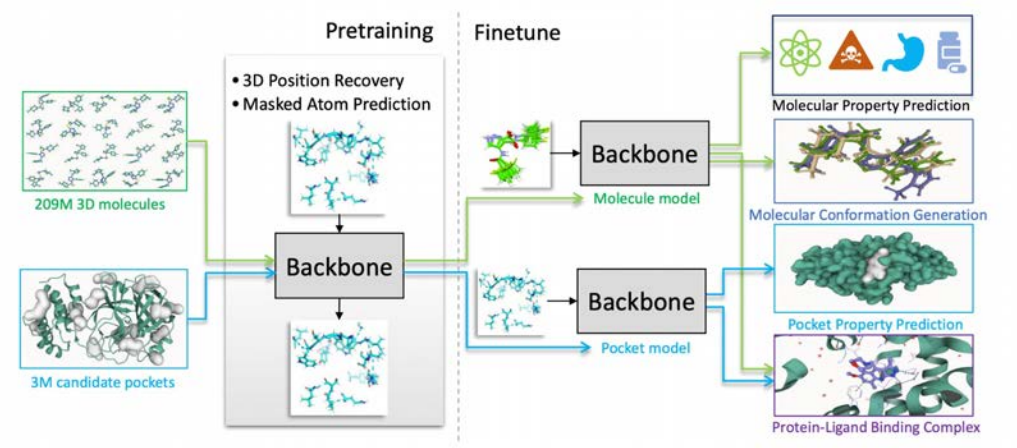
- MGSSL 以 MoleculeNet 作为下游任务，它的两个变种DFS 和BFS（表示不同的遍历节点方式）均取得了 SOTA 的效果

Table 1: Test ROC-AUC (%) performance on molecular property prediction benchmarks using different pre-training strategies with GIN. The rightmost column averages the mean of test performance across the 8 datasets. The best result for each dataset are bolded.

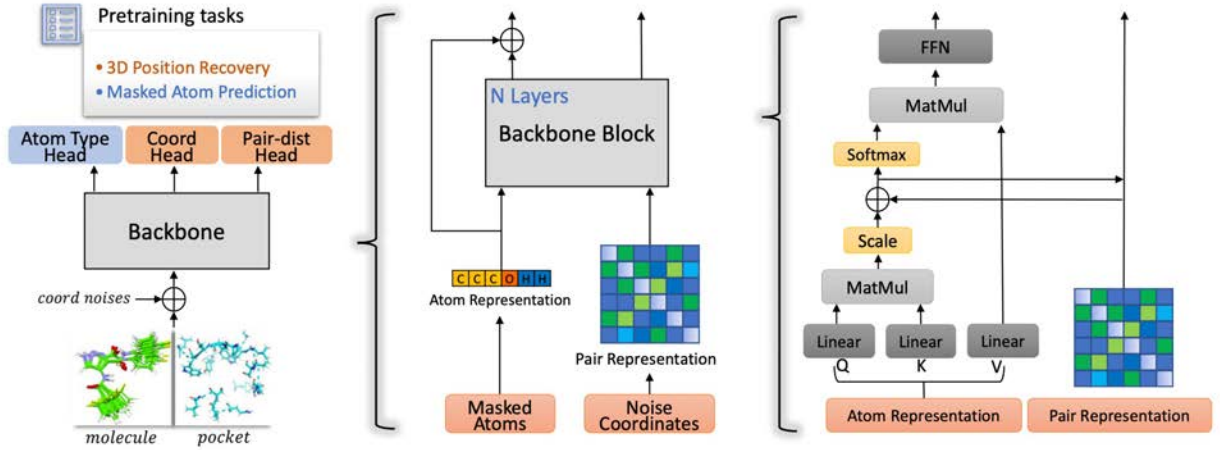
SSL methods	muv	clintox	sider	hiv	tox21	bace	toxcast	bbbp	Avg.
No pretrain	71.7±2.3	58.2±2.8	57.2±0.7	75.4±1.5	74.3±0.5	70.0±2.5	63.3±1.5	65.5±1.8	67.0
Infomax	75.1±2.8	73.0±3.2	58.2±0.5	76.5±1.6	75.2±0.3	75.6±1.0	62.8±0.6	68.1±1.3	70.6
Attribute masking	74.7±1.9	77.5±3.1	59.6±0.7	77.9±1.2	**77.2±0.4**	78.3±1.1	63.3±0.8	65.6±0.9	71.8
GCC	74.1±1.4	73.2±2.6	58.0±0.9	75.5±0.8	76.6±0.5	75.0±1.5	63.5±0.4	66.9±0.7	70.4
GPT-GNN	75.0±2.5	74.9±2.7	59.3±0.8	77.0±1.7	76.1±0.4	78.5±0.9	63.1±0.5	67.5±1.3	71.4
Grover	75.8±1.7	76.9±1.9	60.7±0.5	77.8±1.4	76.3±0.6	79.5±1.1	63.4±0.6	68.0±1.5	72.3
MGSSL (DFS)	78.1±1.8	79.7±2.2	60.5±0.7	**79.5±1.1**	76.4±0.4	**79.7±0.8**	63.8±0.3	**70.5±1.1**	73.5
MGSSL (BFS)	**78.7±1.5**	**80.7±2.1**	**61.8±0.8**	78.8±1.2	76.5±0.3	79.1±0.9	**64.1±0.7**	69.7±0.9	**73.7**

基于3D结构的分子大模型 Uni-Mol

- Uni-Mol以分子3D结构坐标作为输入，对蛋白口袋和分子使用相同的表示学习策略
- Denoising和Masked Atom Prediction作为预训练任务
- 为编码3D结构设计Transformer结构，将pairwise之间的距离关系编码为Pair Representation



对分子和口袋使用相同的训练策略



3D结构Transformer

Uni-Mol: A Universal 3D Molecular Representation Learning Framework. ICLR, 2023. (DP Technology)

基于3D结构的分子大模型 Uni-Mol

- 下游任务：分子性质预测，分子构象生成，口袋性质预测，蛋白-分子docking结构预测

分子性质预测

Classification (ROC-AUC %, higher is better ↑)									
Datasets	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	PCBA	MUV
# Molecules	2039	1513	1478	7831	8575	1427	41127	437929	93087
# Tasks	1	1	2	12	617	27	1	128	17
D-MPNN	71.0(0.3)	80.9(0.6)	90.6(0.6)	75.9(0.7)	65.5(0.3)	57.0(0.7)	77.1(0.5)	86.2(0.1)	78.6(1.4)
Attentive FP	64.3(1.8)	78.4(0.022)	84.7(0.3)	76.1(0.5)	63.7(0.2)	60.6(3.2)	75.7(1.4)	80.1(1.4)	76.6(1.5)
N-Gram$_{RF}$	69.7(0.6)	77.9(1.5)	77.5(4.0)	74.3(0.4)	-	66.8(0.7)	77.2(0.1)	-	76.9(0.7)
N-Gram$_{XGB}$	69.1(0.8)	79.1(1.3)	87.5(2.7)	75.8(0.9)	-	65.5(0.7)	78.7(0.4)	-	74.8(0.2)
PretrainGNN	68.7(1.3)	84.5(0.7)	72.6(1.5)	78.1(0.6)	65.7(0.6)	62.7(0.8)	79.9(0.7)	86.0(0.1)	81.3(2.1)
GROVER$_{base}$	70.0(0.1)	82.6(0.7)	81.2(3.0)	74.3(0.1)	65.4(0.4)	64.8(0.6)	62.5(0.9)	76.5(2.1)	67.3(1.8)
GROVER$_{large}$	69.5(0.1)	81.0(1.4)	76.2(3.7)	73.5(0.1)	65.3(0.5)	65.4(0.1)	68.2(1.1)	83.0(0.4)	67.3(1.8)
GraphMVP	72.4(1.6)	81.2(0.9)	79.1(2.8)	75.9(0.5)	63.1(0.4)	63.9(1.2)	77.0(1.2)	-	77.7(0.6)
MolCLR	72.2(2.1)	82.4(0.9)	91.2(3.5)	75.0(0.2)	-	58.9(1.4)	78.1(0.5)	-	79.6(1.9)
GEM	72.4(0.4)	85.6(1.1)	90.1(1.3)	78.1(0.1)	69.2(0.4)	**67.2(0.4)**	80.6(0.9)	86.6(0.1)	81.7(0.5)
Uni-Mol	**72.9(0.6)**	**85.7(0.2)**	**91.9(1.8)**	**79.6(0.5)**	**69.6(0.1)**	65.9(1.3)	**80.8(0.3)**	**88.5(0.1)**	**82.1(1.3)**

优于之前的2D方法

蛋白-分子docking结构预测

Ligand RMSD % Below Threshold ↑					
Methods	1.0 Å	1.5 Å	2.0 Å	3.0 Å	5.0 Å
Autodock Vina	44.21	57.54	64.56	73.68	84.56
Vinardo	41.75	57.54	62.81	69.82	76.84
Smina	**47.37**	59.65	65.26	74.39	82.11
Autodock4	21.75	31.58	35.44	47.02	64.56
Uni-Mol$_{no_pretrained}$	39.65	63.16	72.98	83.51	91.58
Uni-Mol	43.16	**68.42**	**80.35**	**87.02**	**94.04**

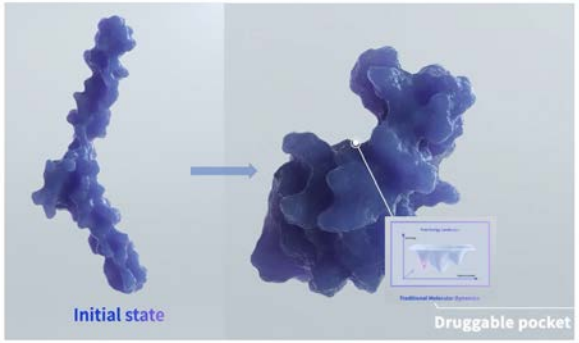
口袋性质预测

Dataset	Classification (higher is better ↑) NRDLD						Regression (lower is better ↓) Our Created		
Methods	Cavity-DrugScore	Volsite	DrugPred	PockDrug	TRAPP-CNN	**Uni-Mol**	Methods	Uni-Mol$_{no_pretrained}$	**Uni-Mol**
Accuracy	0.82	0.89	0.89	0.865	0.946	**0.973**	RMSE$_{Fpocket}$	0.1155(0.002)	**0.1140(0.001)**
Recall	-	-	-	0.957	0.913	**1.000**	RMSE$_{Druggability}$	0.1117(0.002)	**0.1001(0.001)**
Precision	-	-	-	0.846	**1.000**	0.958	RMSE$_{Total\ SASA}$	22.010(0.460)	**20.734(0.015)**
F1-score	-	-	-	0.898	0.955	**0.979**	RMSE$_{Hydrophobicity}$	1.4144(0.034)	**1.2847 (0.005)**

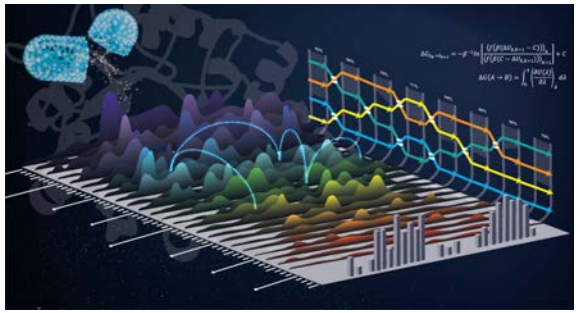
物理感知的分子大模型 Frad

- 物理规律在AI4Science领域中无处不在，影响和决定分子的各类性质
- 而目前常用的从NLP或CV借鉴的自监督学习方法无法和分子底层物理规律产生联系，不具有物理可解释性

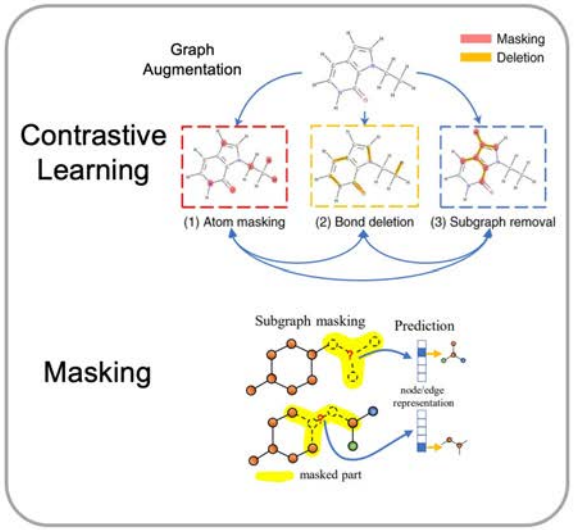
MD辅助靶点发现



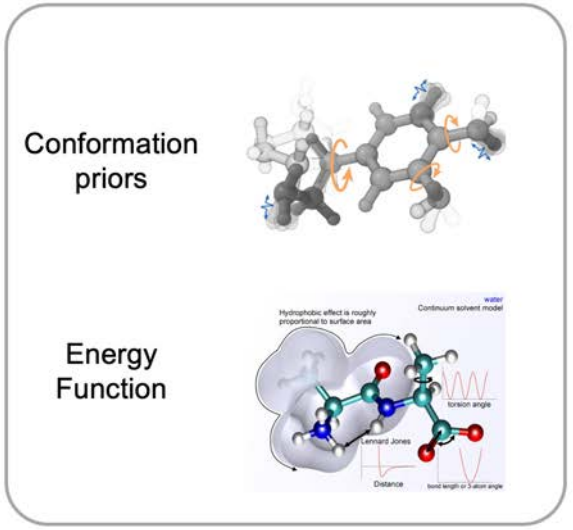
基于物理的方法FEP测定Binding Affinity



物理规律的重要性



≠

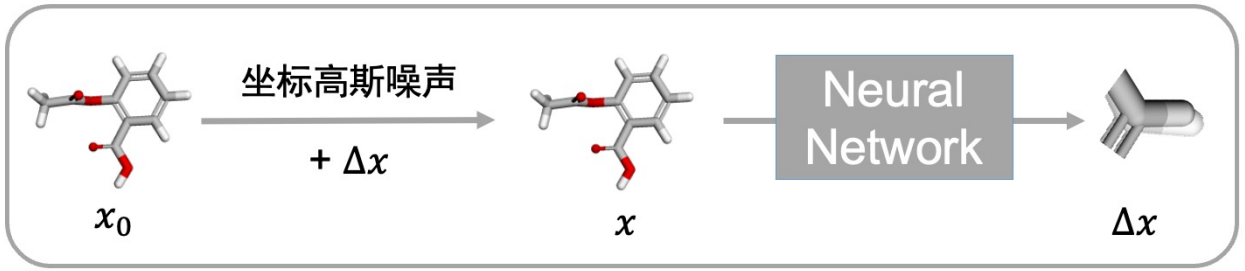


目前常用的对比学习/Masking策略缺乏物理可解释性

Fractional Denoising for 3D Molecular Pre-training. ICML, 2024. (Tsinghua AIR)

物理感知的分子大模型 Frad

- Denoising策略在玻尔兹曼和高斯分布的假设下，具有近似学习力场的物理可解释性
- 简单的高斯假设会导致**低采样覆盖率**和学习**各向同性**的力场限制



