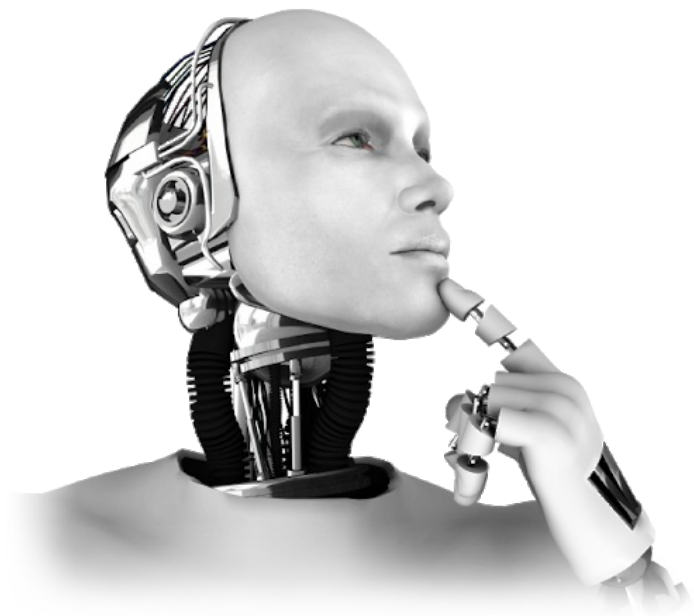




第十四周 人工智能助力科学

李泽椿，复旦大学生物医学工程与技术创新学院

目录

1 AI For Science介绍

2 AI应用于微观科学

3 AI应用于气象学

4 AI应用于生物医学

5 AI应用于其他领域

- **科学研究四范式**

- **实验归纳**：伽利略时代

- **模型推演**：牛顿三定律 -> 经典力学，薛定谔方程 -> 量子力学

- **计算机仿真**：计算物理、计算化学等

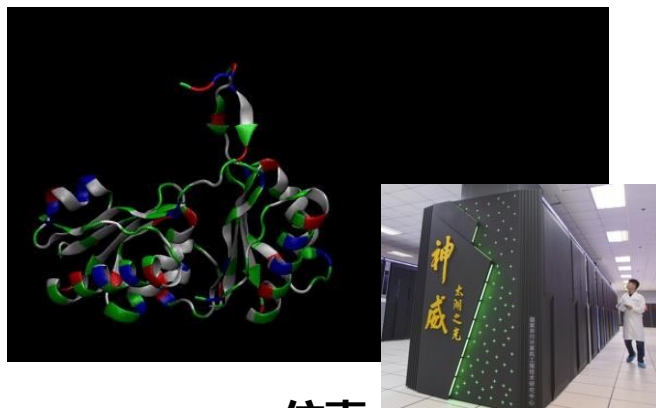
- **AI4Science**：人工智能（微软称为第五范式）



实验

$$F = ma$$
$$-\frac{\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi$$

理论



仿真



数据

AI4Science 与 Science的关系



$$\mathbf{F} = m\mathbf{a}$$

$$-\frac{\hbar^2}{2m}\nabla^2\psi + V\psi = E\psi$$



实验

理论

仿真

是什么

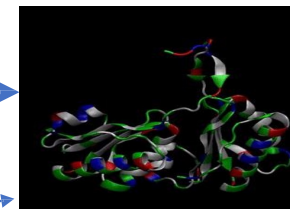
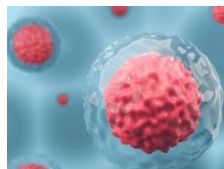
为什么

怎么样

感知

认知

反演



AI For Science



- 科学家们的优势和重点：高通量实验和计算；薄弱项：人工智能和机器学习。
- 有非常好的原材料，但是没有好的厨师
- AI + Science

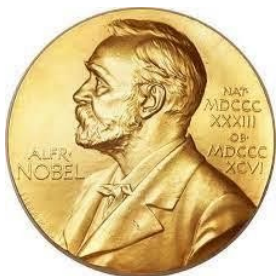


+



=





科研价值

计算是理解实验现象不可缺少的工具

诺贝尔奖：**密度泛函理论**（1998年）、多尺度建模（2013年）
狄拉克奖：第一性原理模拟方法（2009年）



社会价值

加速**电池材料**研发，助力实现碳达峰、碳中和

加速**新药物设计**：感冒特效药、抗癌药物等，惠及社会大众



经济价值

2021年中国锂离子电池行业产值突破**6000亿元**

2021年中国医药工业总营收超**3万亿元**

AI for Science 2025 — 全球最新进展

● 范式转变：AI驱动科学发现

- 深度学习、生成模型、强化学习识别海量数据中的隐藏规律
- AI自动生成假说、自主设计实验，加速跨学科融合

● 2024-2025标志性突破

- AlphaFold3 (Nature 2024): 预测所有生物分子结构，推进药物发现
- GenCast (Nature 2025): AI概率气象，超越欧洲中期气象预报中心
- A-Lab (UC Berkeley): 机器人+机器学习，全自主材料合成实验室
- Princeton / DeepMind: 强化学习稳定托卡马克等离子体，推进核聚变



- **AI共性的**

- 小样本
- 少标注
- 数据形式多样（原子、分子、基因、蛋白、大气）
- 可解释性

- **独特的**

- 同时需要人工智能和科学背景知识
- 理论、实验、计算三足鼎立

AI4Science 核心挑战

● 跨尺度建模

- 现有AI模型局限单一尺度，缺乏从原子到宏观的多尺度耦合
- 路径：物理定律嵌入AI（灰盒模型）；跨尺度统一神经网络架构

● 模型泛化能力

- 高质量科学数据稀缺，模型难适应新领域
- 路径：生成模型合成训练数据；迁移学习与少样本学习

● 可解释性

- 深度学习黑盒限制科学家信任，高风险场景尤为突出
- 路径：物理信息神经网络（PINN）；统计因果推断

● 算力与效率

- 超大模型训练算力需求攀升，以算力换算法不可持续
- 路径：混合精度训练、模型蒸馏；神经形态与量子计算协同

目录

1 AI For Science介绍

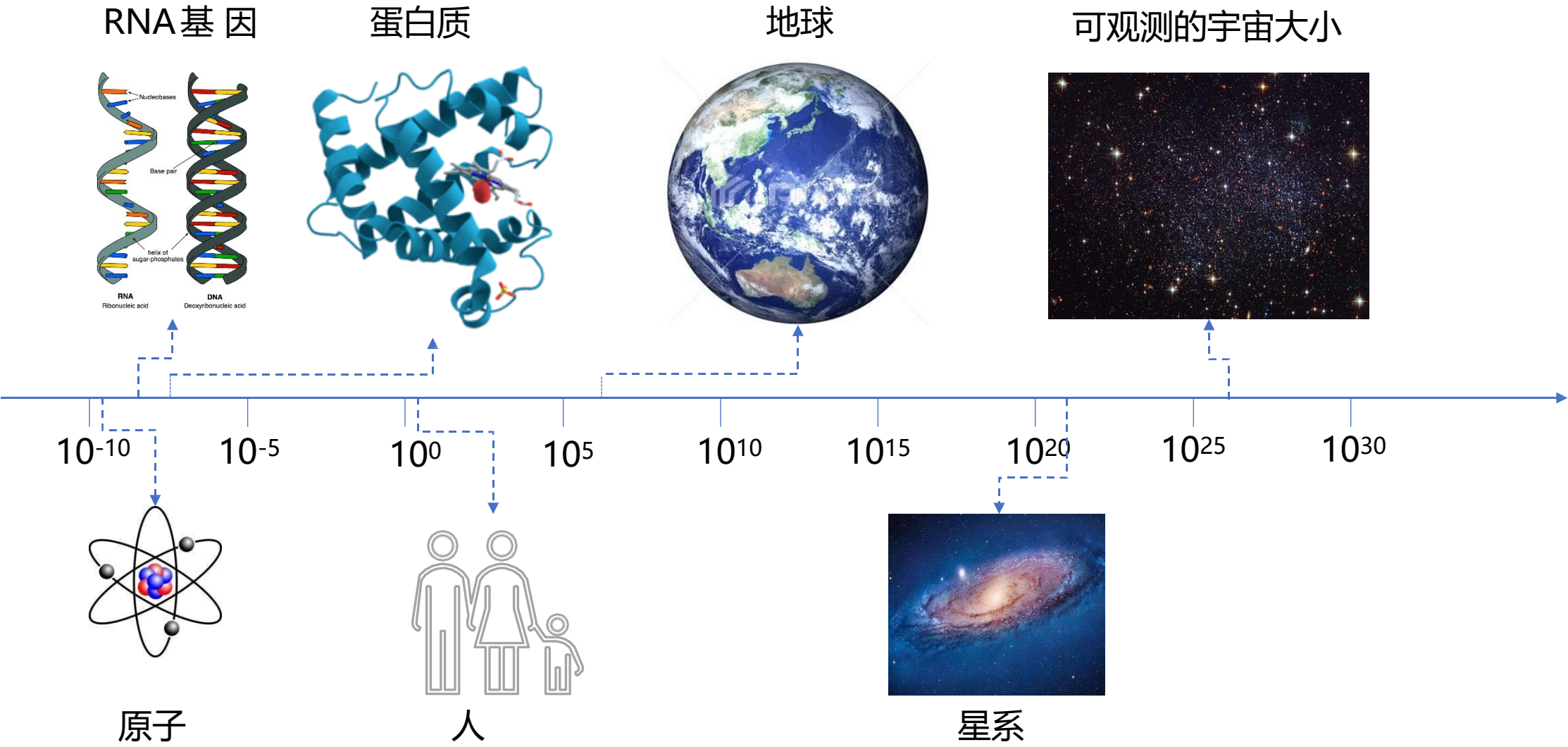
2 **AI应用于微观科学**

3 AI应用于气象学

4 AI应用于生物医学

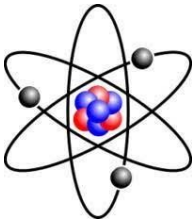
5 AI应用于其他领域

研究内容 – 探寻微观到宏观跨学科的共性AI算法

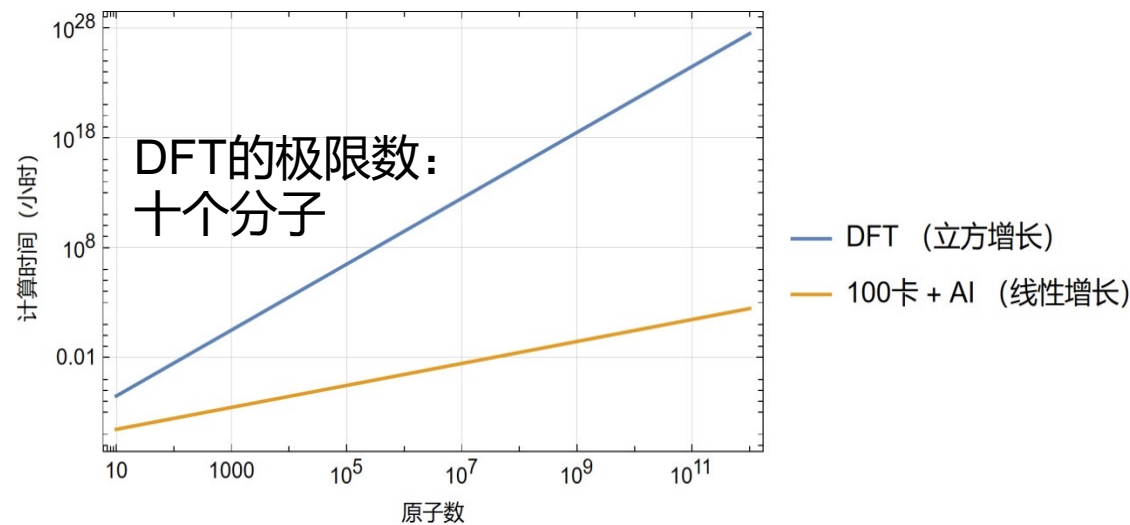
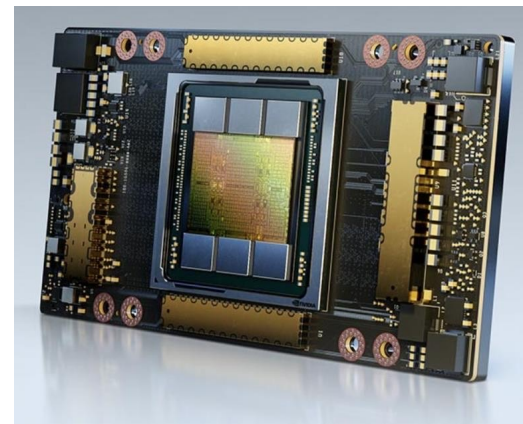
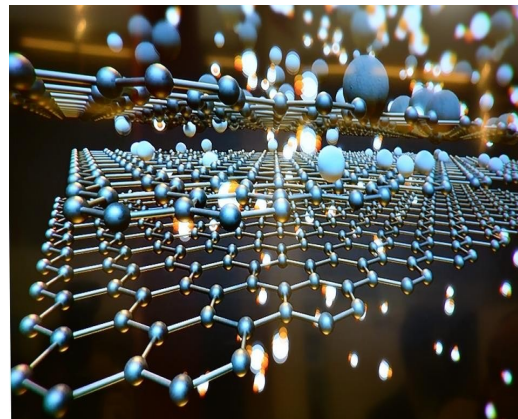