坐标去噪策略

假设条件

$$\nabla \log p(x|x_0) = -\frac{x - x_0}{\tau_c^2} \longleftarrow p(\boldsymbol{x}|\boldsymbol{x}_0) \sim \mathcal{N}(\boldsymbol{x}_0, \tau_c^2 I_{3N})$$

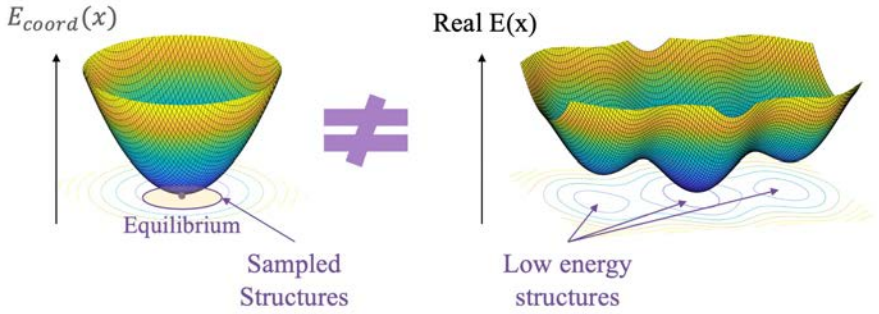
$$\nabla \log p(x) = -\nabla E_{coord}(x) \longleftarrow p(\boldsymbol{x}) \propto exp(-E_{Coord}(\boldsymbol{x}))$$

等价关系

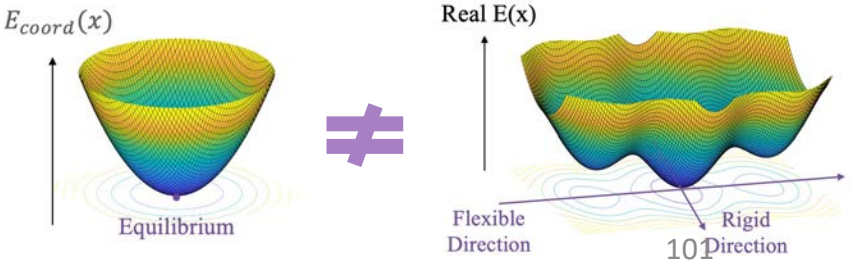
$$\mathcal{L}_{Coord}(\mathcal{M}) = E_{p(\boldsymbol{x}|\boldsymbol{x}_0)p(\boldsymbol{x}_0)}||GNN_\theta(\boldsymbol{x}) - (\boldsymbol{x} - \boldsymbol{x}_0)||^2$$

$$\simeq E_{p(\boldsymbol{x})}||GNN_\theta(\boldsymbol{x}) - (-\nabla_{\boldsymbol{x}} E_{Coord}(\boldsymbol{x}))||^2$$

低采样覆盖率

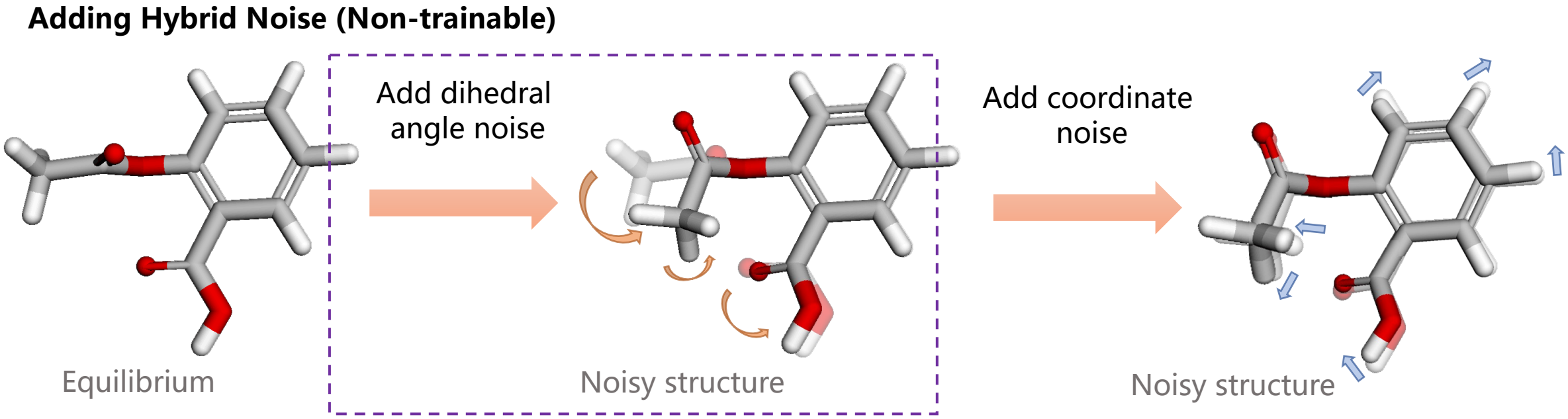


各向同性力场



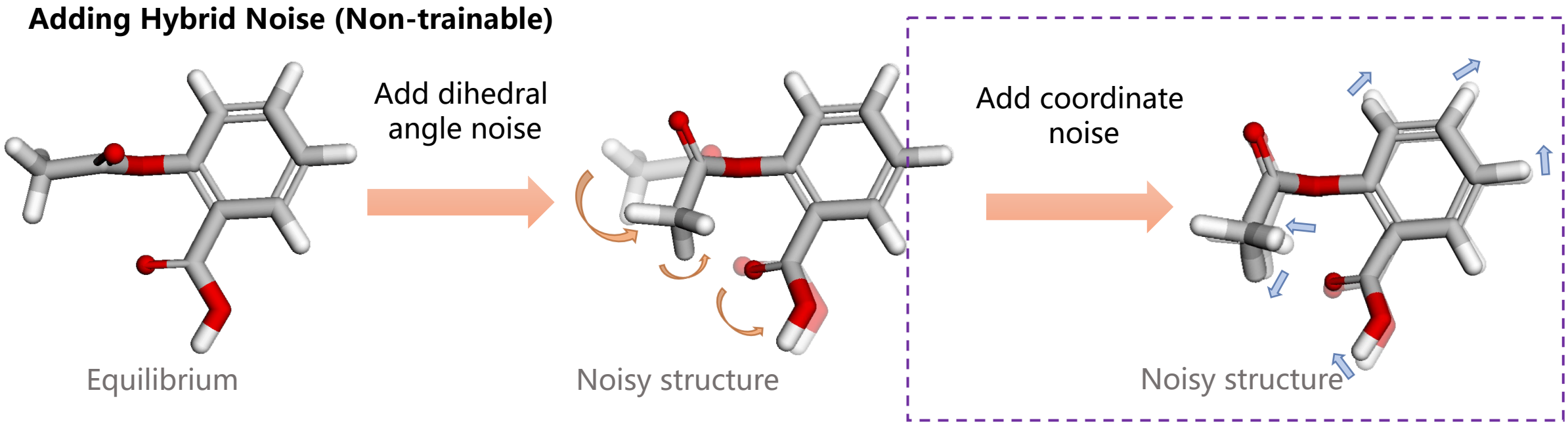
物理感知的分子大模型Frad

- 扩大采样范围：设计组合噪声，在坐标噪声之前加入可旋转二面角噪声，同时建模分子柔性和刚性部分，其中二面角噪声可以赋予较大方差，以扩大采样范围



物理感知的分子大模型Frad

- 学习各向异性分子力场：只去噪坐标部分，可证明和学习分子各向异性力场目标等价。



物理感知的分子大模型Frad

- QM9 是一个量子化学数据集，包含稳态分子的几何、能量、电子和热力学性质。分数阶去噪Frad在9/12个子任务上取得了最优结果
- Frad 超过了其它的去噪预训练方法，表明了去噪过程引入化学先验分布的优势

Models	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$< R^2 >$ (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{molK}$)
SchNet	0.033	0.235	41.0	34.0	63.0	0.07	1.70	14.00	19.00	14.00	14.00	0.033
E(n)-GNN	0.029	0.071	29.0	25.0	48.0	0.11	1.55	11.00	12.00	12.00	12.00	0.031
DimeNet++	0.030	0.043	24.6	19.5	32.6	0.33	1.21	6.32	6.28	6.53	7.56	0.023
PaiNN	0.012	0.045	27.6	20.4	45.7	0.07	1.28	5.85	5.83	5.98	7.35	0.024
SphereNet	0.027	0.047	23.6	18.9	32.3	0.29	**1.120**	6.26	7.33	6.40	8.00	0.022
TorchMD-NET	0.011	0.059	20.3	18.6	36.1	**0.033**	1.840	6.15	6.38	6.16	7.62	0.026
Transformer-M	0.037	0.041	17.5	16.2	**27.4**	0.075	1.18	9.37	9.41	9.39	9.63	0.022
SE(3)-DDM	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026
DP-TorchMD -NET($\tau = 0.04$)	0.012	0.0517	17.7	14.3	31.8	0.4496	1.71	6.57	6.11	6.45	6.91	**0.020**
Frad ($\sigma = 2, \tau = 0.04$)	**0.010**	**0.0374**	**15.3**	**13.7**	27.8	0.3419	1.418	**5.33**	**5.62**	**5.55**	**6.19**	**0.020**

等变网络

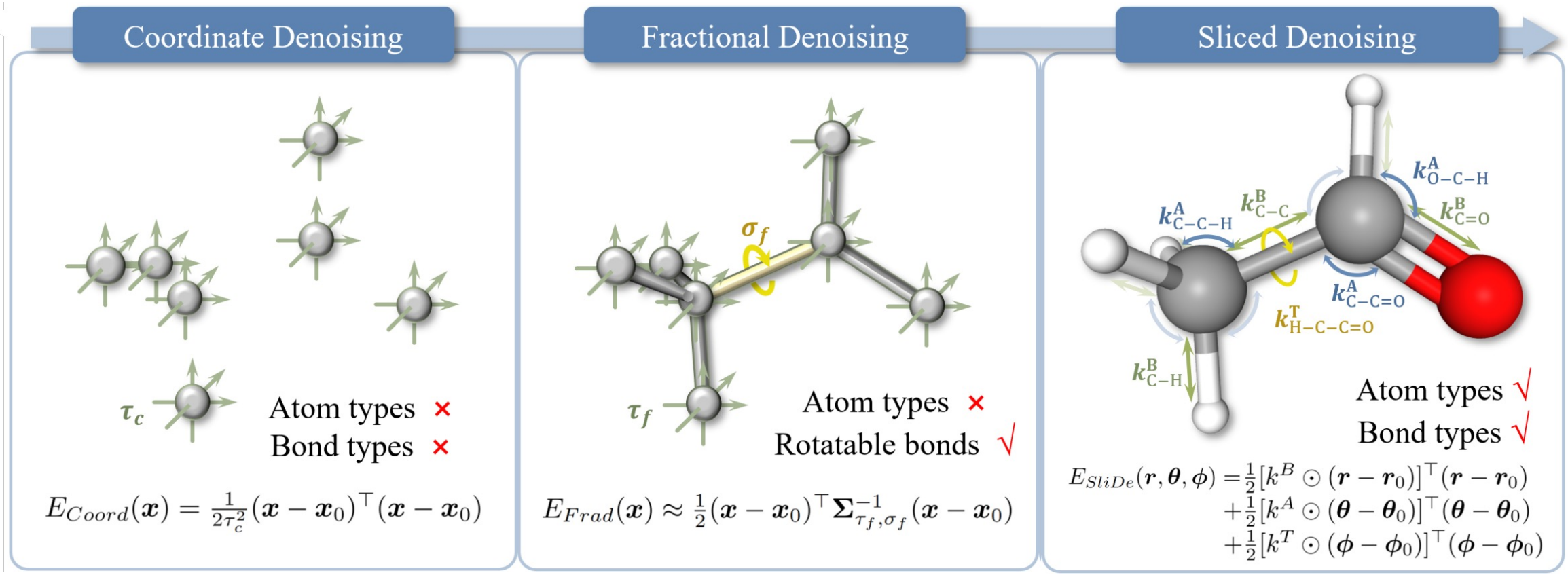
其它去噪预训练方法

物理感知的分子大模型Frad

- MD17 数据集包含了8个有机小分子的分子动力学数据，预测任务是给定分子结构，预测原子受力，与预训练目标一致
- Frad 在两种不同的split设置中， 8 个分子中的 7 个上优于相应的预训练和非预训练基线模型

Training set size	Models	Aspirin	Benzene	Ethanol	Malonaldehyde	Naphthalene	Salicylic Acid	Toluene	Uracil
9500	TorchMD-NET	0.1216	0.1479	0.0492	0.0695	0.0390	0.0655	0.0393	0.0484
	3D-EMGP	0.1560	0.1648	0.0389	0.0737	0.0829	0.1187	0.0619	0.0773
	3D-EMGP (TorchMD-NET)	0.1124	0.1417	0.0445	0.0618	0.0352	0.0586	0.0385	0.0477
	DP-TorchMD -NET(τ = 0.04)	0.0920	**0.1397**	0.0402	0.0661	0.0544	0.0790	0.0495	0.0507
	Frad (σ = 2, τ = 0.04)	**0.0680**	0.1606	**0.0332**	**0.0427**	**0.0277**	**0.0410**	**0.0305**	**0.0323**
1000	SphereNet	0.430	**0.178**	0.208	0.340	0.178	0.360	0.155	0.267
	SchNet	1.35	0.31	0.39	0.66	0.58	0.85	0.57	0.56
	DimeNet	0.499	0.187	0.230	0.383	0.215	0.374	0.216	0.301
	SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
950	PaiNN*	0.338	-	0.224	0.319	0.077	0.195	0.094	0.139
	TorchMD-NET	0.2450	0.2187	0.1067	0.1667	0.0593	0.1284	0.0644	0.0887
	Frad (σ = 2, τ = 0.04)	**0.2087**	0.1994	**0.0910**	**0.1415**	**0.0530**	**0.1081**	**0.0540**	**0.0760**

物理感知的分子大模型SliDe



从为去噪预训练寻找物理解释（Coord, Frad）到为学习精确分子力场设计自监督任务（SliDe）

Sliced Denoising: A Physics-Informed Molecular Pre-Training Method. ICLR, 2024. (Tsinghua AIR)

物理感知的分子大模型SliDe

SliDe从能量函数出发推导自监督损失的形式

A three-step journey to Sliced Denoising (SliDe) method

1. Energy Function

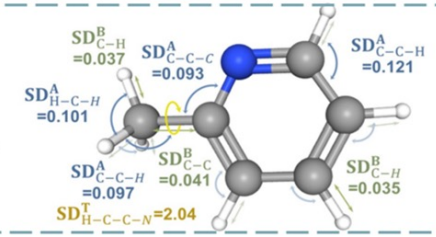
$$E_{BAT}(\boldsymbol{r},\boldsymbol{\theta},\boldsymbol{\phi}) = \frac{1}{2}\sum_{i\in\mathbb{B}} k_i^B (r_i - r_{i,0})^2 + \frac{1}{2}\sum_{i\in\mathbb{A}} k_i^A (\theta_i - \theta_{i,0})^2 + \frac{1}{2}\sum_{i\in\mathbb{T}} k_i^T \omega_i^2 (\phi_i - \phi_{i,0})^2.$$

Boltzmann Distribution

2. Noise Design

$$\boldsymbol{r} \sim \mathcal{N}(\boldsymbol{r}_0, diag(\frac{1}{\boldsymbol{k}^B})),$$

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, diag(\frac{1}{\boldsymbol{k}^A})),\ \boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\phi}_0, diag(\frac{1}{\boldsymbol{k}^T \odot \boldsymbol{\omega}^2}))$$



3. Loss Derivation

Force Learning

$$E_{p(\boldsymbol{x}|\boldsymbol{x}_0)}||GNN_\theta(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} E_{BAT}(\boldsymbol{d}(\boldsymbol{x}))||^2$$

Cartesian coordinates　Relative coordinates $\boldsymbol{d} = (\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{\phi})$

Variable Change

Efficient Jacobian Estimation

$$E_{p(\boldsymbol{x}|\boldsymbol{x}_0)}||GNN_\theta(\boldsymbol{x}) - \nabla_{\boldsymbol{d}} E_{BAT}(\boldsymbol{d})^\top \cdot J(\boldsymbol{x})||^2$$

Random Slicing $\boldsymbol{v}_i \sim \mathcal{N}(\boldsymbol{0}, I_{3N})$

$$E_{p(\boldsymbol{x}|\boldsymbol{x}_0)} \frac{1}{N_v}\sum_{i=1}^{N_v}\left[GNN_\theta(\boldsymbol{x})^\top \cdot \boldsymbol{v}_i - \frac{1}{\sigma}\nabla_{\boldsymbol{d}} E_{BAT}(\boldsymbol{d})^\top \cdot J(\boldsymbol{x}) \cdot \boldsymbol{v}_i\right]^2$$

SliDe Loss

$$E_{p(\boldsymbol{x}|\boldsymbol{x}_0)} \frac{1}{N_v}\sum_{i=1}^{N_v}\left[GNN_\theta(\boldsymbol{x})^\top \cdot \boldsymbol{v}_i - \frac{1}{\sigma}\nabla_{\boldsymbol{d}} E_{BAT}(\boldsymbol{d})^\top \cdot \left(f^{\mathcal{M}}(\boldsymbol{x}+\sigma\boldsymbol{v}_i) - f^{\mathcal{M}}(\boldsymbol{x})\right)\right]^2$$

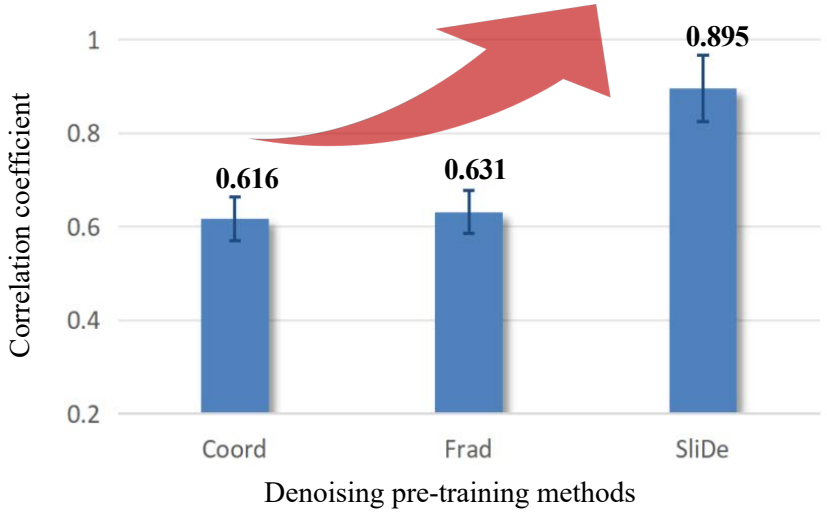
$O(n^2)$复杂度

↓

O(n)复杂度

物理感知的分子大模型SliDe

预训练目标的物理一致性



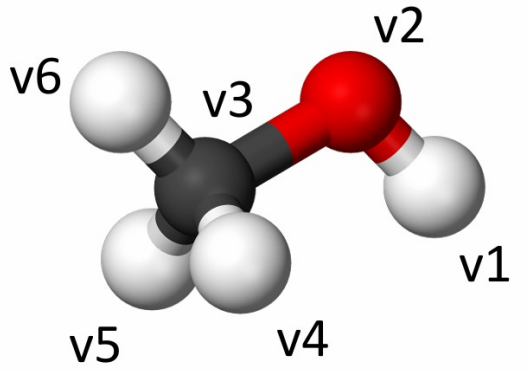
QM9	μ (D)	α (a_0^3)	homo (meV)	lumo (meV)	gap (meV)	R^2 (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{mol \cdot K}$)
SchNet	0.033	0.235	41.0	34.0	63.0	0.07	1.70	14.00	19.00	14.00	14.00	0.033
E(n)-GNN	0.029	0.071	29.0	25.0	48.0	0.11	1.55	11.00	12.00	12.00	12.00	0.031
DimeNet++	0.030	0.044	24.6	19.5	32.6	0.33	1.21	6.32	6.28	6.53	7.56	0.023
PaiNN	0.012	0.045	27.6	20.4	45.7	0.07	1.28	5.85	5.83	5.98	7.35	0.024
SphereNet	0.025	0.045	22.8	18.9	31.1	0.27	**1.120**	6.26	6.36	6.33	7.78	0.022
ET	0.011	0.059	20.3	17.5	36.1	**0.033**	1.840	6.15	6.38	6.16	7.62	0.026
TM	0.037	0.041	17.5	16.2	27.4	0.075	1.18	9.37	9.41	9.39	9.63	0.022
SE(3)-DDM	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026
Coord	0.012	0.0517	17.7	14.3	31.8	0.4496	1.71	6.57	6.11	6.45	6.91	0.020
Frad	0.010	0.0374	15.3	13.7	27.8	0.3419	1.418	5.33	5.62	5.55	6.19	0.020
SliDe	**0.0087**	**0.0366**	**13.6**	**12.3**	**26.2**	0.3405	1.521	**4.28**	**4.29**	**4.26**	**5.37**	**0.019**

MD17	Aspirin	Benzene	Ethanol	Malonal -dehyde	Naphtha -lene	Salicy -lic Acid	Toluene	Uracil
SphereNet	0.430	0.178	0.208	0.340	0.178	0.360	0.155	0.267
SchNet	1.35	0.31	0.39	0.66	0.58	0.85	0.57	0.56
DimeNet	0.499	0.187	0.230	0.383	0.215	0.374	0.216	0.301
PaiNN*	0.338	-	0.224	0.319	0.077	0.195	0.094	0.139
ET	0.2450	0.2187	0.1067	0.1667	0.0593	0.1284	0.0644	0.0887
SE(3)-DDM*	0.453	-	0.166	0.288	0.129	0.266	0.122	0.183
Coord	0.2108	0.1692	0.0959	**0.1392**	0.0529	0.1087	0.0582	**0.0742**
Frad	0.2087	0.1994	0.0910	0.1415	0.0530	0.1081	**0.0540**	0.0760
SliDe	**0.1740**	**0.1691**	**0.0882**	0.1538	**0.0483**	**0.1006**	**0.0540**	0.0825

	Noneq	SliDe
w/o pre-train	1.50	1.362
pre-train	1.01	0.786
pre-train improvement	32.7%	**42.3%**

多模态的分子大模型MoleBlend

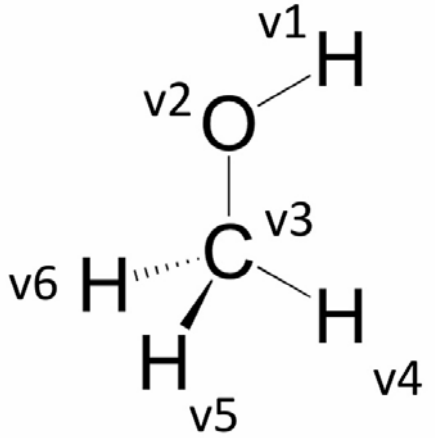
3D结构

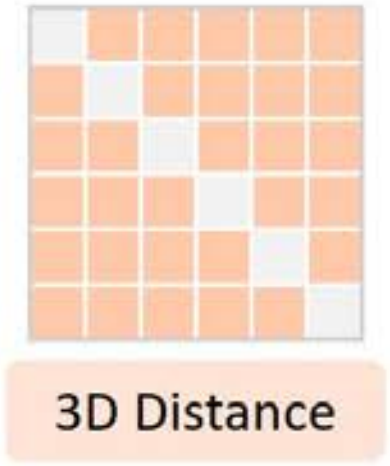


2D/3D具有相同节点表示的图结构

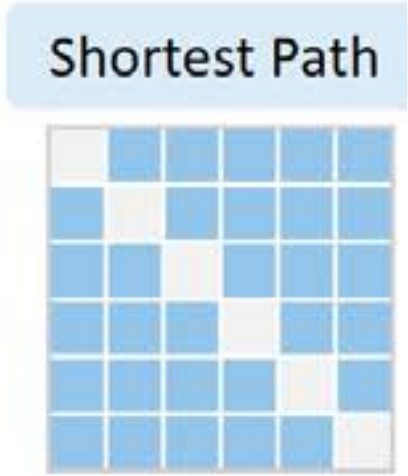
原子间关系表示不同

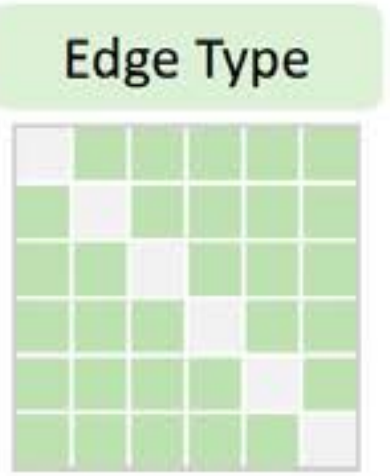
2D分子图





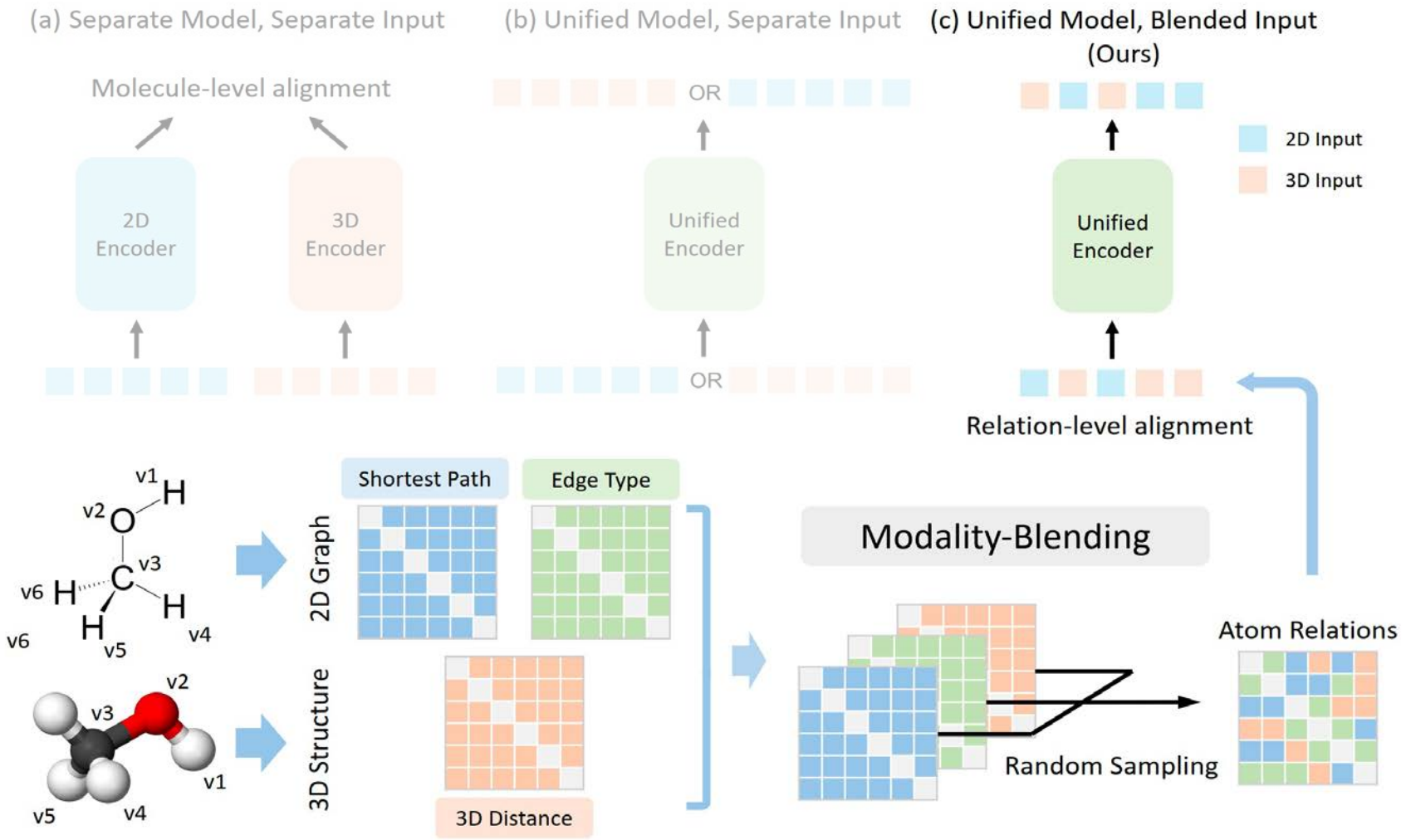
能否使用**细粒度对齐**来同时增强二维和三维表示?