我们的目标：利用人工智能加速材料研发



原子级 ($\sim 10^{-10}$) 

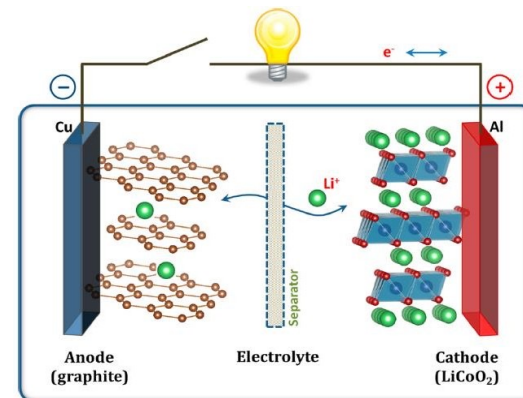
- 材料是科学技术进步的重要载体
 - 减少实验的成本
- 当前材料研究领域面临的瓶颈问题
 - 高精度计算模拟
 - 提速



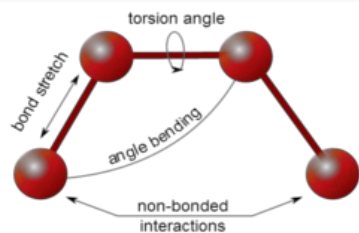
人工智能如何加速电池的研究?



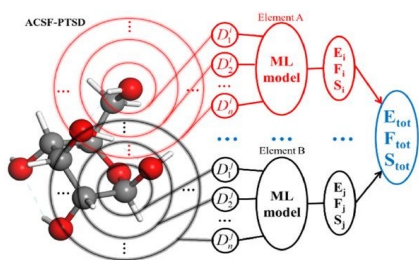
常见的锂离子电池



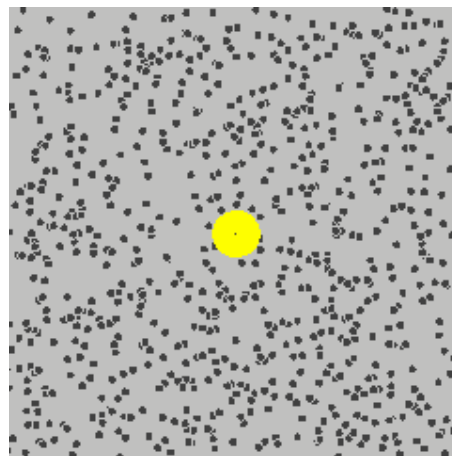
锂离子电池的内部结构示意图



传统方法：用经验公式计算受力信息



人工智能：用机器学习模型学习受力信息



利用势函数进行分子动力学模拟

溶剂化结构
电导率
离子迁移率
化学反应路径
.....

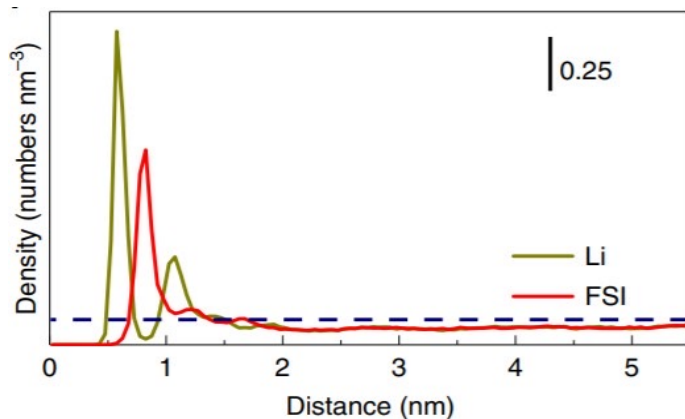
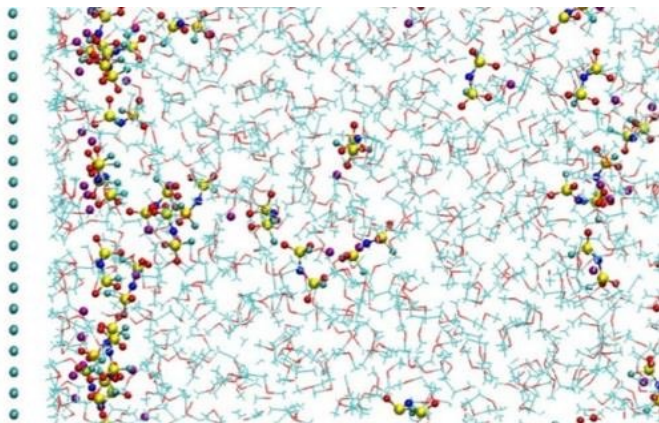
计算性质，用于指导电池设计



计算物理在电池研究中的应用



1. 分子动力学模拟研究电解质。
电压：2.8 V，存在电双层结构。

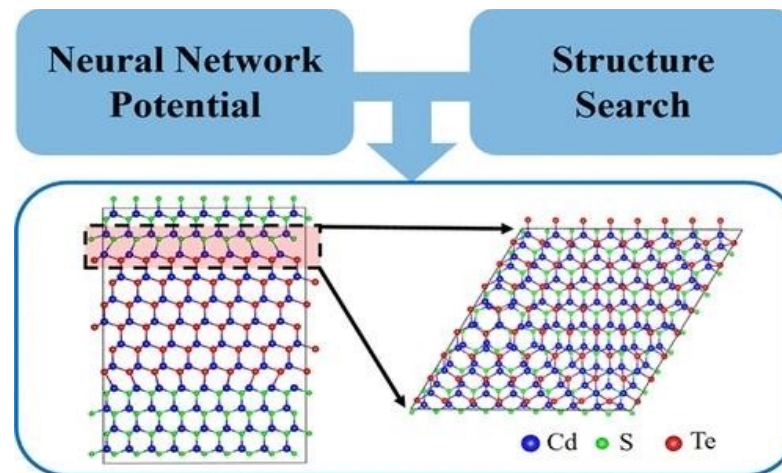


1. Zhou Y. et al. Real-time mass spectrometric characterization of the solid–electrolyte interphase of a lithium-ion battery. Nat. Nanotechnol, 2020.

2. 考虑关联效应修正的泊松-玻尔兹曼方程：
理论计算更加符合模拟结果。

$$-\nabla^2 u_{\pm}(r, r') = 4\pi l_B [z_+ \rho_+(r) \left(e^{-z_+ [u_{\pm}(r, r') + G_{0\pm}(r, r')] - v(|r-r'|)} - 1 \right) + z_- \rho_-(r) \left(e^{-z_- [u_{\pm}(r, r') + G_{0\pm}(r, r')] - v(|r-r'|)} - 1 \right)].$$

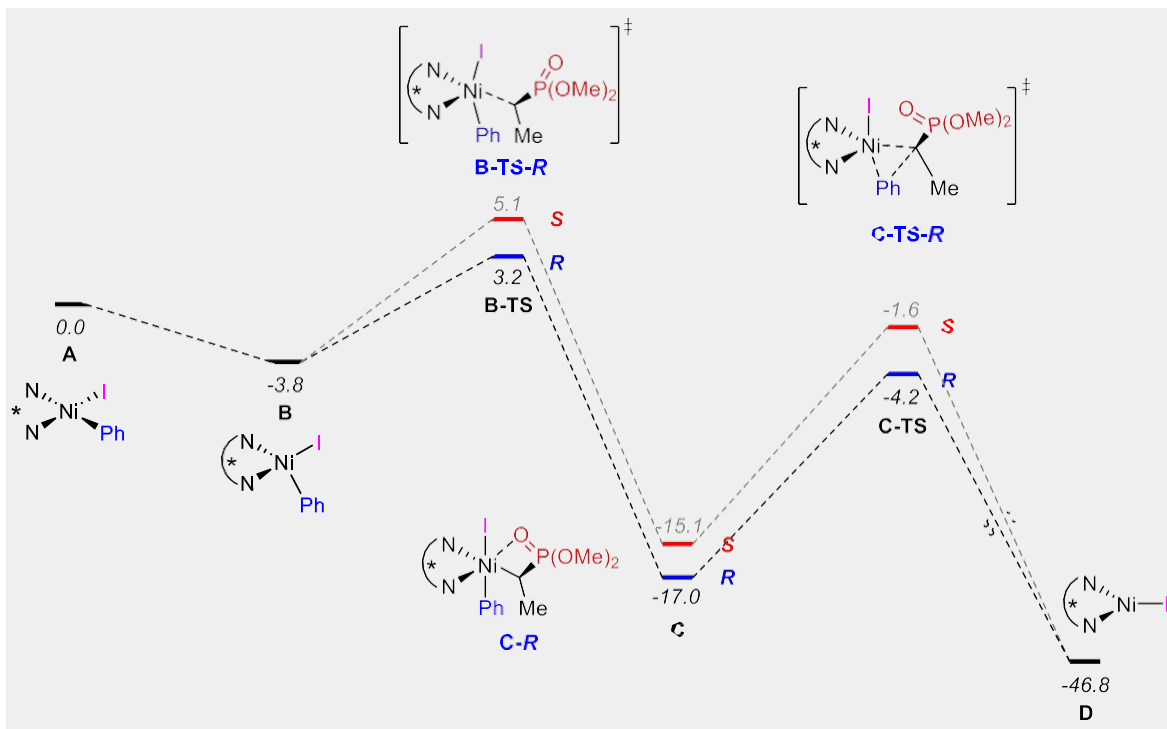
3. 利用神经网络势函数描述原子间相互作用力，
兼顾速度和精度。



2. Su M. et al. Poisson–Boltzmann theory with non-linear ion correlations. Journal of Physics: Condensed Matter, 2019.
3. Xu Y. et al. J. Phys. Chem. C, 2014.