Multimodal Molecular Pretraining via Modality Blending. ICLR, 2024. (Tsinghua AIR)

多模态的分子大模型MoleBlend

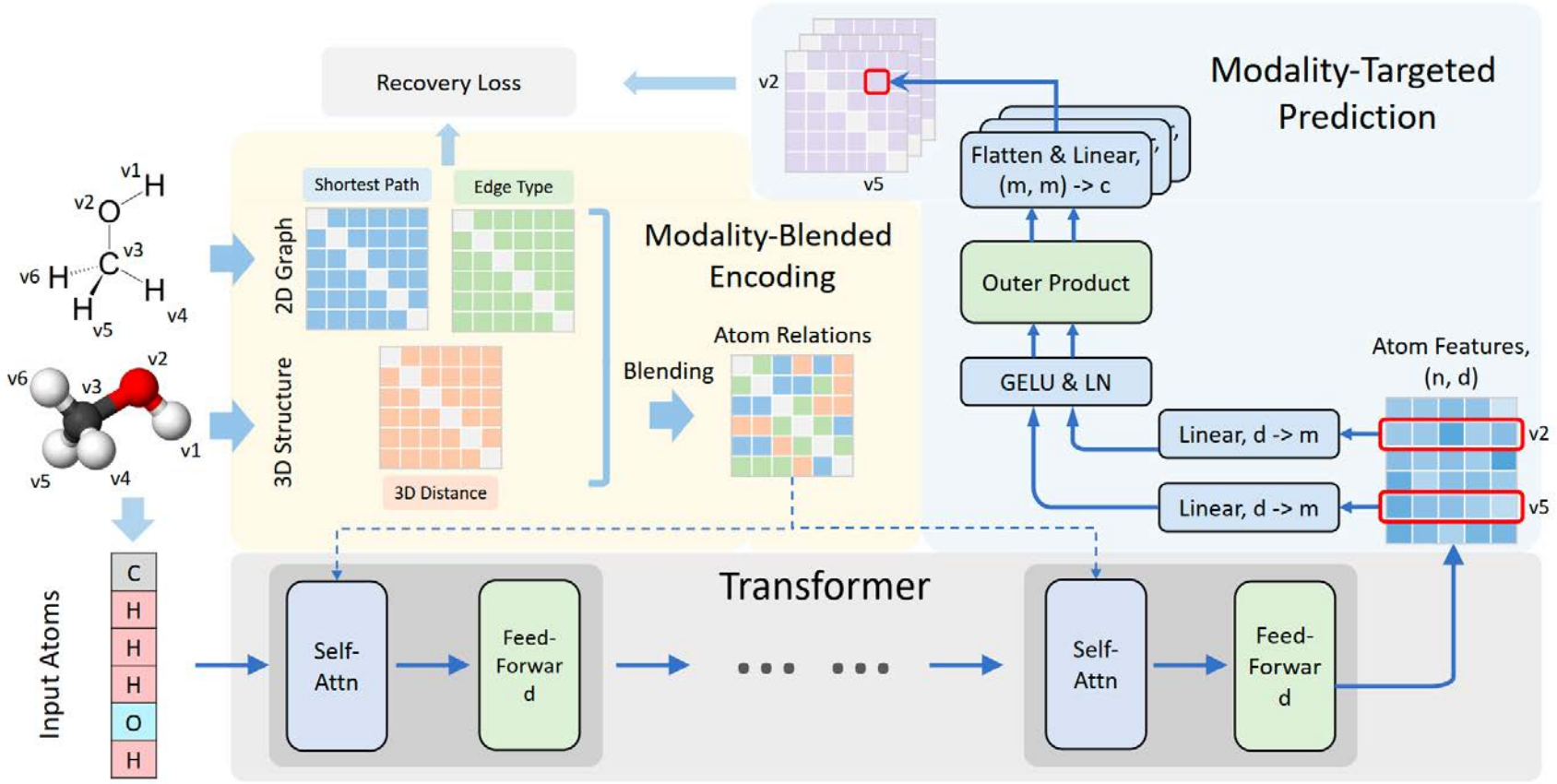


将2D/3D多模态输入**混合（Blend）**为一个统一的数据结构，使用单塔模型来建模

Multimodal Molecular Pretraining via Modality Blending. ICLR, 2024. (Tsinghua AIR)

多模态的分子大模型MoleBlend

- 融合分子2D和3D模态得到分子的统一表示
- 提出Modality-Blended Encoding融合分子2D和3D的信息，设计Modality-Targeted Prediction进行跨模态信息重建



Multimodal Molecular Pretraining via Modality Blending. *ICLR, 2024. (Tsinghua AIR)*

多模态的分子大模型MoleBlend

MoleculeNet (Only 2D)

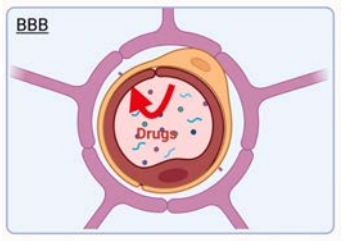
Pre-training Methods	BBBP ↑	Tox21 ↑	ToxCast ↑	SIDER ↑	ClinTox ↑	MUV ↑	HIV ↑	Bace ↑	Avg ↑
AttrMask (Hu et al., 2020)	65.0±2.3	74.8±0.2	62.9±0.1	61.2±0.1	87.7±1.1	73.4±2.0	76.8±0.5	79.7±0.3	72.68
ContextPred (Hu et al., 2020)	65.7±0.6	74.2±0.0	62.5±0.3	62.2±0.5	77.2±0.8	75.3±1.5	77.1±0.8	76.0±2.0	71.28
GraphCL (You et al., 2020)	69.7±0.6	73.9±0.6	62.4±0.5	60.5±0.8	76.0±2.6	69.8±2.6	78.5±1.2	75.4±1.4	70.78
InfoGraph (Sun et al., 2020)	67.5±0.1	73.2±0.4	63.7±0.5	59.9±0.3	76.5±1.0	74.1±0.7	75.1±0.9	77.8±0.8	70.98
GROVER (Rong et al., 2020)	70.0±0.10	74.3±0.1	65.4±0.4	64.8±0.6	81.2±3.0	67.3±1.8	62.5±0.9	82.6±0.7	71.01
MolCLR (Wang et al., 2022b)	66.6±1.8	73.0±0.1	62.9±0.3	57.5±1.7	86.1±0.9	72.5±2.3	76.2±1.5	71.5±3.1	70.79
GraphMAE (Hou et al., 2022)	72.0±0.6	75.5±0.6	64.1±0.3	60.3±1.1	82.3±1.2	76.3±2.4	77.2±1.0	83.1±0.9	73.85
Mole-BERT (Xia et al., 2023)	71.9±1.6	76.8±0.5	64.3±0.2	62.8±1.1	78.9±3.0	78.6±1.8	78.2±0.8	80.8±1.4	74.04
3D InfoMax (Stärk et al., 2022)	69.1±1.0	74.5±0.7	64.4±0.8	60.6±0.7	79.9±3.4	74.4±2.4	76.1±1.3	79.7±1.5	72.34
GraphMVP (Liu et al., 2022b)	68.5±0.2	74.5±0.4	62.7±0.1	62.3±1.6	79.0±2.5	75.0±1.4	74.8±1.4	76.8±1.1	71.69
MoleculeSDE (Liu et al., 2023)	71.8±0.7	76.8±0.3	65.0±0.2	60.8±0.3	87.0±0.5	**80.9±0.3**	78.8±0.9	79.5±2.1	75.07
Transformer from scratch	69.4±1.1	74.2±0.3	62.6±0.3	**65.8±0.3**	**90.3±0.9**	71.3±0.8	76.2±0.6	79.5±0.2	73.66
3D InfoMax (Transformer impl.)	70.4±1.0	75.5±0.5	63.1±0.7	64.1±0.1	89.8±1.2	72.8±1.0	74.9±0.3	80.7±0.6	73.91
GraphMVP (Transformer impl.)	71.5±1.3	76.1±0.9	64.3±0.6	64.7±0.7	89.9±0.9	74.9±1.2	76.0±0.6	81.5±1.2	74.86
MOLEBLEND	**73.0±0.8**	**77.8±0.8**	**66.1±0.0**	64.9±0.3	87.6±0.7	77.2±2.3	**79.0±0.8**	**83.7±1.4**	**76.16**

QM9 (3D + 2D)

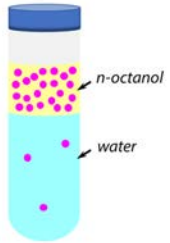
Pre-training Methods	Alpha ↓	Gap ↓	HOMO ↓	LUMO ↓	Mu ↓	Cv ↓	G298 ↓	H298 ↓	R2 ↓	U298 ↓	U0 ↓	Zpve ↓
Distance Prediction (Liu et al., 2022a)	0.065	45.87	27.61	23.34	0.031	0.033	14.83	15.81	0.248	15.07	15.01	1.837
3D InfoGraph (Sun et al., 2020)	0.062	45.96	29.29	24.60	0.028	0.030	13.93	13.97	**0.133**	13.55	13.47	1.644
3D InfoMax (Stärk et al., 2022)	0.057	42.09	25.90	21.60	0.028	0.030	13.73	13.62	0.141	13.81	13.30	1.670
GraphMVP (Liu et al., 2022b)	0.056	41.99	25.75	21.58	0.027	0.029	13.43	13.31	0.136	13.03	13.07	1.609
MoleculeSDE (Liu et al., 2023)	**0.054**	41.77	25.74	21.41	**0.026**	**0.028**	13.07	12.05	0.151	12.54	12.04	1.587
MOLEBLEND	0.060	**34.75**	**21.47**	**19.23**	0.037	0.031	**12.44**	**11.97**	0.417	**12.02**	**11.82**	**1.580**

MoleBlend 在 MoleculeNet（2D 模态）和 QM9（2D/3D 模态）任务中均达到了 SOTA 表现

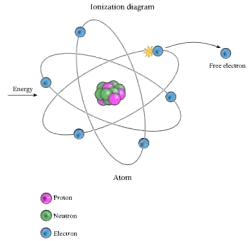
多任务统一的分子大模型UniCorn



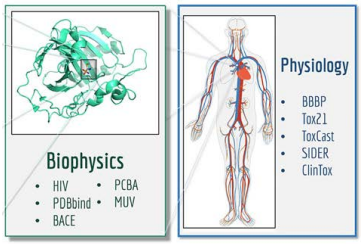
血脑屏障通透性



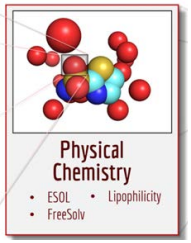
脂溶性



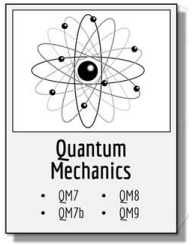
电离能



生理性质



物理化学性质



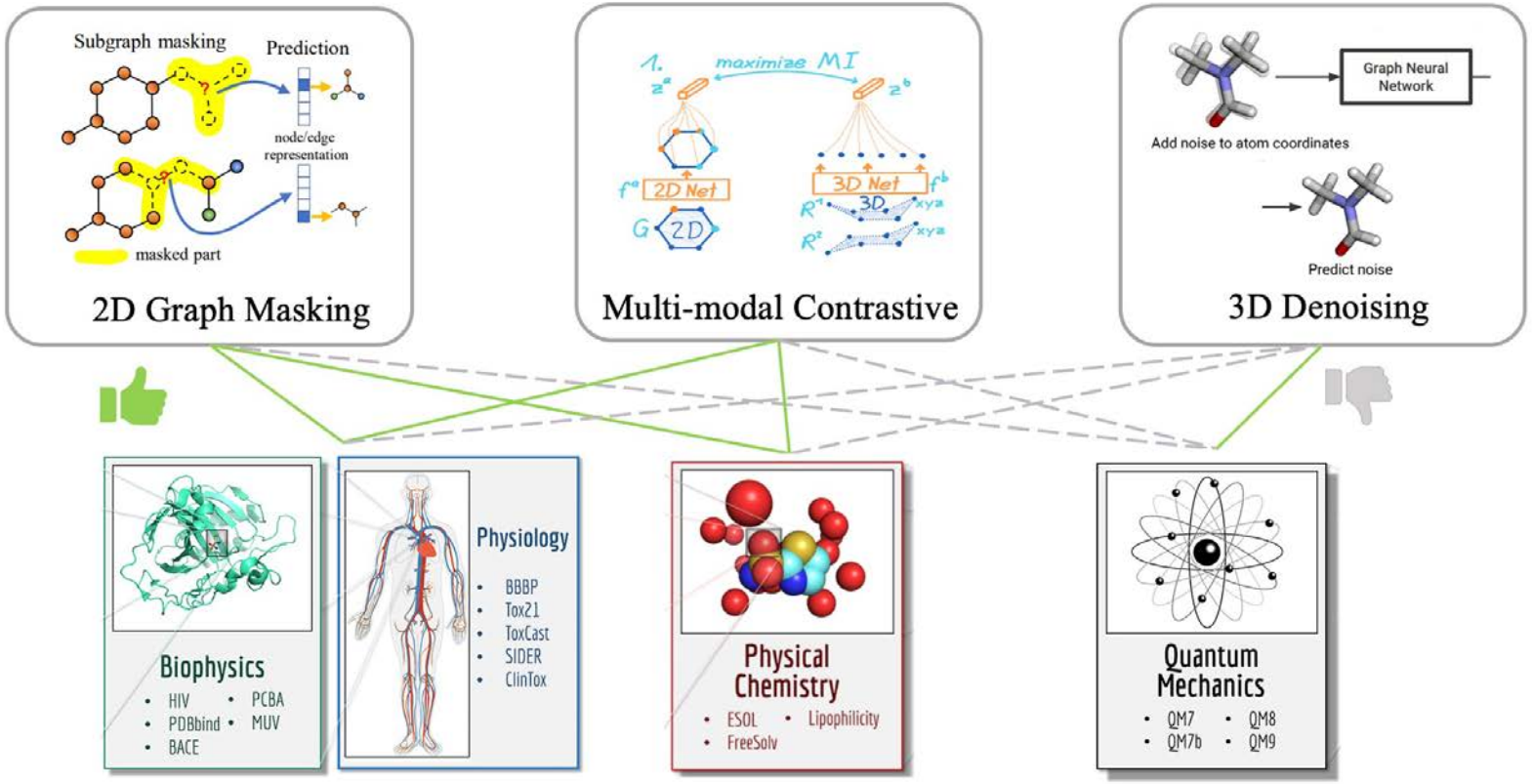
量子力学性质

宏观

微观

分子性质具有层次性，微观和宏观性质对应分子的不同的尺度，微观性质影响和决定宏观性质

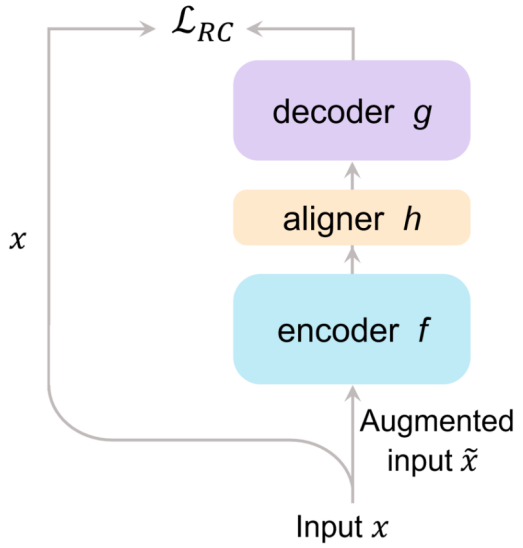
多任务统一的分子大模型UniCorn



不同层次的任务虽然具有很强的相关性，但是现有预训练方法对特定任务存在**偏好性**，目前缺乏有效的统一多任务的表示学习方法

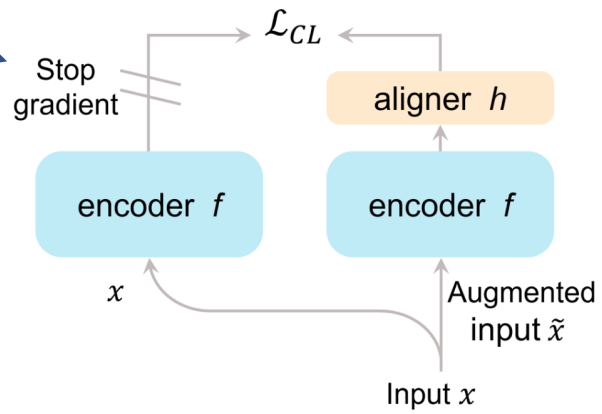
多任务统一的分子大模型UniCorn

- 将Masking和Denoising总结为重建方法，对重建方法和对比学习的目标函数进行分析，证明在一定的正则条件下，两种优化的loss具有优化等价性
- 三种SSL方法都可归纳为对比学习方法，本质区别在于样本增强方式的不同



$$\mathcal{L}_{CL} + \mathcal{L}_{reg} \gtrsim \mathcal{L}_{RC} \gtrsim \mathcal{L}_{CL} - \mathcal{L}_{reg}$$

$$\mathcal{L}_{reg} = \mathbb{E}_{p(x)} ||g_\phi(\mathcal{SG}(f_\theta(x))) - x||_2$$