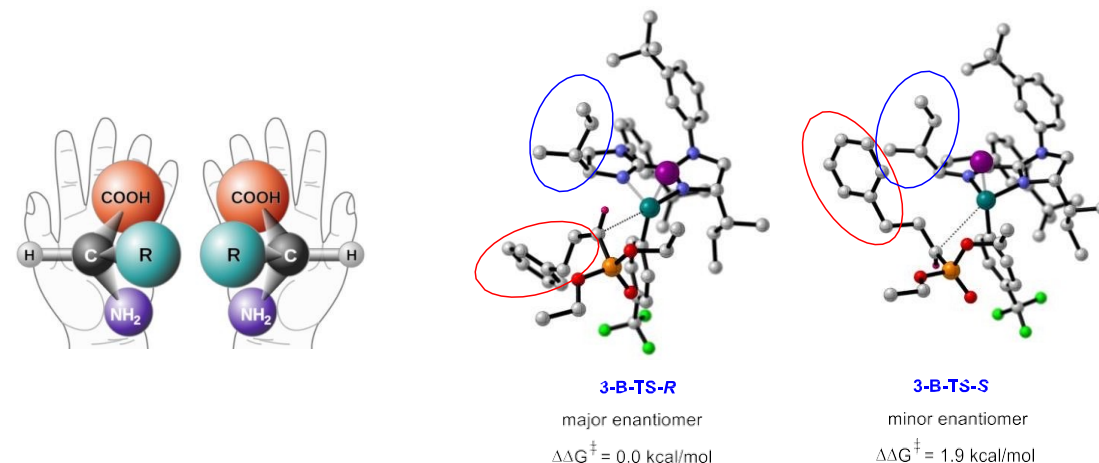
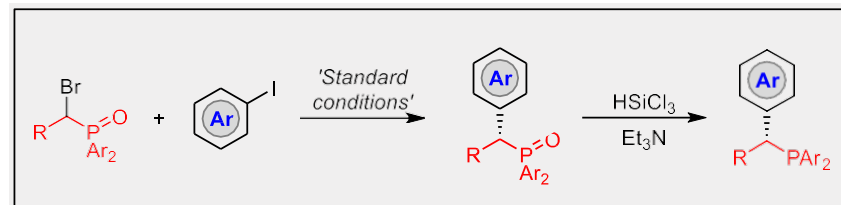


计算化学在化学反应研究中的应用



对化学反应的路径进行研究

原理：通过研究化学反应中分子几何结构和能量的关系，获得化学反应的势能面



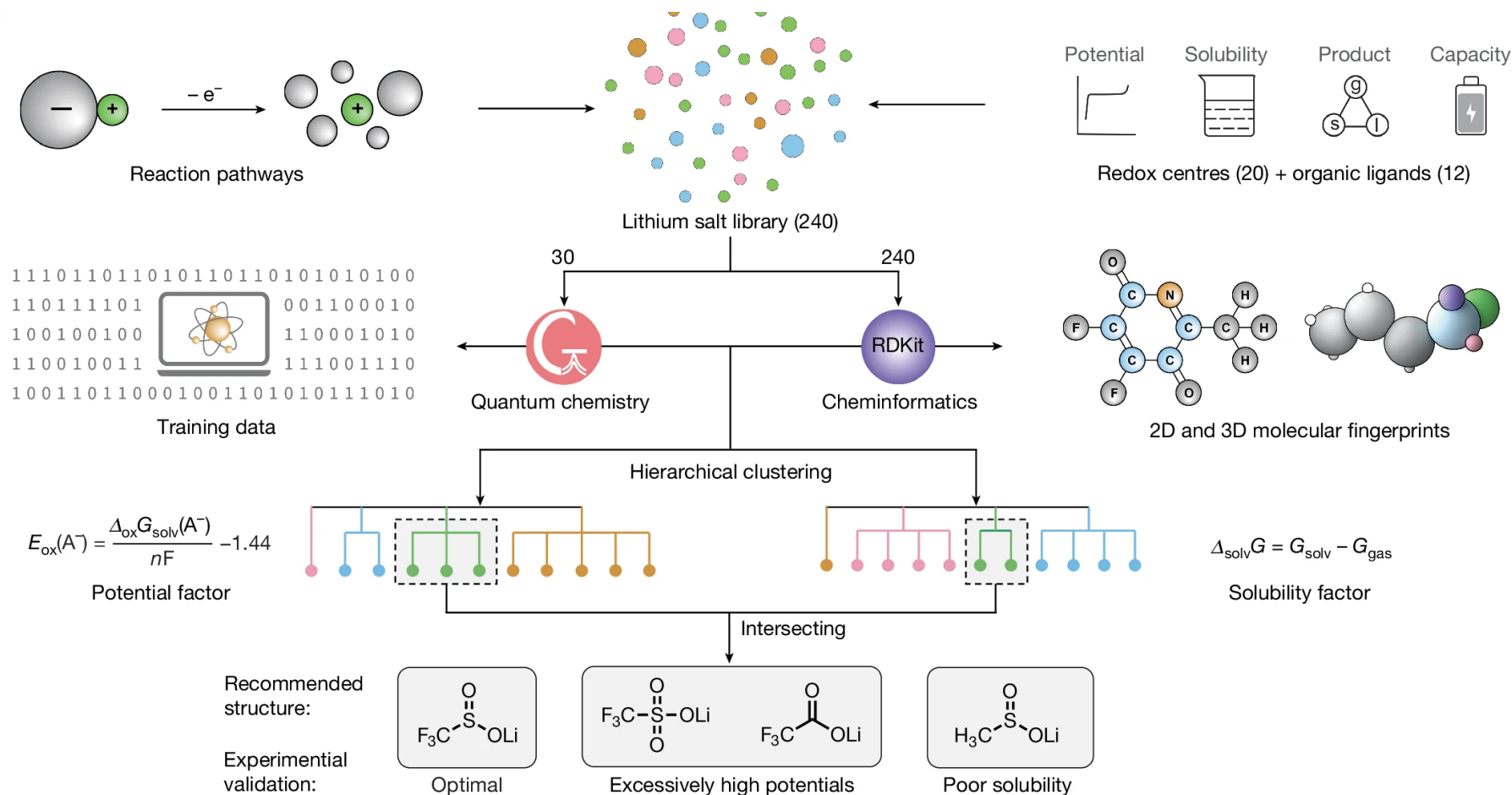
解释产物子类型

通过对过渡态能量——结构关系的分析，解释产物手性的控制

计算化学在电池研究中的应用 – 复旦研究



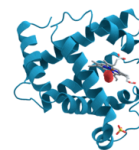
复旦大学：AI设计分子注入老化电池补充锂离子，突破性延长电池寿命至10000+次循环



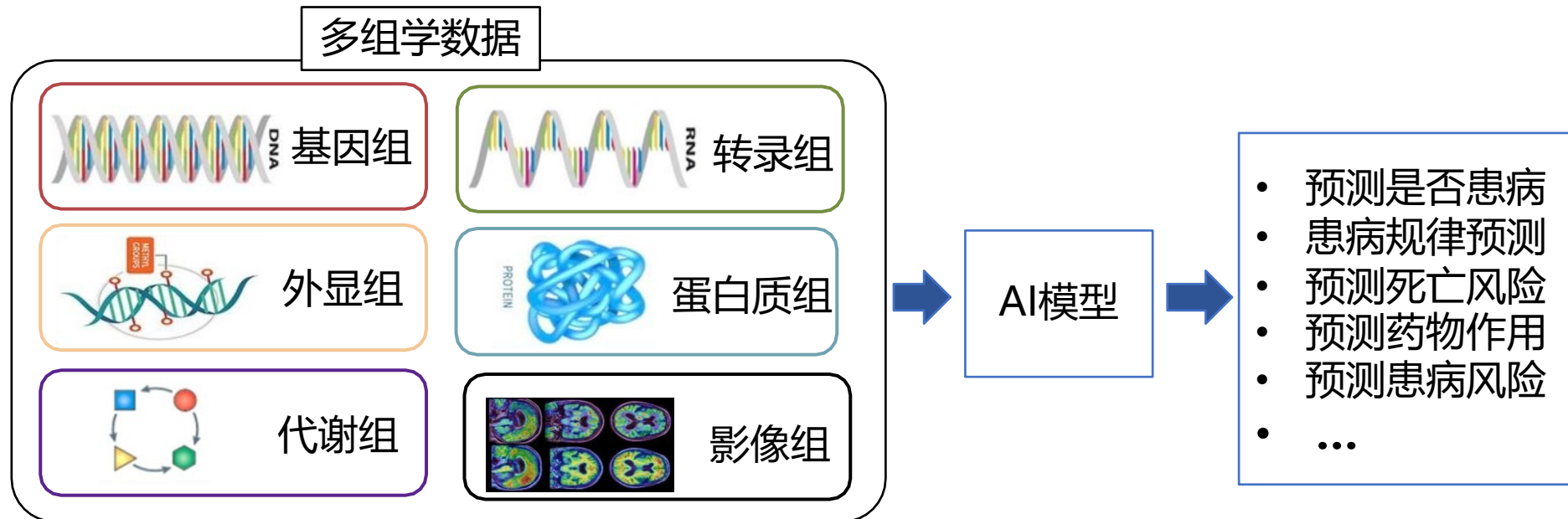
Chen S. et al. External Li supply reshapes Li deficiency and lifetime limit of batteries. Nature, 2025.



RNA级、蛋白质级 ($10^{-10} \sim 10^{-5}$)



应用：疾病诊断、病程预测、生存分析、治疗作用分析、风险预测 ...



基于多组学数据的疾病分析

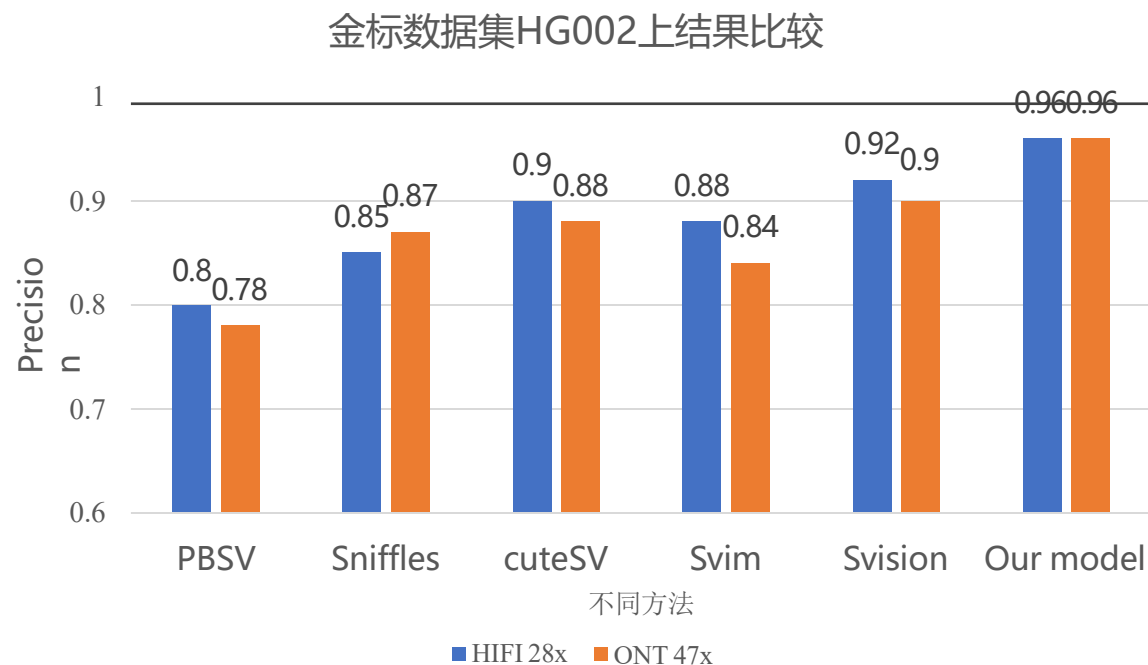
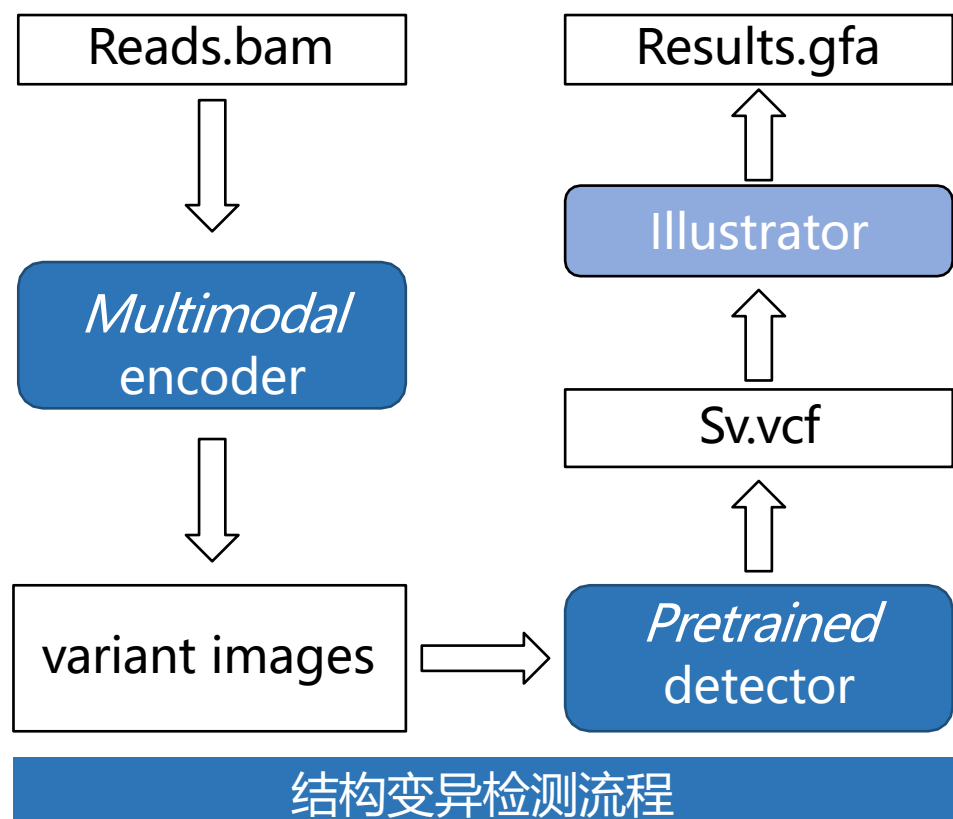
AI + 基因组学：基因结构变异检测



重要性：结构变异被证明与肿瘤等疾病有密切关系

问题：由于结构变异具有类型复杂、多尺度等特点，现有平台对结构变异检测率低

方法：采用多模态自监督学习策略，实现对复杂结构变异更精准的检测



实验结果

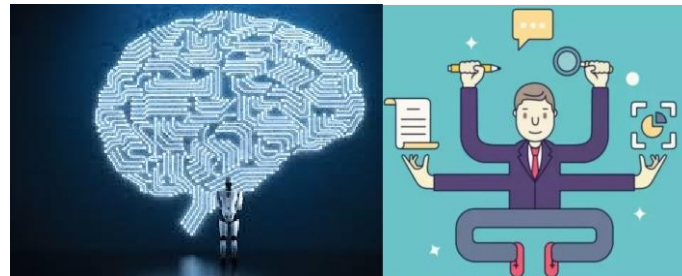


AI + 基因组学：通用DNA大模型

目标: 构建DNA大模型，服务大量的DNA任务，推动DNA研究



+



支持的DNA下游任务： DNA

大模型

- | | |
|---------------------------------------|---|
| 1. 启动子预测 (2019 Frontiers in genetics) | 7. 增强子活性预测 (2022 NG) |
| 2. 转录因子结合位点预测 (2015 NBT) | 8. 基因表达量预测 (2020 Cell reports, 2021 NM) |
| 3. 变异效应预测 (2001 NAR, cell 2023) | 9. 染色质谱预测 (2015 NM) |
| 4. DNA甲基化预测 (2022 Genome Biology) | 10. 模型跨物种分析 |
| 5. 表观遗传标记预测 (2005 Cell) | 11. 预测DNA结构 |
| 6. 增强子预测 (2016 Bio) | 12. 细胞特异性任务 |

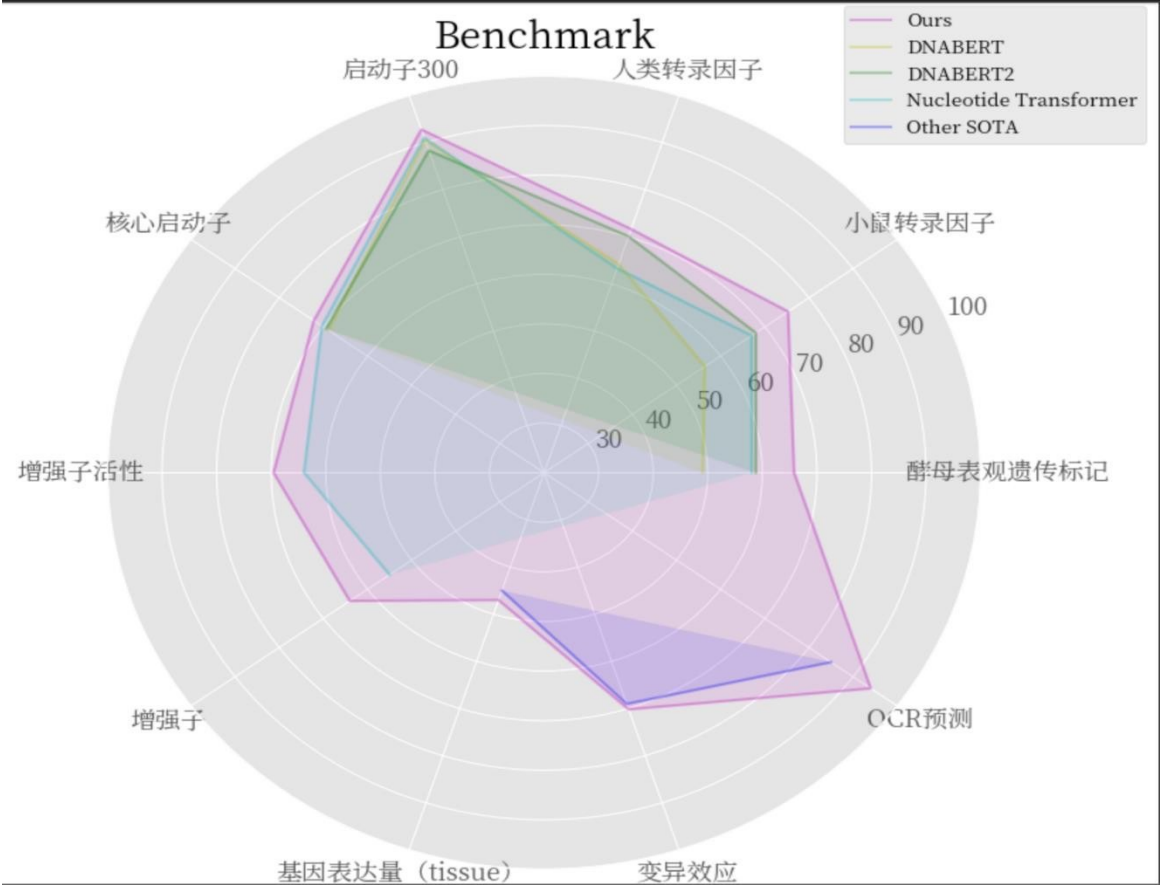
.....

AI + 基因组学：通用DNA大模型



task	酵母表观遗传标记	小鼠转录因子	启动子300
metric	mcc	mcc	mcc
sota	61.007	67.996	91.01
method	HyenaDNA	DNABERT2	NT-multi
our	65.826	75.278	92.74

task	增强子	变异效应	OCR预测
metric	mcc	F1	AUC
sota	0.593	0.6903	0.85
method	NT	NT-1000g	HyenaDNA
our	0.64	0.7018	0.94

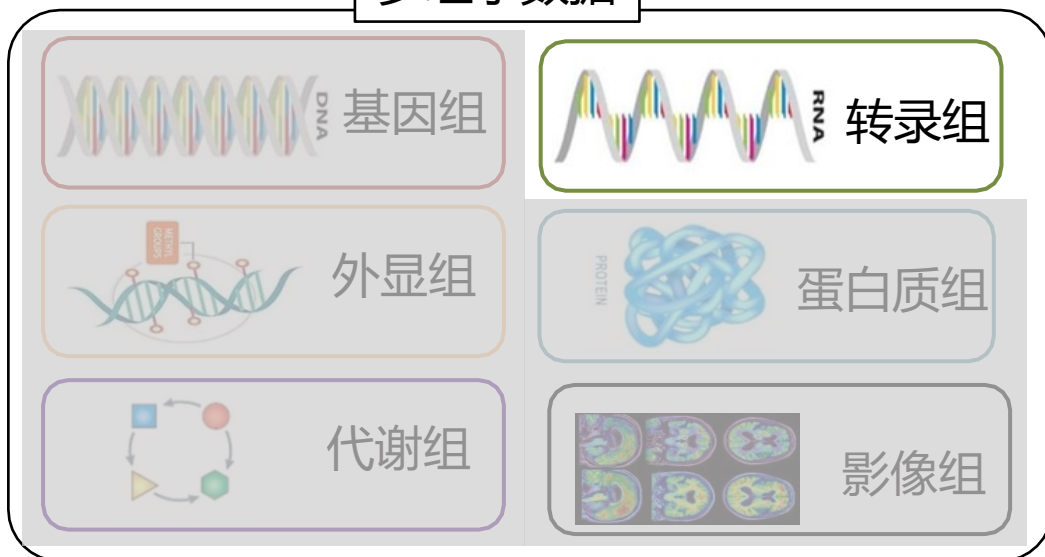


英伟达Nucleotide Transformer: 25亿模型参数, 10亿数据
复旦大学DNA大模型: 8千万参数, 5千万数据



应用：疾病诊断、病程预测、生存分析、治疗作用分析、风险预测 ...

多组学数据



AI模型



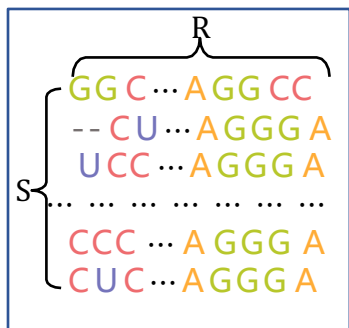
- 预测是否患病
- 患病规律预测
- 预测死亡风险
- 预测药物作用
- 预测患病风险
- ...

基于多组学数据的疾病分析

AI + 转录组学：RNA自监督学习方法

□ 问题描述

RNA多重序列



自监督方法

预训练
模型

下游任务

RNA二维结构 (二级结构) 预测
RNA溶剂可及表面积 (可及性) 预测
RNA 功能预测
...

标注少?

数据即标注

□ 问题必要性

- 绝大多数RNA功能未知，生物“暗物质”；
- 已知结构的RNA数目不足6千；
- 已知RNA序列数目超过2亿条；

□ 初步结果

- 在RNA二级结构预测任务上，相比SPOT-RNA2¹提升8%；
- 在RNA溶剂可及性任务上，相比RNASnap2²提升5.8%；

1. Singh, J. et al. Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. Bioinformatics, 2021.
2. Hanumanthappa, A. K. et al. Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network. Bioinformatics, 2020.