重建方法(*Masking, Denoising*)

$$\mathcal{L}_{\mathrm{RC}} = \mathbb{E}_{p(x)} \mathbb{E}_{p(\tilde{x}|x)} ||g_\phi(h_\psi(f_\theta(\tilde{x}))) - x||_2,$$

对比学习(*2D-3D Contrastive loss*)

$$\mathcal{L}_{\mathrm{CL}} = \mathbb{E}_{p(x)} \mathbb{E}_{p(\tilde{x}|x)} ||h_\psi(f_\theta(\tilde{x})) - \mathcal{SG}(f_\theta(x))||_2,$$

UniCorn: A Unified Contrastive Learning Approach for Multi-view Molecular Representation Learning . ICML, 2024. (Tsinghua AIR)

多任务统一的分子大模型UniCorn

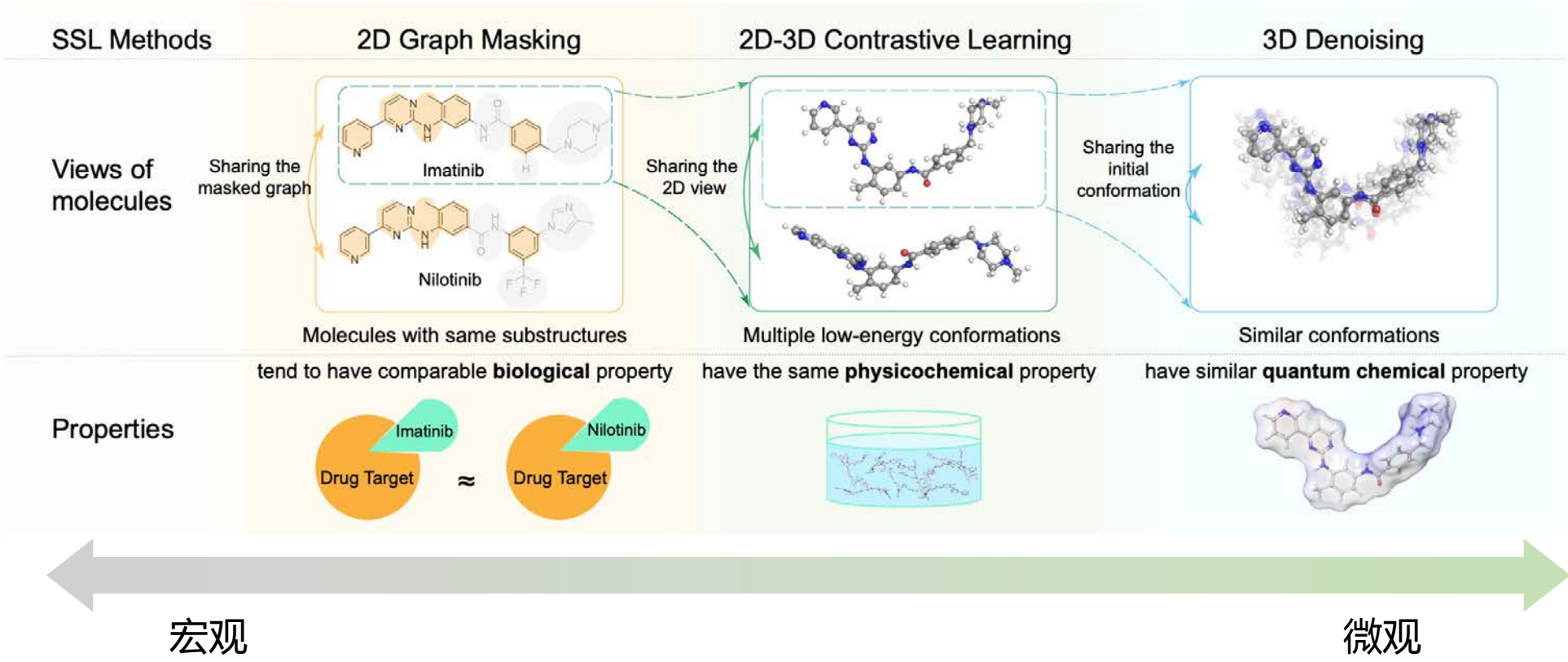
不同的自监督任务

作用

分子不同层次的聚类效果

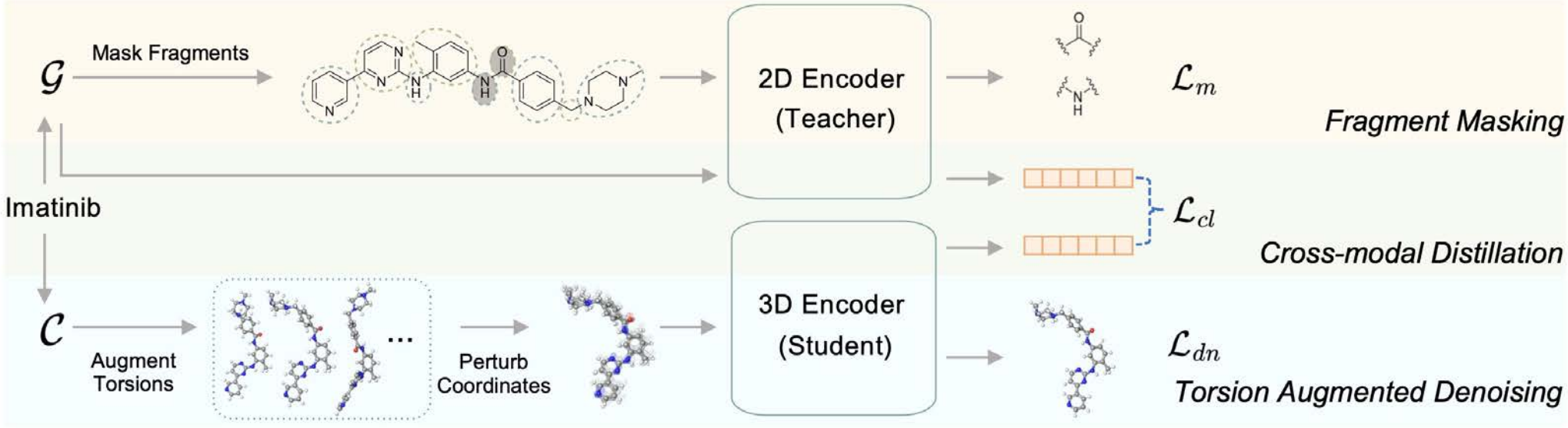
结合归纳偏置

偏好与特定的性质预测任务



统一的对比学习分析一方面提供了对于自监督任务偏好性的解释，另一方面不同尺度下聚类模式的兼容性启发我们设计多视图分子表示学习框架UniCorn

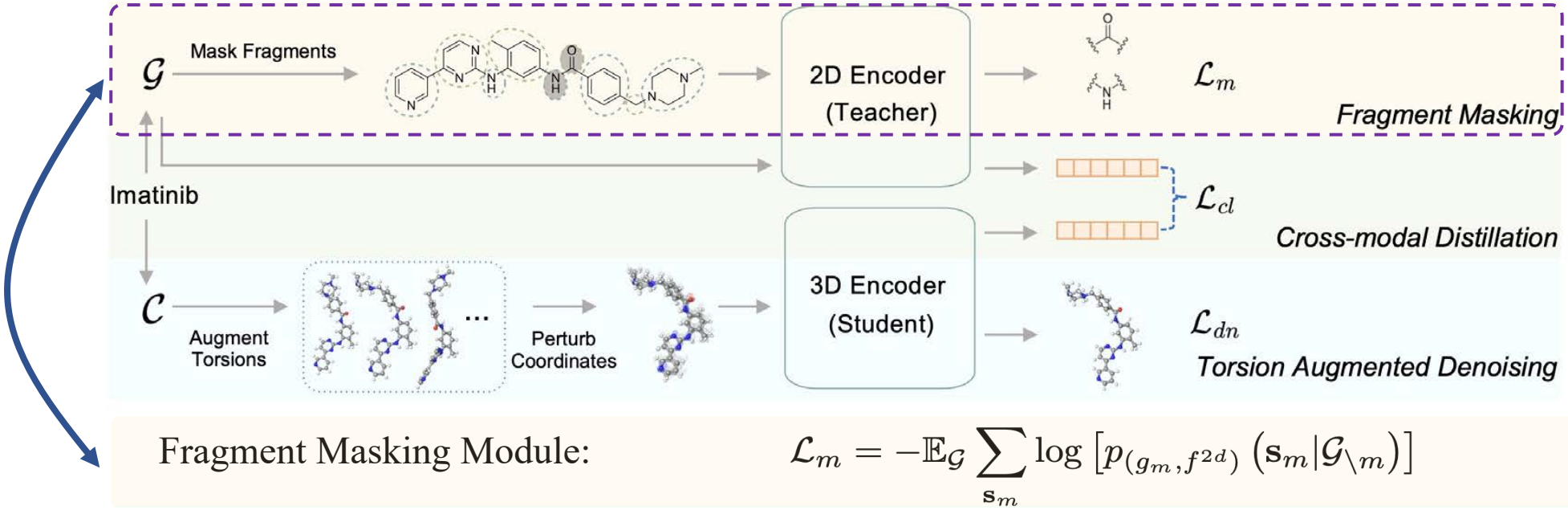
多任务统一的分子大模型UniCorn



基于以上分析提出UniCorn，包含了2D分子图片段遮挡、二面角增强去噪和跨模态蒸馏模块，这些模块捕捉了不同尺度的分子信息，学习层次化的多视图分子表示。

多任务统一的分子大模型UniCorn

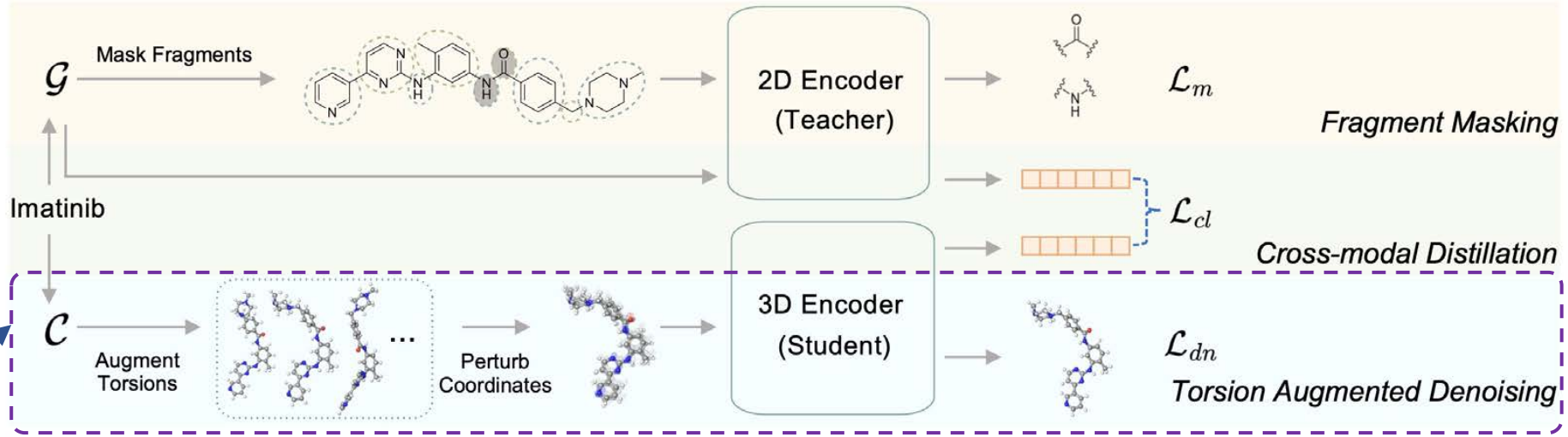
相似分子



Fragment Masking Module: $\mathcal{L}_m = -\mathbb{E}_{\mathcal{G}} \sum_{\mathbf{s}_m} \log \left[p_{(g_m, f^{2d})} \left(\mathbf{s}_m | \mathcal{G}_{\backslash m} \right) \right]$

基于以上分析提出UniCorn，包含了2D分子图片段遮挡、二面角增强去噪和跨模态蒸馏模块，这些模块捕捉了不同尺度的分子信息，学习层次化的多视图分子表示。

多任务统一的分子大模型UniCorn



相似构象

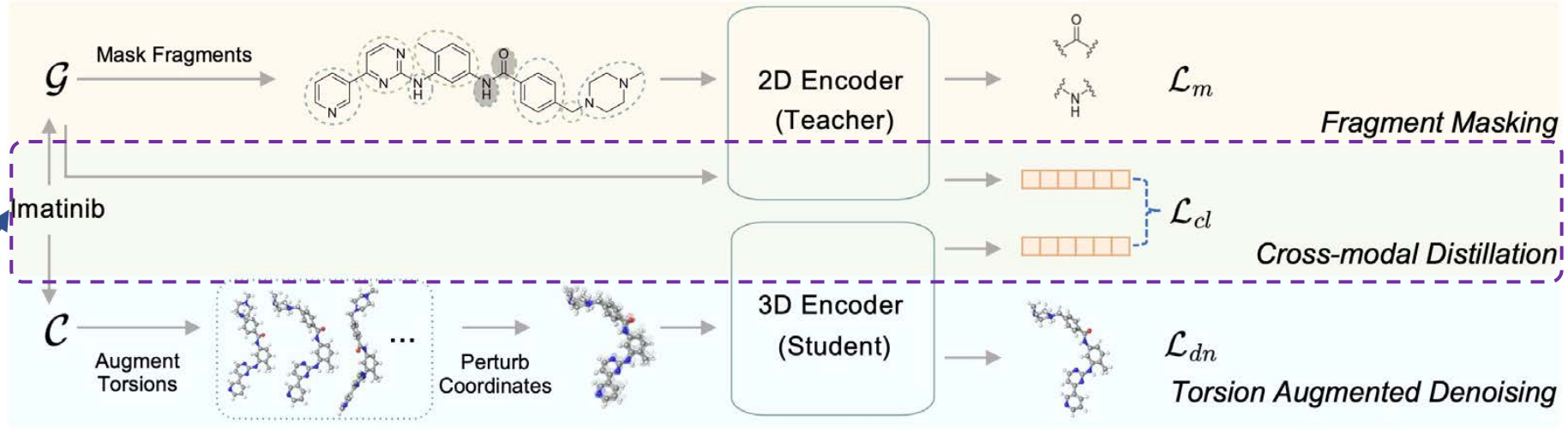
Fragment Masking Module: $\mathcal{L}_m = -\mathbb{E}_{\mathcal{G}} \sum_{\mathbf{s}_m} \log \left[p_{(g_m, f^{2d})} \left(\mathbf{s}_m | \mathcal{G}_{\backslash m} \right) \right]$

Torsion Augmented Denoising Module: $\mathcal{L}_{dn} = \mathbb{E}_{\mathcal{C}} || g_d(f^{3d}(\tilde{\mathcal{C}})) - (\tilde{\mathcal{X}} - \mathcal{X}_a) ||_2^2$

基于以上分析提出UniCorn，包含了2D分子图片段遮挡、二面角增强去噪和跨模态蒸馏模块，这些模块捕捉了不同尺度的分子信息，学习层次化的多视图分子表示。

多任务统一的分子大模型UniCorn

同一分子的不同构象&从2D蒸馏知识到3D Encoder



Fragment Masking Module: $\mathcal{L}_m = -\mathbb{E}_{\mathcal{G}} \sum_{\mathbf{s}_m} \log \left[p_{(g_m, f^{2d})} \left(\mathbf{s}_m | \mathcal{G}_{\backslash m} \right) \right]$

Torsion Augmented Denoising Module: $\mathcal{L}_{dn} = \mathbb{E}_{\mathcal{C}} || g_d(f^{3d}(\tilde{\mathcal{C}})) - (\tilde{\mathcal{X}} - \mathcal{X}_a) ||_2^2$

Cross-modal Distillation Module: $\mathcal{L}_{cl} = -\mathbb{E}_{(\mathcal{G},\mathcal{C})} [\log \frac{e^{\cos(\boldsymbol{z}^g, \boldsymbol{z}^c)/\tau}}{\sum_{\boldsymbol{z}_j^c} e^{\cos(\boldsymbol{z}^g, \boldsymbol{z}_j^c)/\tau}} + \log \frac{e^{\cos(\boldsymbol{z}^c, \boldsymbol{z}^g)/\tau}}{\sum_{\boldsymbol{z}_j^g} e^{\cos(\boldsymbol{z}^c, \boldsymbol{z}_j^g)/\tau}}]$

基于以上分析提出UniCorn，包含了2D分子图片段遮挡、二面角增强去噪和跨模态蒸馏模块，这些模块捕捉了不同尺度的分子信息，学习层次化的多视图分子表示。

多任务统一的分子大模型UniCorn

生理性质

Methods	Models	BBBP	Tox21	MUV	BACE	ToxCast	SIDER	ClinTox	HIV	Avg.
Graph Masking	AttrMask	65.0±2.3	74.8±0.2	73.4±2.0	79.7±0.3	62.9±0.1	61.2±0.1	87.7±1.1	76.8±0.5	72.7
	GROVER	70.0±0.1	74.3±0.1	67.3±1.8	82.6±0.7	65.4±0.4	64.8±0.6	81.2±3.0	62.5±0.9	71.0
	GraphMAE	72.0±0.6	75.5±0.6	76.3±2.4	83.1±0.9	64.1±0.3	60.3±1.1	82.3±1.2	77.2±1.0	73.9
	Mole-BERT	71.9±1.6	76.8±0.5	78.6±1.8	80.8±1.4	64.3±0.2	62.8±1.1	78.9±3.0	78.2±0.8	74.0
Multimodal	3D InfoMax	69.1±1.0	74.5±0.7	74.4±2.4	79.7±1.5	64.4±0.8	60.6±0.7	79.9±3.4	76.1±1.3	72.3
	GraphMVP	68.5±0.2	74.5±0.4	75.0±1.4	76.8±1.1	62.7±0.1	62.3±1.6	79.0±2.5	74.8±1.4	71.7
	MoleculeSDE	71.8±0.7	76.8±0.3	80.9±0.3	79.5±2.1	65.0±0.2	60.8±0.3	87.0±0.5	78.8±0.9	75.1
	MoleBLEND	73.0±0.8	77.8±0.8	77.2±2.3	83.7±1.4	66.1±0.0	**64.9±0.3**	87.6±0.7	79.0±0.8	76.2
	UniCorn	**74.2±1.1**	**79.3±0.5**	**82.6±1.0**	**85.8±1.2**	**69.4±1.1**	64.0±1.8	**92.1±0.4**	**79.8±0.9**	**78.4**

物理化学性质

Models	ESOL	FreeSolv	Lipo
AttrMask	1.112±0.048	-	0.730±0.004
GROVER	0.983±0.090	2.176±0.052	0.817±0.008
3D InfoMax	0.894±0.028	2.337±0.227	0.695±0.012
GraphMVP	1.029±0.033	-	0.681±0.010
MoleBLEND	0.831±0.026	1.910±0.163	0.638±0.004
UniCorn	**0.817±0.034**	**1.555±0.075**	**0.591±0.016**

UniCorn在MoleculeNet上的**生理性质**、**物理化学性质**任务都能取得SOTA的结果

多任务统一的分子大模型UniCorn

在QM9的量子力学性质任务中，UniCorn超过了之前的Denoising预训练方法，取得了10/12个子任务的最佳性能

Methods	Models	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$< R^2 >$ (a_0^2)	ZPVE (meV)	U_0 (meV)	U (meV)	H (meV)	G (meV)	C_v ($\frac{cal}{molK}$)
Multimodal	3D InfoMax	0.0280	0.057	25.9	21.6	42.1	0.141	1.67	13.30	13.81	13.62	13.73	0.030
	GraphMVP	0.0270	0.056	25.8	21.6	42.0	0.136	1.61	13.07	13.03	13.31	13.43	0.029
	MoleculeSDE	0.0260	0.054	25.7	21.4	41.8	0.151	1.59	12.04	12.54	12.05	13.07	0.028
	MoleculeJAE	0.0270	0.056	26.0	21.6	42.7	0.141	1.56	10.70	10.81	10.70	11.22	0.029
	MoleBLEND	0.0370	0.060	21.5	19.2	34.8	0.417	1.58	11.82	12.02	11.97	12.44	0.031
3D Denoising	Transformer-M	0.0370	0.041	17.5	16.2	27.4	0.075	**1.18**	9.37	9.41	9.39	9.63	0.022
	SE(3)-DDM	0.0150	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024
	3D-EMGP	0.0200	0.057	21.3	18.2	37.1	**0.092**	1.38	8.60	8.60	8.70	9.30	0.026
	Frad	0.0100	0.037	15.3	13.7	27.8	0.342	1.42	5.33	5.62	5.55	6.19	0.020
	UniCorn	**0.0085**	**0.036**	**13.0**	**11.9**	**24.9**	0.326	1.40	**3.99**	**3.95**	**3.94**	**5.09**	**0.019**

未来发展趋势

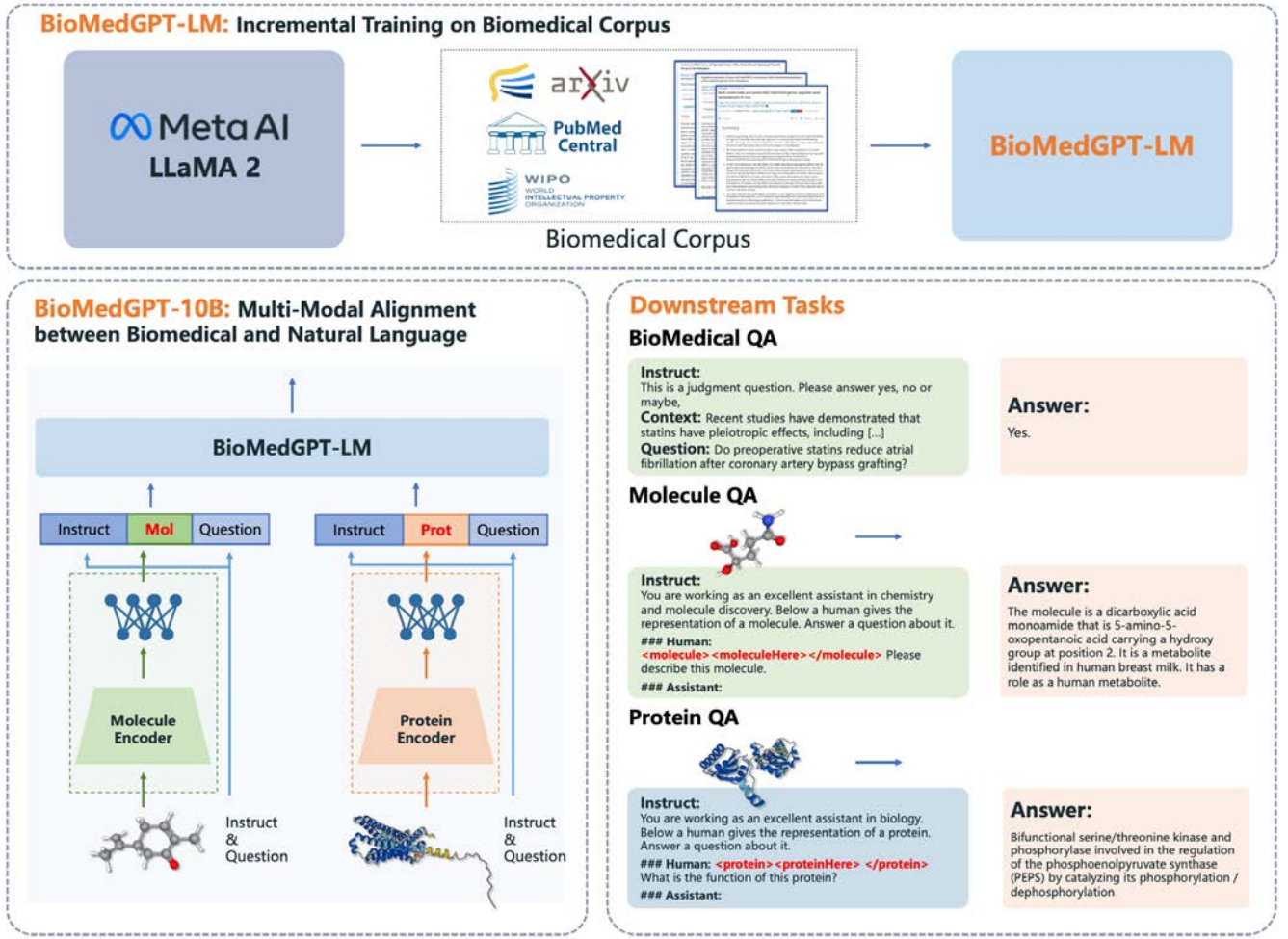
- **大语言模型和分子专有模型的结合**
  - 大语言模型蕴含的海量经验知识可以和精细建模分子的专有模型产生互补

- **分子和蛋白一体化建模**
  - 大小分子共享基本的组成单元，遵循统一的物理规律，一体化建模可以提升模型在不同domain的泛化能力

- **统一分子生成和理解**
  - 分子生成和理解都依赖于分子表示学习，统一生成和理解的模型可以同时提升两类任务的性能

多模态小分子大语言模型BioMedGPT

- BioMedGPT基于LLaMa2语言大模型，专注于医疗分子领域的多模态问答任务
- 构建了两个跨模态QA数据集 PubChem QA 和 Uniprot QA来对LLaMa2语言大模型进行指令微调



BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. 2023. (Tsinghua AIR)

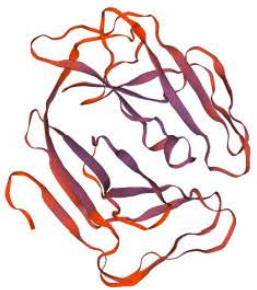
多模态小分子大语言模型BioMedGPT

在分子QA任务上， BioMedGPT显著超过了ChatGPT和Lalam2

Table 3: Performance comparison on molecule QA.

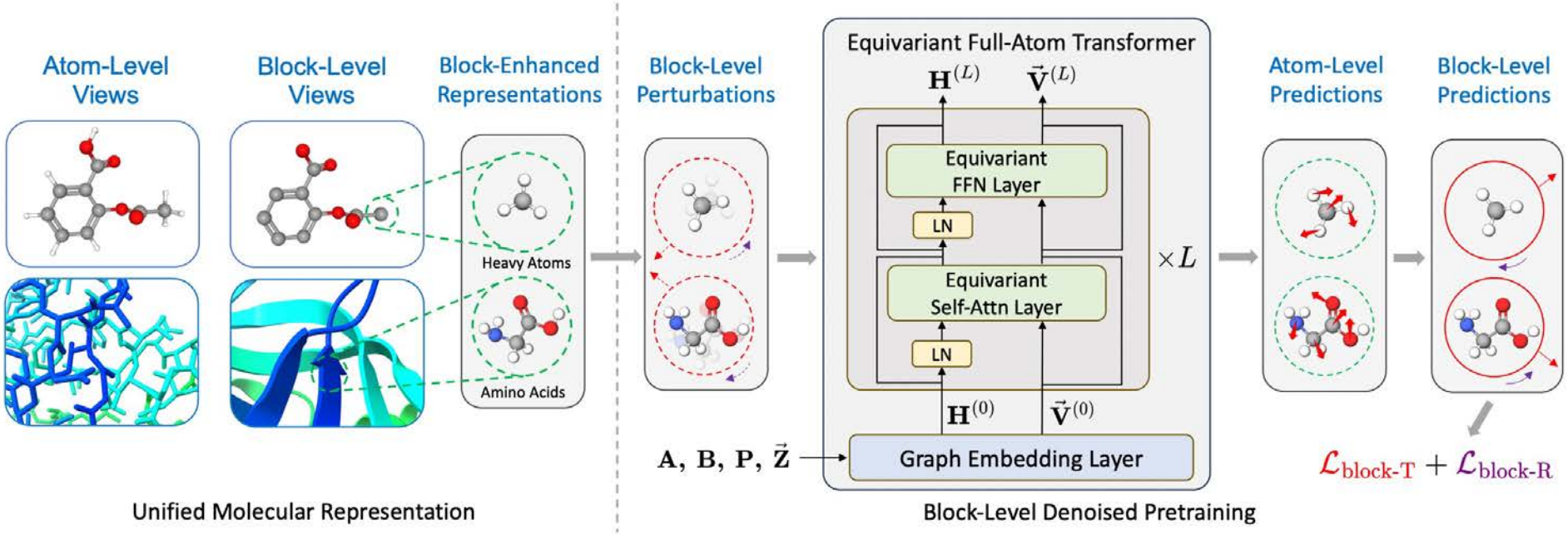
Method	Alignment	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	MEATOR
ChatGPT [Li et al., 2023]	w/o	0.103	0.050	0.261	0.088	0.204	0.161
Llama2-7B-Chat		0.075	0.009	0.184	0.043	0.142	0.149
BioMedGPT-10B	w/	**0.234**	**0.141**	**0.386**	**0.206**	**0.332**	**0.308**

在蛋白QA任务上，只有BioMedGPT给出了准确包含信息量的回答

Protein Input	Llama2-7B-chat (w/o Alignment)	Llama2-7B-chat (w/ Alignment)	BioMedGPT-10B	Ground Truth
P52341 · DUT_HHV7J (EC:3.6.1.23) 	🔍 To determine the function of the protein, we need to analyze its structure and sequence. Can you provide me with more information about the protein, such as its molecular weight, isoelectric point, and any post-translational modifications? 🤔	Has a role in the repair of DNA double-strand breaks. Can also act as a transcriptional repressor.	Involved in the biosynthesis of thymidine monophosphate (dTMP), a precursor of thymidine triphosphate (dTTP), which is a substrate for DNA synthesis.	Involved in nucleotide metabolism: produces dUMP, the immediate precursor of thymidine nucleotides and decreases the intracellular concentration of dUTP to avoid uracil incorporation into viral DNA.