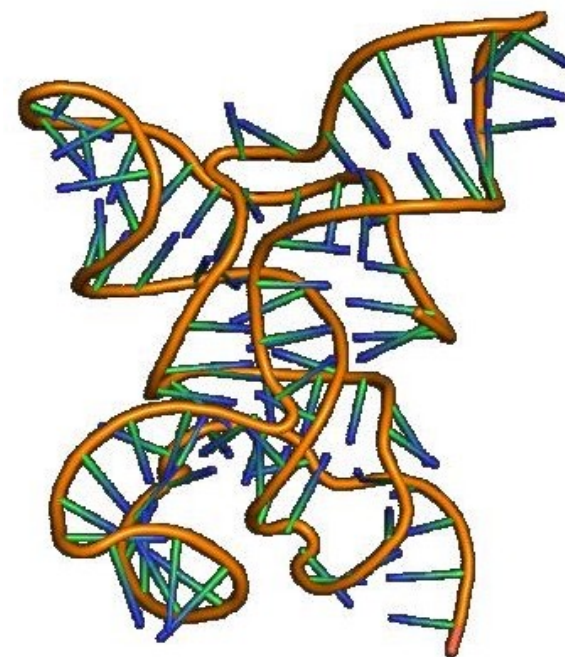
- 利用RNA语言模型和全新的结构预测模块预测三维结构，从而提高RNA三维结构样本利用率

GGAUCACGAGG
GGGAGACCCCG
GCAACCUGGGA
CGGACACCCAA
GGUGCUCACAC
CGGAGACGGUG
GAUCCGGCCCG
AGAGGGCAACG
AAGUCCGU

Puzzle 8
PDB ID: 4I81



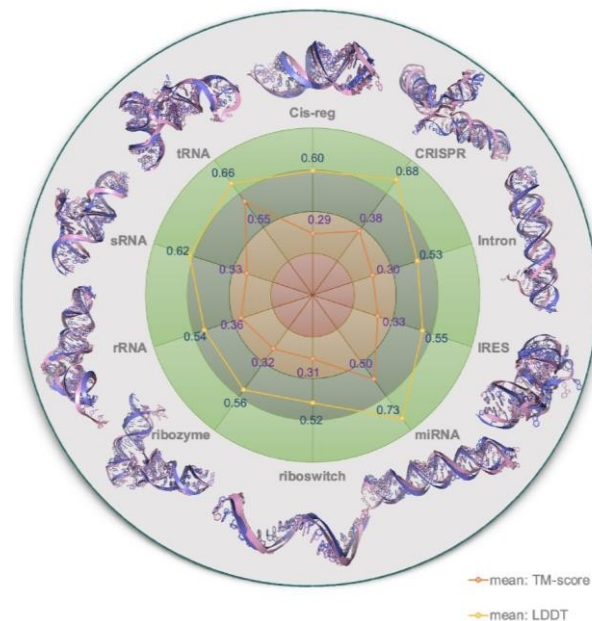
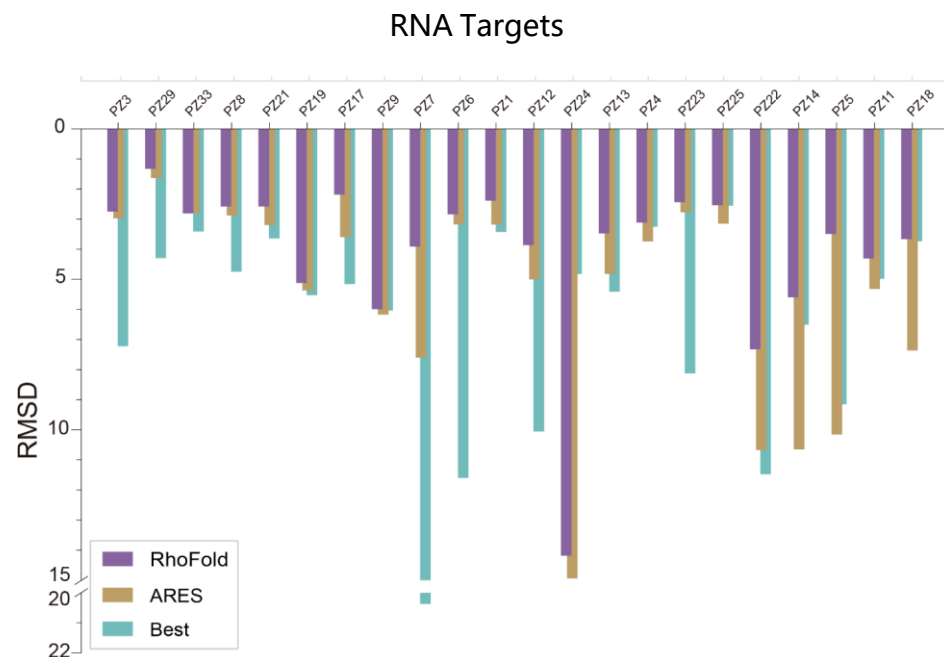
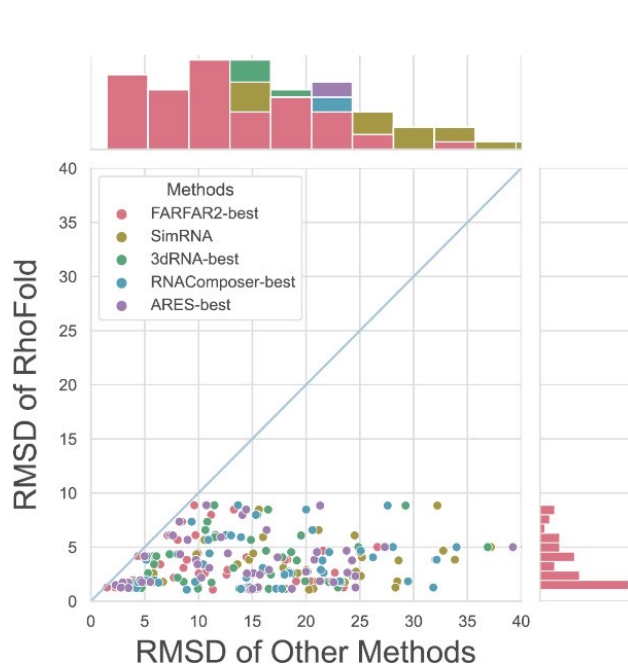
AI模型



AI + 转录组学：RNA三维结构预测



- 利用RNA语言模型和全新的结构预测模块，显著提高RNA三维结构样本利用率
- 首个端到端的RNA三维结构预测模型，误差 (RMSD) 从 $\sim 17.3 \text{ \AA}$ 改进至 $\sim 3.5 \text{ \AA}$
- 在RNA Puzzle上显著提高最优模型 (state-of-the-art) 的精度
- 模型具有跨RNA种类的预测能力 (结果采用留一交叉验证取得)



AI + 合成生物学

● AI驱动从头蛋白质设计

- AlphaFold: 覆盖98.5%人类蛋白质组, 原子级精度结构预测
- RFdiffusion (Nature 2023): 扩散模型实现从头蛋白质全新设计
- EVOLVEpro (Science 2025): 蛋白质语言模型驱动酶与抗体定向进化

● AI转型生物制造

- Evo (Science 2024): 全基因组设计, 生成全新功能DNA序列
- AI精准设计合成路径, 构建高产量微生物工厂



AI + 医学大语言模型

- **医学LLM突破 (2023-2024)**

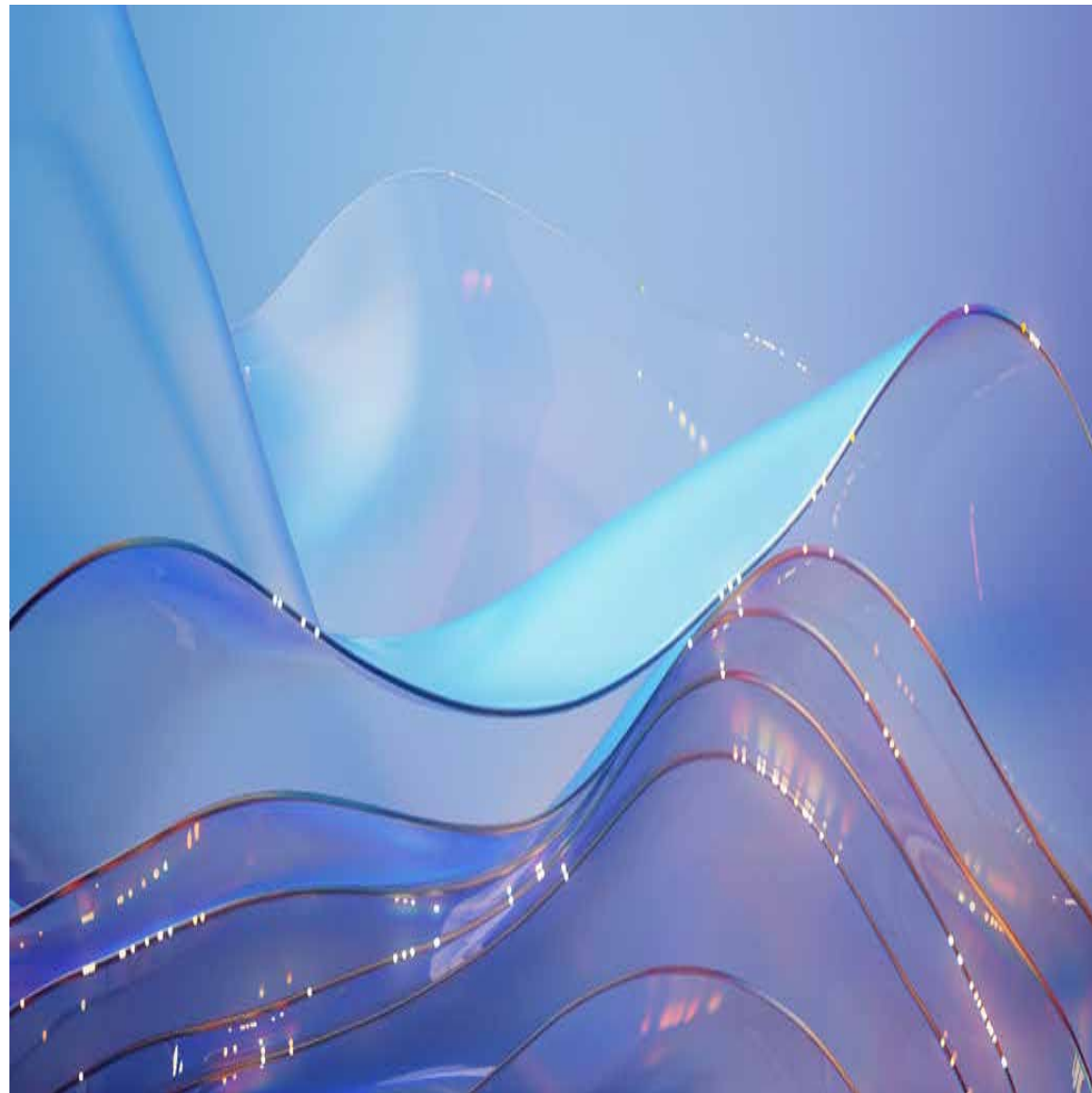
- Google Med-PaLM2: 通过美国医学执照考试, 达到专家级水平
- Microsoft BioGPT: 生物医学文本生成与关系提取

- **医学多模态大模型 (MLLM)**

- 融合影像、基因组和临床数据, 覆盖筛查→诊断→个性化治疗
- 北大MuMo: 多模态预测HER2阳性胃癌免疫治疗反应
- MedFound-DX-PA: 零样本学习疾病诊断准确率80.7%

- **关键挑战**

- 黑盒本质: 高风险场景需可解释证据链
- 伦理隐私: 患者数据使用、误诊责任、医疗公平性



目录

1 AI For Science介绍

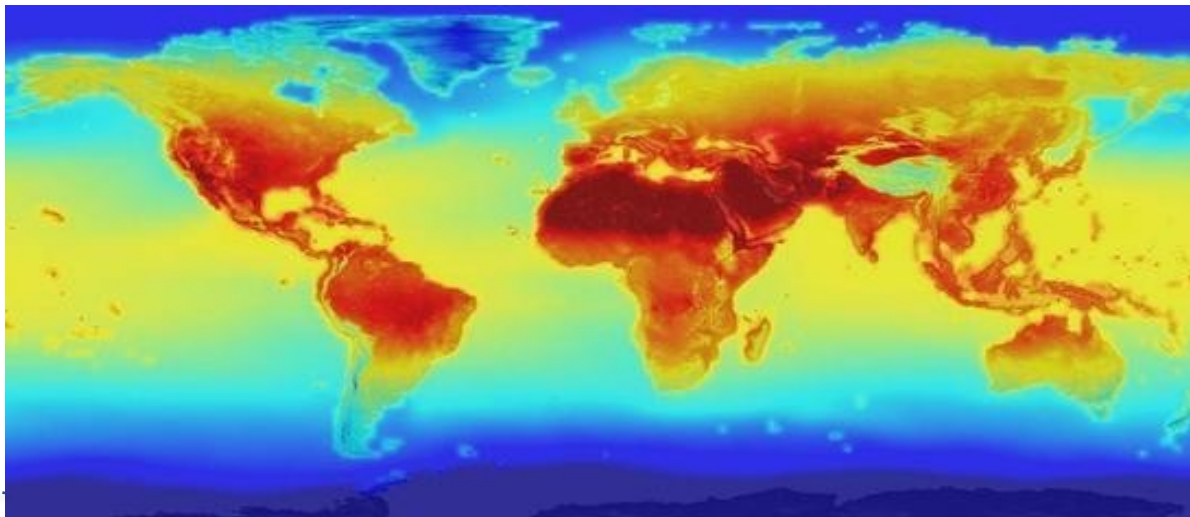
2 AI应用于微观科学

3 **AI应用于气象学**

4 AI应用于生物医学

5 AI应用于其他领域

地球级 ($10^5 \sim 10^{10}$)