全原子统一大模型EPT

统一小分子和蛋白质的分子表示学习模型EPT，对大小分子定义不同的输入block，对于小分子block是重原子，对于蛋白质则是氨基酸，以坐标去噪作为预训练任务



An Equivariant Pretrained Transformer for Unified 3D Molecular Representation Learning. ICLR GEM Workshop, 2024. (Tsinghua AIR)

全原子统一大模型EPT

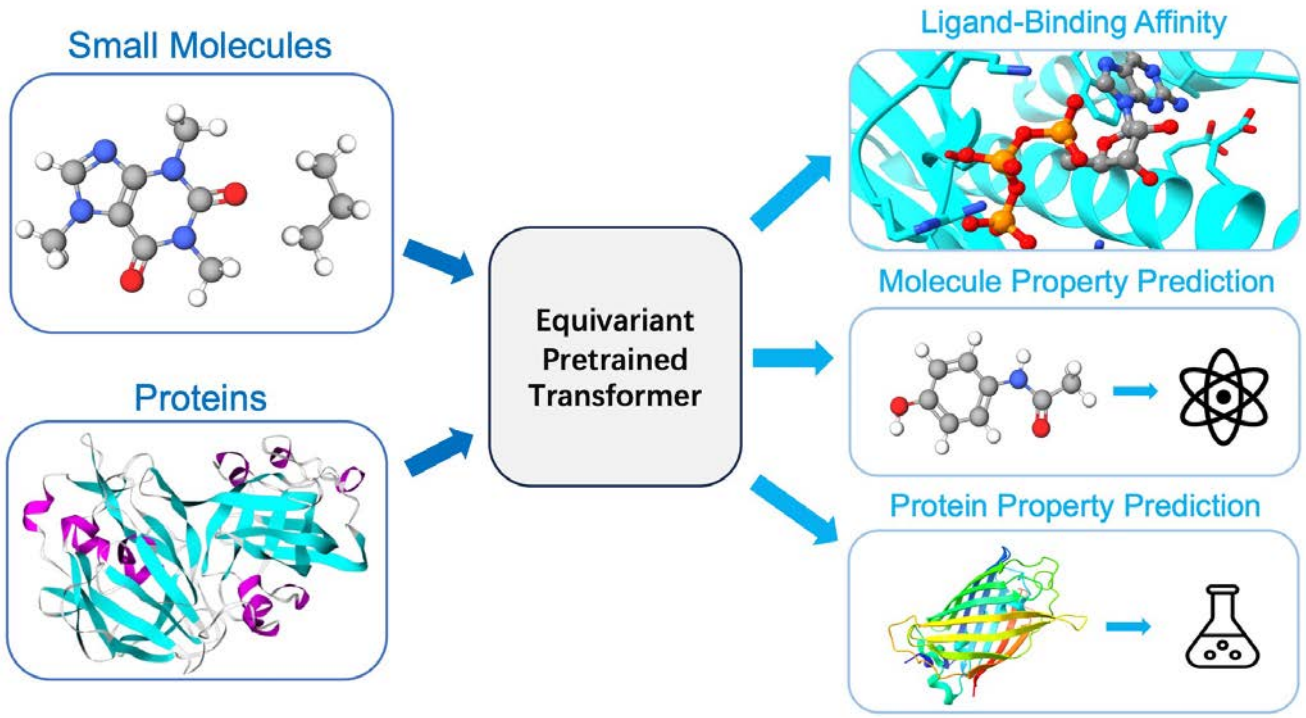


Figure 1. Equivariant Pretrained Transformer (EPT) aims at pre-training one model for multiple domains.

优势：统一模型可以同时兼顾小分子，蛋白性质预测，以及小分子蛋白交互任务（亲和力预测）

全原子统一大模型EPT

分子性质预测

Model	$\mu\downarrow$ (D)	$\alpha\downarrow$ (a_0^3)	$\epsilon_{\text{HOMO}}\downarrow$ (meV)	$\epsilon_{\text{LUMO}}\downarrow$ (meV)	$\Delta\epsilon\downarrow$ (meV)	$< R^2 >\downarrow$ (a_0^2)	ZPVE$\downarrow$ (meV)	$U_0\downarrow$ (meV)	$U\downarrow$ (meV)	$H\downarrow$ (meV)	$G\downarrow$ (meV)	$C_v\downarrow$ ($\frac{cal}{molK}$)	Avg.$\downarrow$ Rank
SchNet	0.033	0.235	41.0	34.0	63.0	0.070	1.70	14.00	19.00	14.00	14.00	0.033	11.83
E(n)-GNN	0.029	0.071	29.0	25.0	48.0	0.110	1.55	11.00	12.00	12.00	12.00	0.031	11.17
DimeNet++	0.030	0.043	24.6	19.5	32.6	0.330	1.21	6.32	6.28	6.53	7.56	0.023	7.17
PaiNN	0.012	0.045	27.6	20.4	45.7	0.070	1.28	5.85	5.83	5.98	7.35	0.024	6.33
TorchMD-Net	0.011	0.059	20.3	18.6	36.1	**0.033**	1.84	6.15	6.38	6.16	7.62	0.026	7.08
Equiformer	0.011	0.046	**15.0**	14.0	30.0	0.251	1.26	6.59	6.74	6.63	7.63	0.023	6.00
Transformer-M	0.037	0.041	17.5	16.2	**27.4**	0.075	1.18	9.37	9.41	9.39	9.63	0.022	6.92
GeoSSL	0.015	0.046	23.5	19.5	40.2	0.122	1.31	6.92	6.99	7.09	7.65	0.024	8.42
3D-EMGP	0.020	0.057	21.3	18.2	37.1	0.092	1.38	8.60	8.60	8.70	9.30	0.026	8.83
DP-TorchMD-Net	0.012	0.052	17.7	14.3	31.8	0.450	1.71	6.57	6.11	6.45	6.91	**0.020**	6.67
Frad	**0.010**	**0.037**	15.3	13.7	27.8	0.342	1.42	**5.33**	5.62	5.55	**6.19**	**0.020**	3.17
EPT	0.011	0.045	16.2	14.1	29.6	0.122	1.14	5.53	5.70	5.52	6.42	**0.020**	3.33
EPT-10	**0.010**	0.045	15.2	**13.6**	29.0	0.152	**1.11**	5.44	**5.54**	**5.42**	6.37	**0.020**	**2.33**

蛋白性质预测

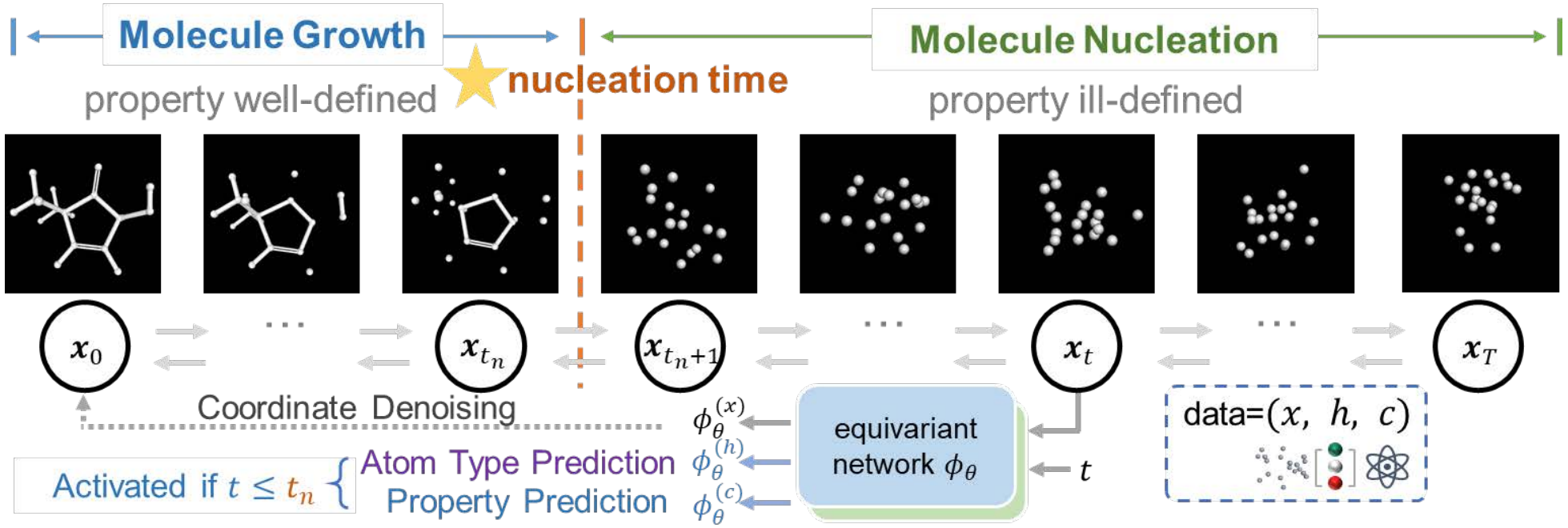
	Model	EC		MSP
		F1 Max	AUPRC	AUROC
w/o Pretrain	GCN	0.320	0.319	0.621
	Atom3D-CNN	-	-	0.574
	Atom3D-ENN	-	-	0.574
	GVP	0.489	0.482	0.680
	GearNet	0.730	0.751	-
	GearNet-Edge	0.810	0.835	0.633
	EPT (ours)	**0.823**	**0.844**	**0.700**
w/ Pretrain	LM-GVP	0.664	0.710	-
	ProtBERT-BFD	0.838	0.859	-
	GearNet-Edge	**0.874**	**0.892**	0.646
	EPT (ours)	0.858	0.871	**0.741**

蛋白分子亲和力预测

DeepAffnity	1.893 ± 0.650	0.415	0.426	–	–	–
GeoSSL	1.451 ± 0.030	0.577 ± 0.020	0.572 ± 0.010	–	–	–
EGNN-PLM	1.403 ± 0.010	0.565 ± 0.020	0.544 ± 0.010	1.559 ± 0.020	0.644 ± 0.020	0.646 ± 0.020
Uni-Mol	1.520 ± 0.030	0.558 ± 0.000	0.540 ± 0.000	1.619 ± 0.040	0.645 ± 0.020	0.653 ± 0.020
ProFSA	1.377 ± 0.010	0.628 ± 0.010	0.620 ± 0.010	1.377 ± 0.010	0.764 ± 0.000	0.762 ± 0.010
EPT-Scratch	1.356 ± 0.041	0.604 ± 0.022	0.591 ± 0.025	1.303 ± 0.015	0.777 ± 0.001	0.776 ± 0.003
EPT-Molecule	1.325 ± 0.007	0.627 ± 0.006	0.618 ± 0.004	1.263 ± 0.022	0.791 ± 0.006	0.783 ± 0.006
EPT-Protein	1.326 ± 0.035	0.628 ± 0.014	0.611 ± 0.019	1.223 ± 0.014	0.805 ± 0.002	0.803 ± 0.004
EPT-MultiDomain	$\mathbf{1.318 \pm 0.020}$	$\mathbf{0.643 \pm 0.005}$	$\mathbf{0.630 \pm 0.005}$	$\mathbf{1.165 \pm 0.007}$	$\mathbf{0.822 \pm 0.002}$	$\mathbf{0.819 \pm 0.002}$

EPT在小分子、蛋白性质预测和蛋白质小分子亲和力预测任务上都达到了SOTA。

理解与生成一体的分子模型UniGEM



- 提出两阶段分子扩散生成: 分子成核前只生成原子位置，扩散生成算法基于EDM或BFN;
- 成核后同时预测原子类型和分子性质, 生成连续与离散分子数据的新范式

$$\mathcal{L}_t^{(x)} = \mathbb{E}_{q(\boldsymbol{x}_0,\boldsymbol{x}_n)}\|\boldsymbol{\phi}_\theta^{(x)}(\boldsymbol{x}_t,t) - \boldsymbol{\epsilon}_t\|^2, \quad \boldsymbol{\epsilon}_t = (\boldsymbol{x}_t - \alpha_t\boldsymbol{x}_0)/\sigma_t, \quad t \in [1,T]$$

$$\mathcal{L}_t^{(h)} = \mathbb{E}_{q(\boldsymbol{x}_0,\boldsymbol{x}_n)}|\boldsymbol{\phi}_\theta^{(h)}(\boldsymbol{x}_t,t) - \boldsymbol{h}|, \quad t \in [1,t_n], \quad \mathcal{L}_t^{(c)} = \mathbb{E}_{q(\boldsymbol{x}_0,\boldsymbol{x}_n)}||\boldsymbol{\phi}_\theta^{(c)}(\boldsymbol{x}_t,t) - c||^2, \quad t \in [1,t_n],$$

UniGEM: A Unified Approach to Generation and Property Prediction for Molecules. ICLR 2025. (AIR)

理解与生成一体的分子模型UniGEM

无条件生成

在EDM、GeoBFN基础上显著提升

#Metrics	QM9 Atom sta(%)	QM9 Mol sta(%)	QM9 Valid(%)	QM9 V*U(%)	GEOM-Drugs Atom sta(%)	GEOM-Drugs Valid(%)
E-NF	85.0	4.9	40.2	39.4	-	-
G-Schnet	95.7	68.1	85.5	80.3	-	-
EDM	98.7	82.0	91.9	90.7	81.3	92.6
GDM	97.6	71.6	90.4	89.5	77.7	91.8
EDM-Bridge	98.8	84.6	92.0	90.7	82.4	92.8
GeoLDM	98.9	89.4	93.8	92.7	84.4	**99.3**
UniGEM	**99.0** +0.3%	**89.8** +7.8%	**95.0** +3.1%	**93.2** +2.5%	**85.1** +3.8%	98.4 +5.8%
GeoBFN	90.9	99.1	95.3	93.0		
UniGEM(w/ BFN)	**93.7** +2.8%	**99.3** +0.2%	**97.3** +2.0%	93.0 +0.0%		

条件生成

用模型自带的性质预测模型对扩散生成做Guidance

$$x_{t-1} = \frac{1}{\alpha_{t|t-1}} x_t - \frac{\sigma^2_{t|t-1}}{\alpha_{t|t-1}\sigma_t}\phi^{(x)}_\theta(x_t, t) \boxed{- \lambda \nabla L^{(c)}_t} + \frac{\sigma_{t|t-1}\sigma_{t-1}}{\sigma_t}\epsilon$$

	ϵ_{LUMO} (eV)	ϵ_{HOMO} (eV)	$\Delta\epsilon$ (eV)	μ (D)	α (bohr^3)	C_v ($\frac{cal}{mol \cdot K}$)
Conditional EDM(Average MAE↓)	0.606	0.356	0.665	1.111	2.76	1.101
Guided UniGEM (Average MAE↓)	**0.592**	**0.233**	**0.511**	**0.805**	**2.22**	**0.873**

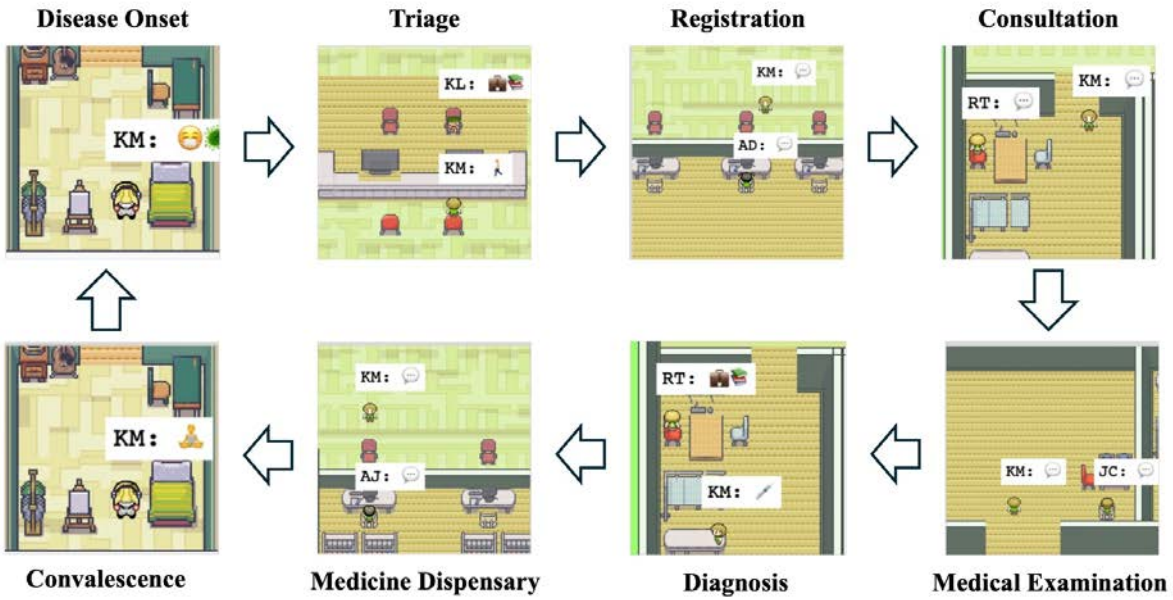
性质预测

超过用额外数据预训练的模型

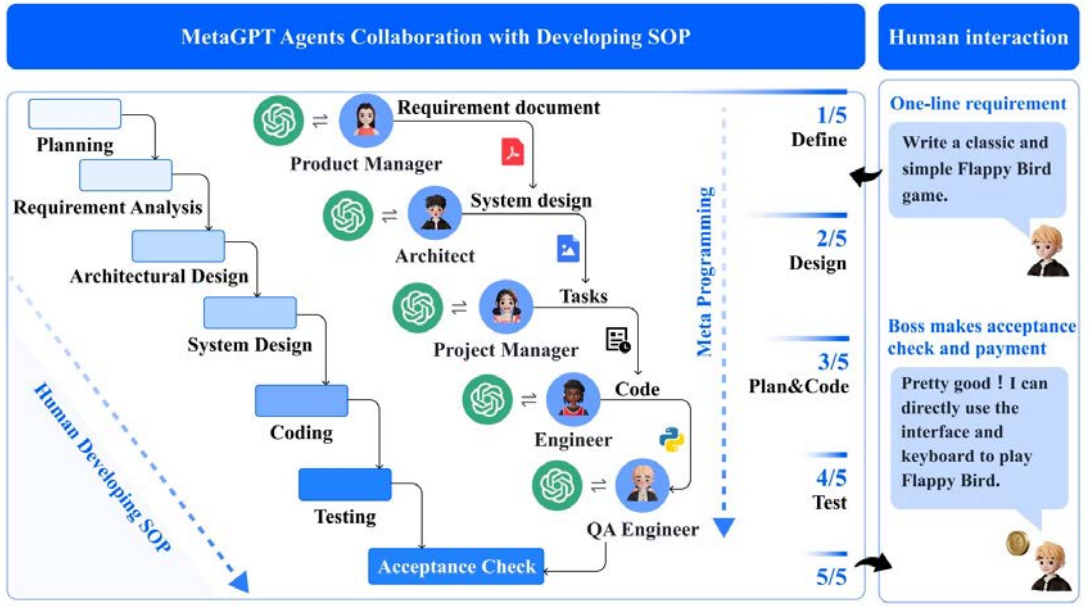
Task (Units)	α (bohr^3)	$\Delta\epsilon$ (meV)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	μ (D)	C_v ($\frac{cal}{mol \cdot K}$)
EGNN	0.071	48	29	25	0.029	0.031
GraphMVP	0.070	46.9	28.5	26.3	0.031	0.033
3D Infomax	0.075	48.8	29.8	25.7	0.034	0.033
GEM	0.081	52.1	33.8	27.7	0.034	0.035
3D-EMGP	**0.057**	37.1	21.3	18.2	0.020	0.026
UniGEM	0.060 -15.5%	**34.5** -28.1%	**20.9** -27.9%	**16.7** -33.2%	**0.019** -34.5%	**0.023** -25.8%

大语言模型多智能体系统

大语言模型多智能体系统（LLM-based Multi-Agent System）： 是指由多个自主智能体组成的系统，每个智能体借助大语言模型（如GPT）理解自然语言、规划任务并协作完成复杂目标，从而实现更高水平的智能行为协调与问题解决。

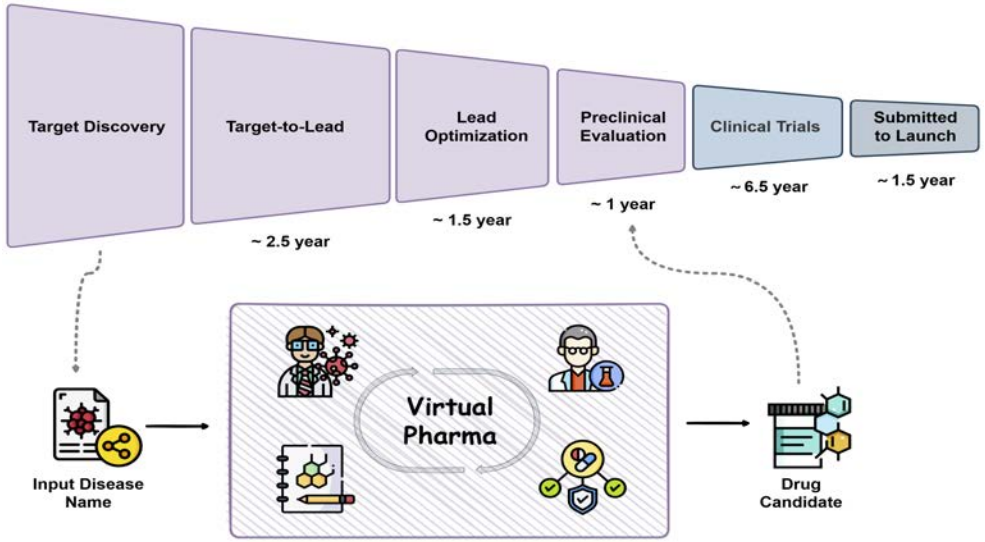


Agent Hospital: 模拟医院诊断治疗疾病的全过程闭环系统



MetaGPT: 模拟软件开发各个环节流程

药物发现智能体

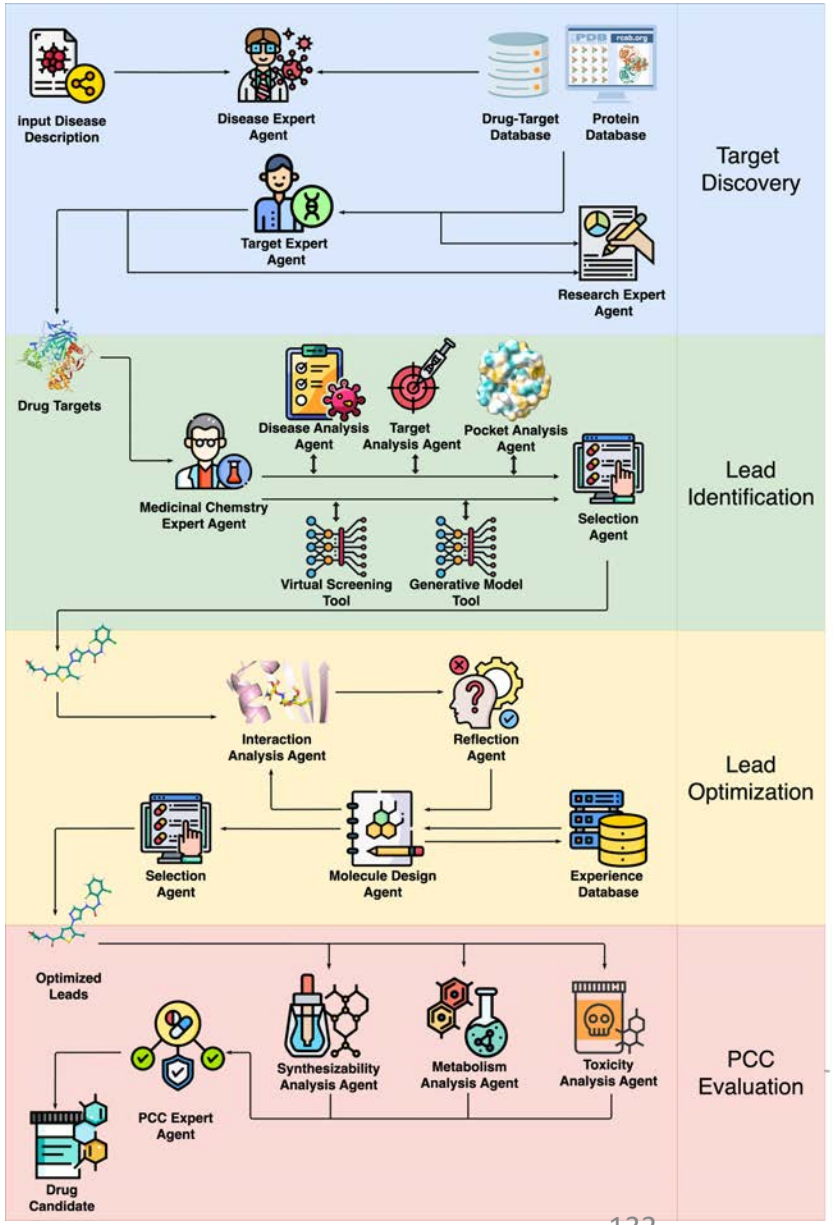


问题：

- 传统小分子药物研发是一个高度复杂、资源密集且周期漫长的过程。
- 现有机器学习模型驱动的药物设计模型多为独立工具，缺乏端到端整合，且“黑箱”特性导致可解释性不足

解决思路：将药物研发划分成若干子任务，每个任务均由LLM驱动的AI智能体完成。

PharmAgents：模拟了完整的虚拟制药公司，模拟小分子药物发现的整个流程，从靶点发现、先导化合物识别、先导优化、临床前评价。



PharmAgents: Building a Virtual Pharma with Large Language Model Agents. 2025. (Tsinghua AIR)

总结：智能药物研发中的AI4S问题

- 蛋白质结构预测 —— 回归/生成
- 大规模药物虚拟筛选 —— 信息检索
- 生成式药物设计 —— 生成式人工智能
- 分子大模型及药物性质预测 —— 大模型
- 药物发现智能体 —— 智能体

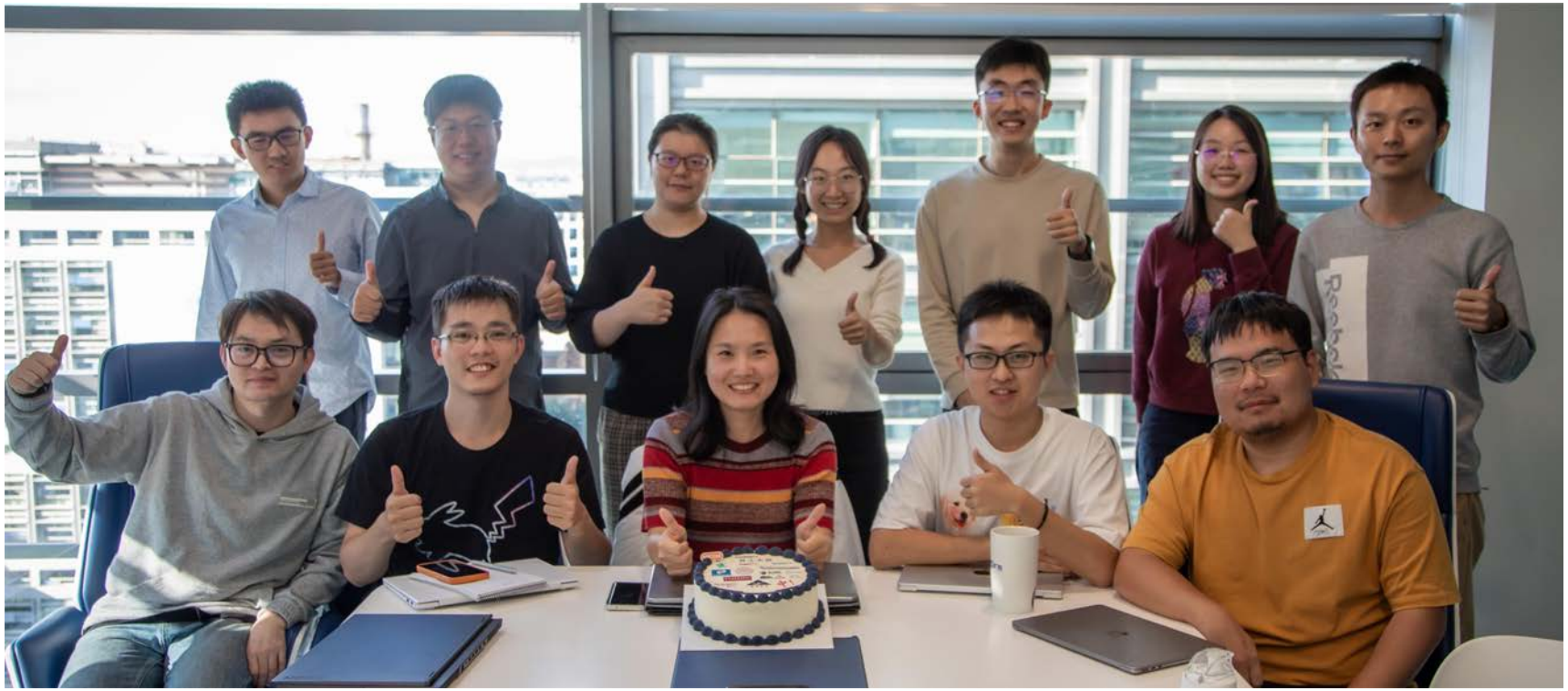
AI for Science认知和感想

- AI的未来：奔赴科学的星辰大海
- 价值驱动：好的问题比创新技术更重要
- Aim high，and stay grounded
- 终极目标：创造新的科学发现
- AIDD重要技术方向

  - 基础：下一代分子生成模型，基于强化学习的科学推理大模型，生命世界模型
  - 应用：先导化合物优化，干湿闭环，AI生命模拟器，药物发现智能体等

Acknowledgement：ATOM Lab

AI Transforming Optimal Medicine



PI：兰艳艳

博士后：洪鑫、贾寅君、张钰莹

科研助理：朱文钰、李红良

博士生：冯世坤、倪雨嫣、谭海川、高博文、林碧澄、吴科霖

实习生：黄彦雯、刘亦乔、樊高凡、谢行思、陈子陶、田子桐、王舰辉、谭好江、臧璇

欢迎申请PhD、博士后、实习生、科研助理！

群聊：AI4S探索

清華大學 TSINGHUA UNIVERSITY

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

谢谢大家！

兰艳艳

lanyanyan@air.tsinghua.edu.cn