大气系统和海洋环流相互作用，深刻地影响地球的气象、气候和生态系统，建模和分析大气与海洋的各种活动，预测未来的大气与海洋变化是地球科学的核心任务之一

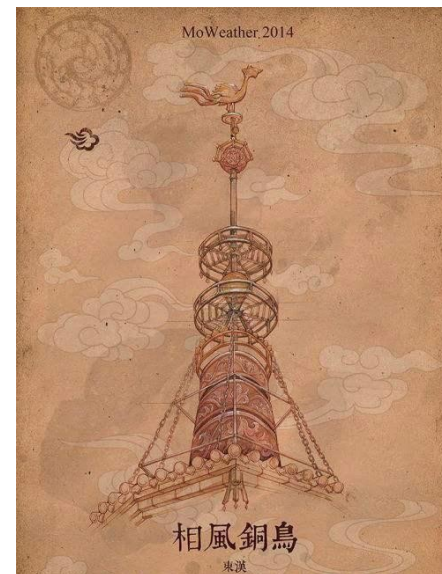
中国古代观天象的历史

通过观察天象预测气候与天气变化
历史悠久，具有丰富经验

现代气象预报与古代观天象的联系

目的相同：预测未来天气状况

方法不同：现代气象学基于科学定律和数值模型



相风铜鸟



诸葛亮



借东风

Abbe和Bjerknes提出物理定律可用于气象预报

当时面临的挑战

- 大气观测数据稀缺

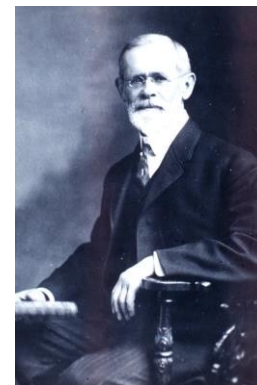
- 无计算机辅助

- 对天气可预测性了解有限

现今气象预报

- 超过100年的发展

- 每十年预测能力增长一天



Cleveland Abbe
(美国)



Vilhelm Bjerknes
(挪威)

发展趋势



国内

- 2016年，**清华大学**成立地球系统科学系
- 2018年，**复旦大学**成立大气与海洋科学系
- 2021年，**中国气象局**成立地球系统数值预报中心
- 2022年，中国基本建成12.5公里分辨率全球模式
- 2022年，**香港科技大学**成立地球与海洋大气科学枢纽



国外



欧洲中期天气预报中心



阿兰图灵研究中心



DeepMind



英伟达



微软



劳伦斯伯克利
国家实验室



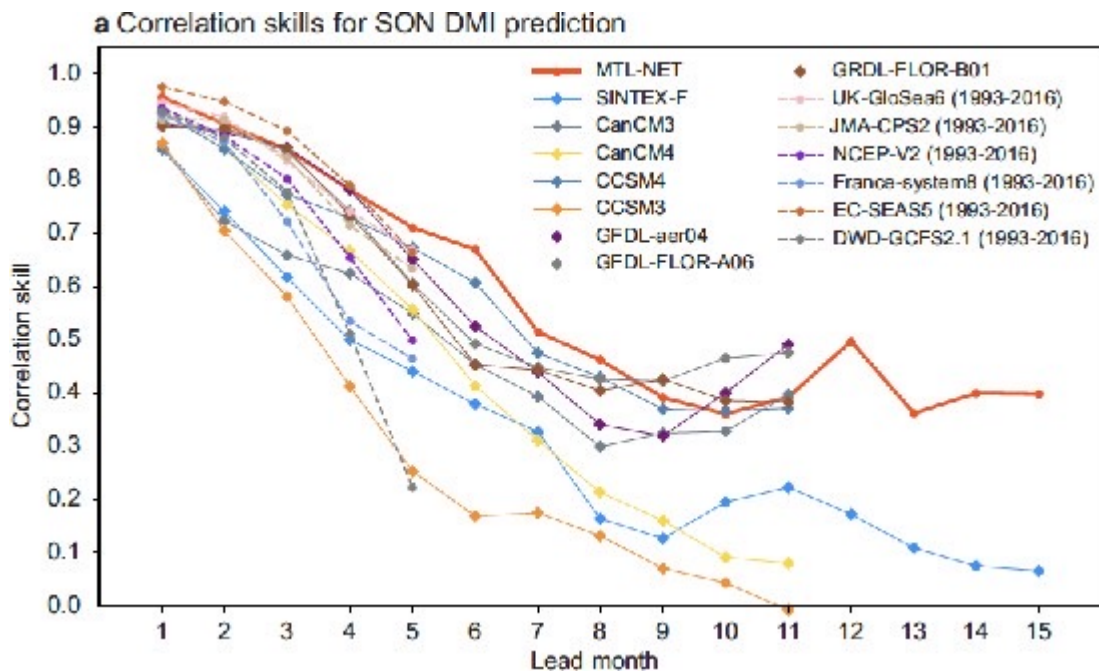
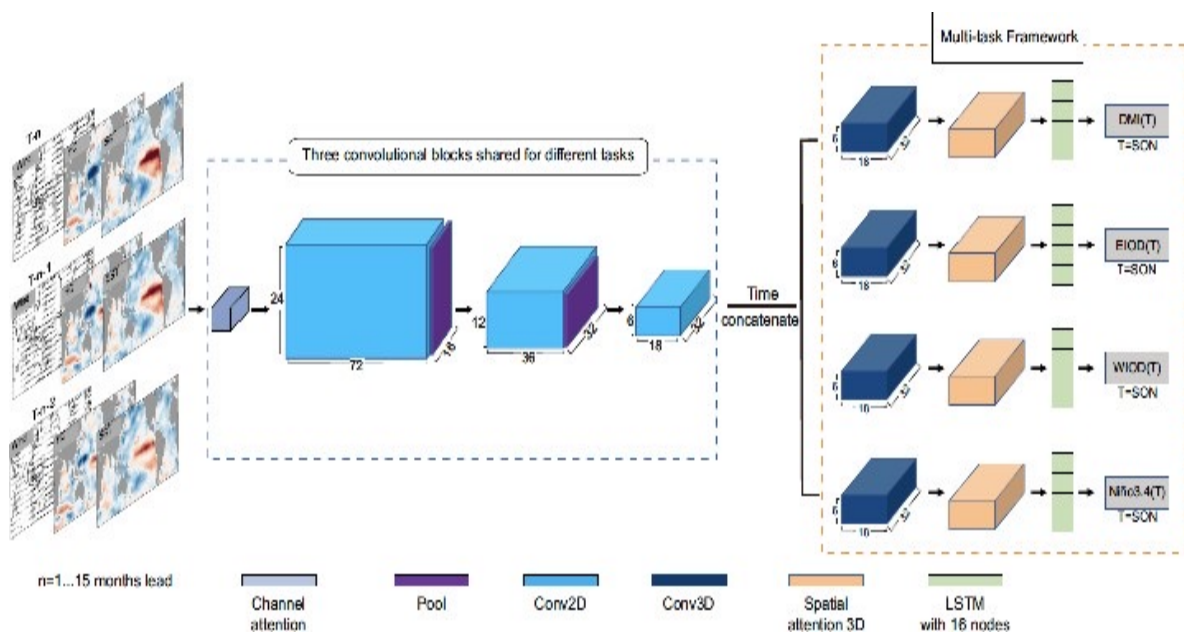
基于多任务学习的印度洋偶极子预测 (Nature Communications)

印度洋偶极子是影响我国夏季气象的重要气候现象之一

提出了一种基于多任务学习的大气-海洋耦合模型来同时建模印度洋和太平洋海温

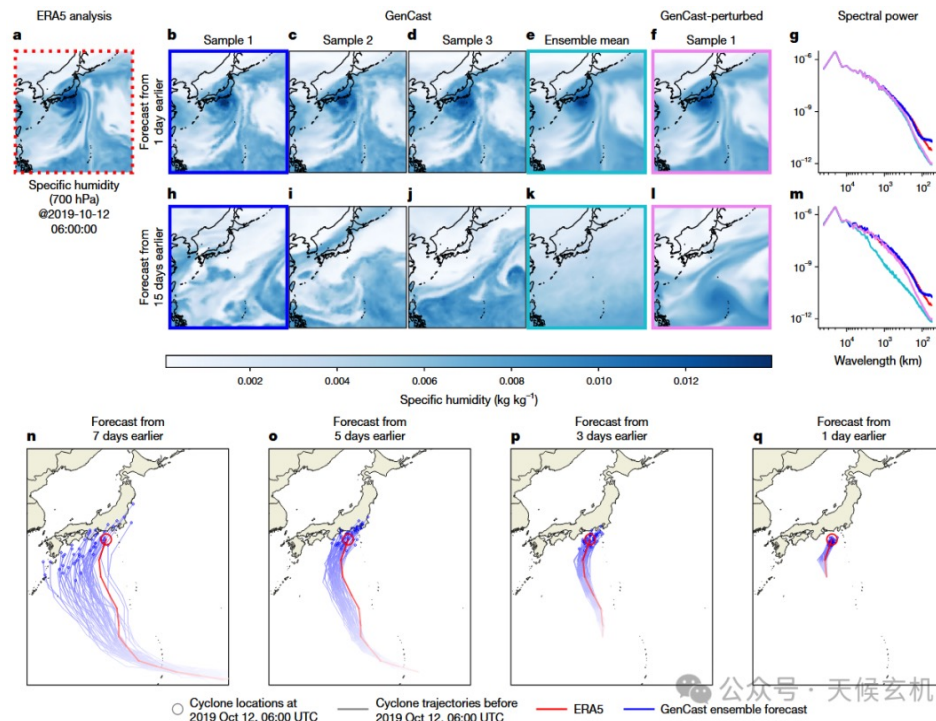
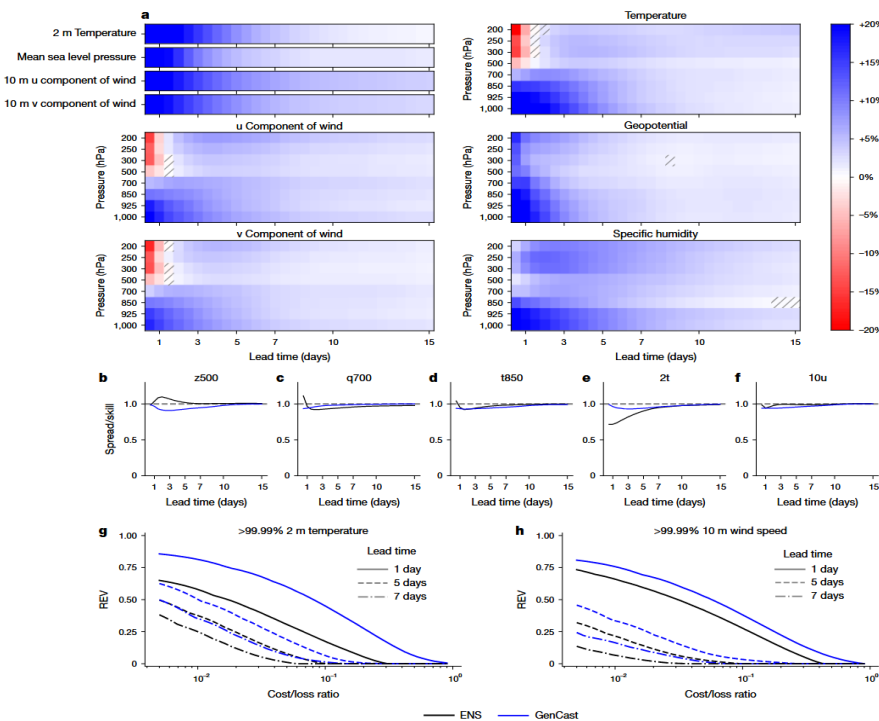
首次印度洋核心气候指标的可用预报技巧提高到**7**个月

通过注意力机制识别出不同预测因子、不同区域的重要性



基于机器学习的概率天气模型 (Nature)

- GenCast概率天气模型精度和速度超欧洲中期天气预报中心 (ECMWF) 的集合预报系统 (ENS)
GenCast 用再分析数据训练, 可生成 15 天全球集合预报, 分辨率 0.25° , 8 分钟完成
GenCast能更好预测极端天气、热带气旋路径和风电生产, 开启天气预报新篇章



Price I. et al. Probabilistic weather forecasting with machine learning. Nature, 2025.

AI + 全球中期气象预报（风乌大模型）



相通问题：疾病走势预测

数据即标注

任务描述：预测未来14天全球全要素大气场

分辨率：0.25经纬度

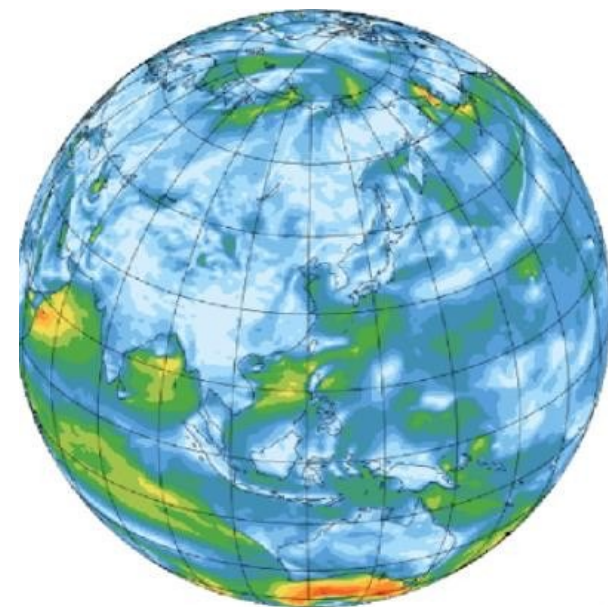
区域：全球

垂直层次：37层

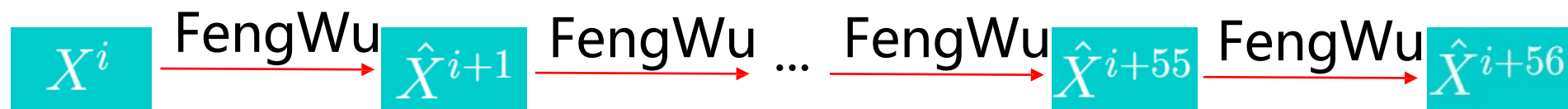
气象要素：位势、温度、湿度、经向风速、纬向风速

问题建模：

得到一个函数（风乌），输入当前时刻大气场，输出未来14天大气场
与物理模型一样，风乌训练的目标是预测下一时刻大气场，然后通过
自回归方式得到所有预测



$$\hat{X}^{i+1} = \text{FengWu}(X^i), \hat{X}^{i+2} = \text{FengWu}(\hat{X}^{i+1}), \dots, \hat{X}^{i+56} = \text{FengWu}(\hat{X}^{i+55})$$

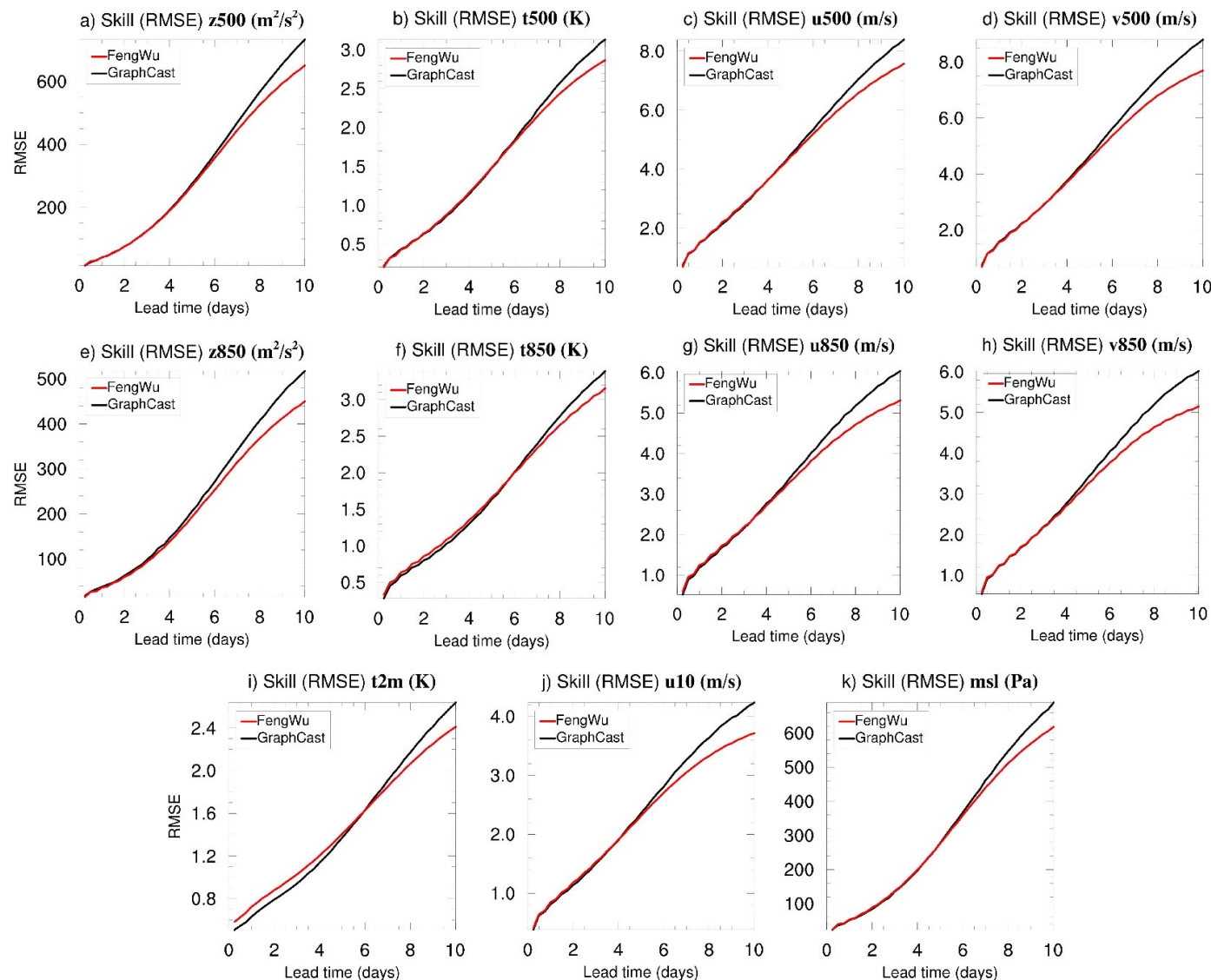


AI + 全球中期气象预报 (风乌大模型)



相比Deepmind于2022年12月发布的GraphCast模型，风乌在80%的指标上获得了更好的性能，尤其是5-10天的预测，风乌具有显著优势（RMSE越小越好）。

十天预报误差相比GraphCast降低10.87%，相比欧洲中期气象预报中心的IFS-HRES降低19.4%。

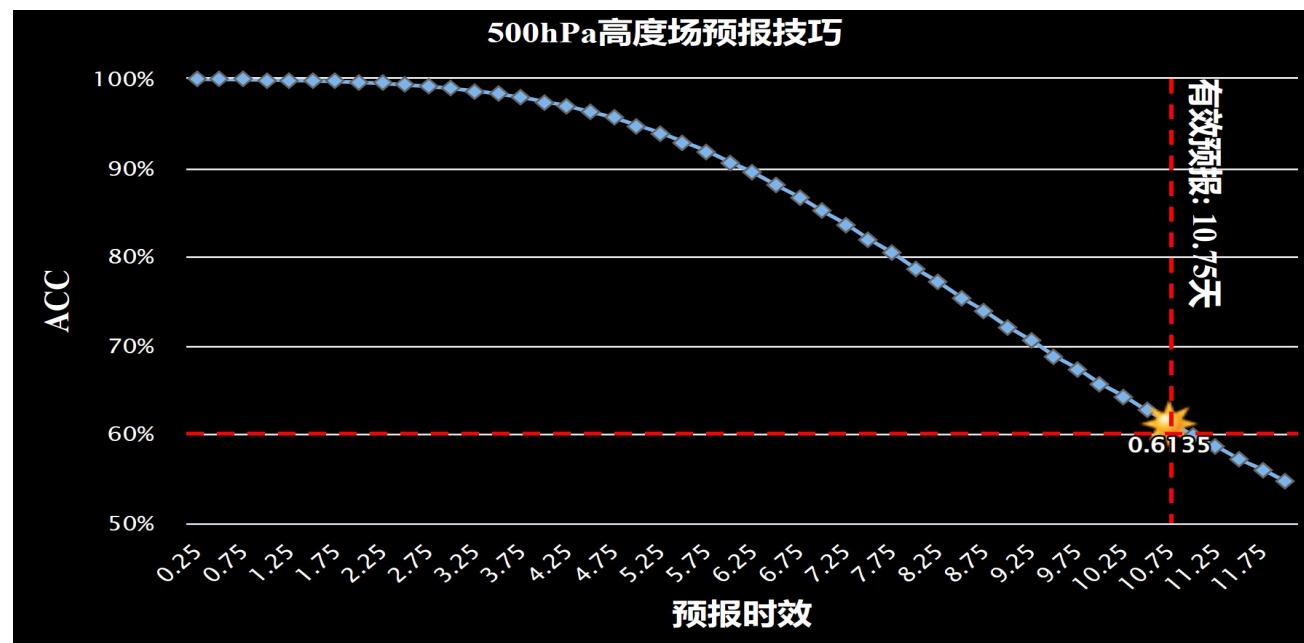


AI + 全球中期气象预报（风乌大模型）



AI的高速迭代：基于2018年回报表明，“风乌”超过了Deepmind、英伟达等公司发布的AI气象大模型，实现了最精准的全球高分辨率气象预报，并突破性的将全球中期气象预报的可用性预报时效提高到**10.75天**（相比GraphCast仅过了4个月便提高1.25天）。

AI的高效性：**一张GPU**就可以在一**分钟**内生成未来14天全球所有地区高精度高分辨率气象预报结果。



- 发布一周内，“风乌”引起了国内外的反响。在国内，新华社、人民网、央广网、文汇报、量子位、机器之心等23家新媒体和近百家新媒体进行了报道，其中**新华社**客户端报道浏览量超**140万次**，新华社公众号推文阅读10万+，点赞1500余次。在国外，**英国气象局**、**ECMWF**、Google、Deepmind、亚马逊等著名研究机构第一时间联系寻求合作或分享。



people.com.cn

<http://sh.people.com.cn> › 上海频道

上海人工智能实验室发布“风乌”大模型，全球气象有效预报时间 ...

2023年4月7日 — 基于多模态和多任务深度学习方法构建，AI大模型“风乌”首次实现在高分辨率上对核心大气变量进行超过10天的有效预报，并在80%的评估指标上超越DeepMind ...



cnr.cn

<https://www.cnr.cn> › shanghai

上海人工智能实验室发布“风乌”大模型 - 央广网

2023年4月8日 — 基于多模态和多任务深度学习方法构建，AI大模型“风乌”首次实现在高分辨率上对核心大气变量进行超过10天的有效预报，并在80%的评估指标上超越DeepMind发布 ...



huanqiu.com

<https://tech.huanqiu.com> › article

“风乌”来了！30秒可生成未来10天全球高精度天气预报 - 环球网

2023年4月9日 — 此外，“风乌”仅需30秒即可生成未来10天全球高精度预报结果，在效率上大幅优于传统模型。实验室领军科学家欧阳万里表示，“风乌”取名自秦汉时期的相风铜 ...



yicai.com

<https://m.yicai.com> › news

高分辨率全球气象预报大模型“风乌”发布有效预报 ... - 第一财经

“风乌”在80%的评估指标上超越谷歌旗下DeepMind发布的气象大模型GraphCast。“风乌”仅需30秒即可生成未来10天全球高精度预报结果，在效率上大幅优于传统模型。

大模型“风乌”实现全球气象有效预报超10天

2023-04-08 20:31:13

来源：新华社

浏览量：140.8万



新华视点

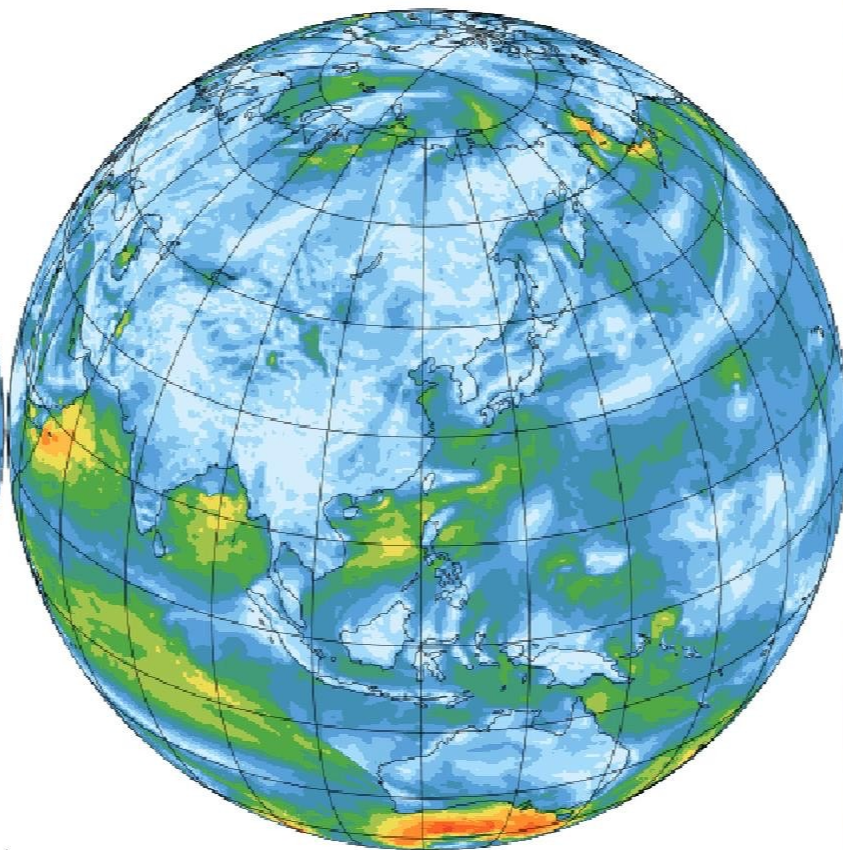
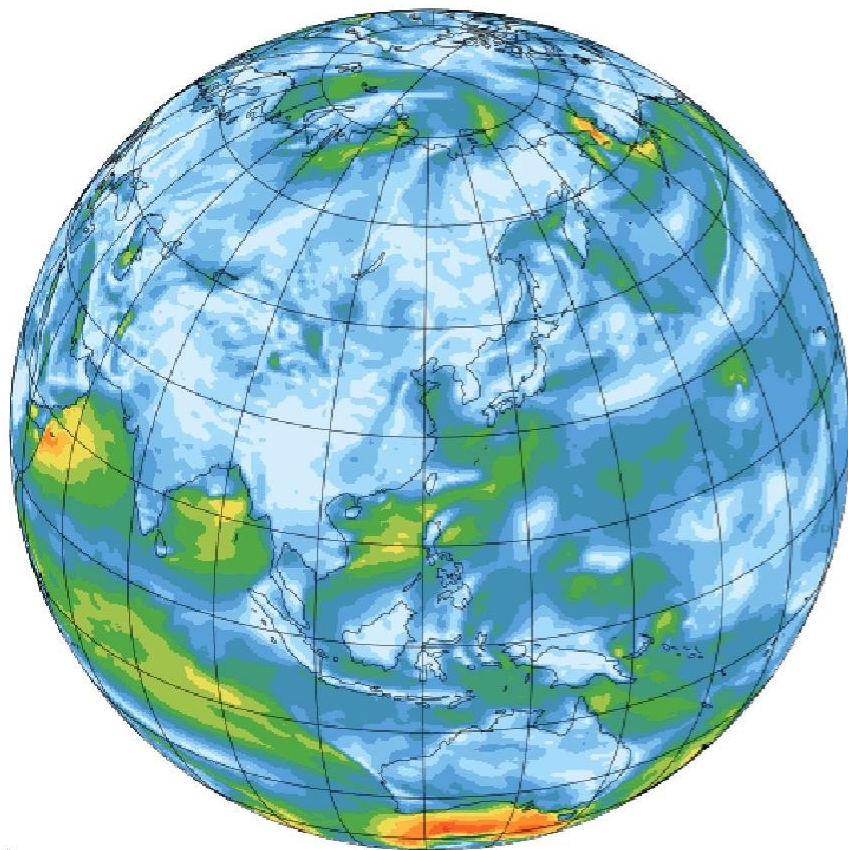
[查看详情 >](#)

AI + 全球中期气象预报 (风乌大模型)

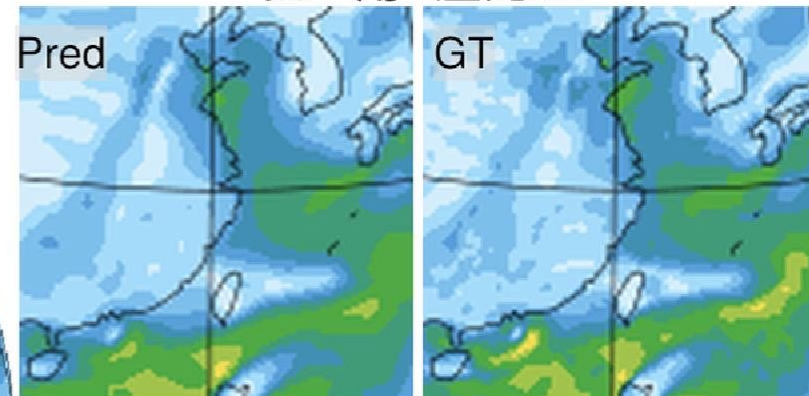


Pred: 2018-08-14T06:00:00

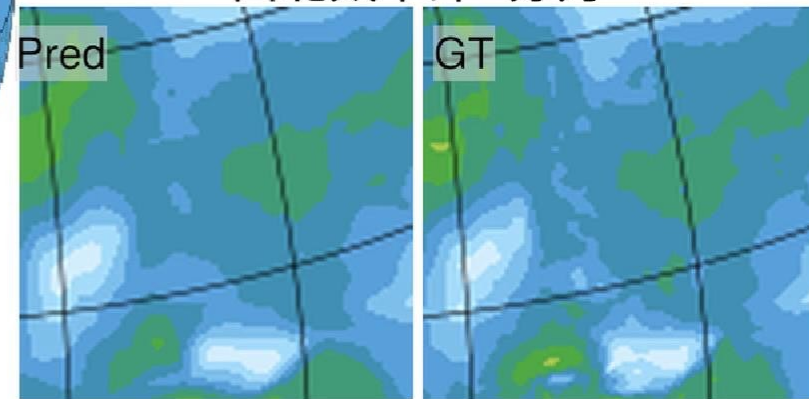
GT: 2018-08-14T06:00:00



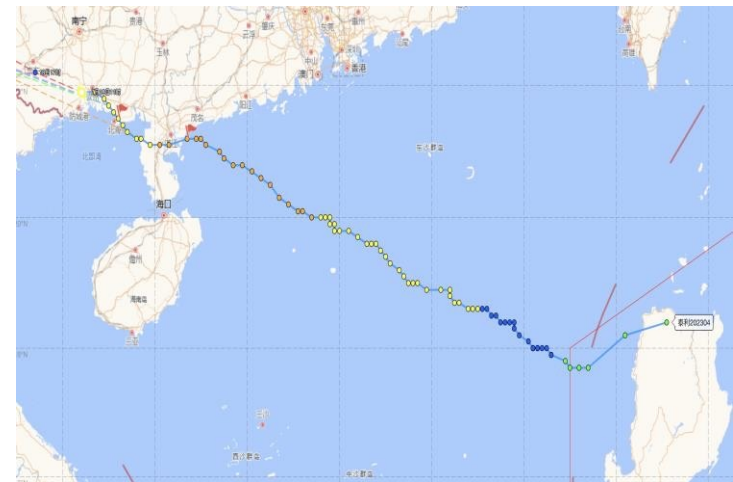
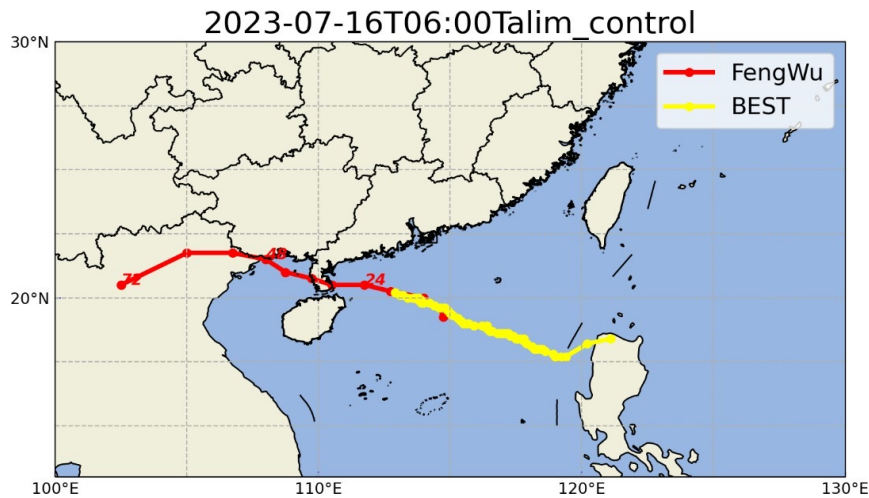
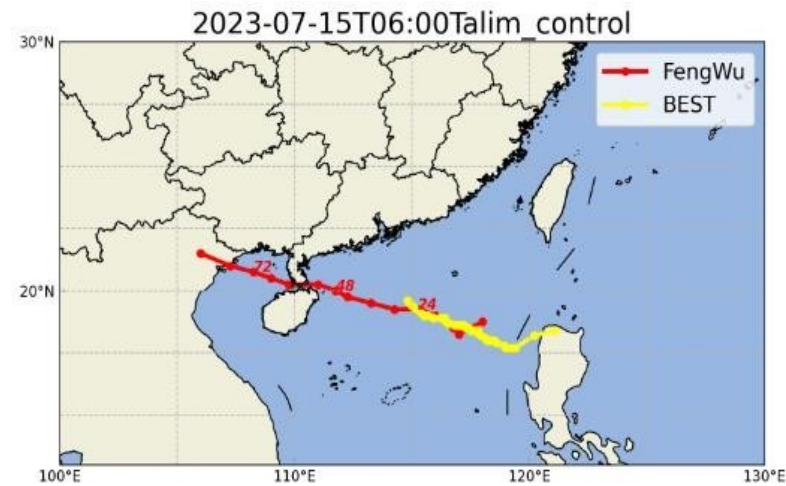
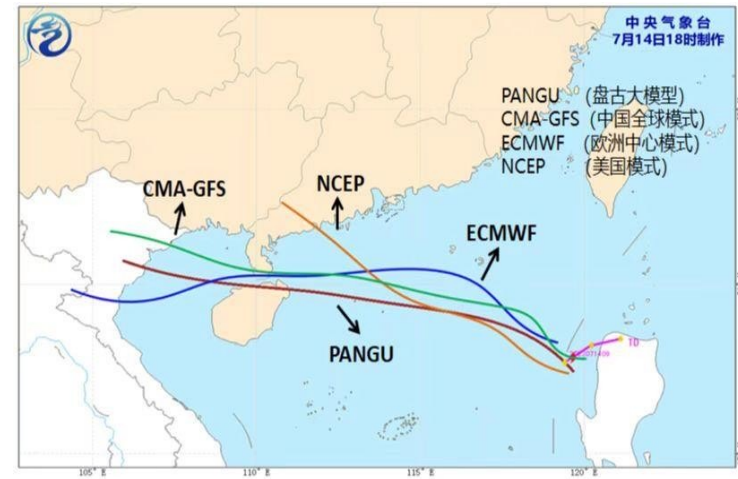
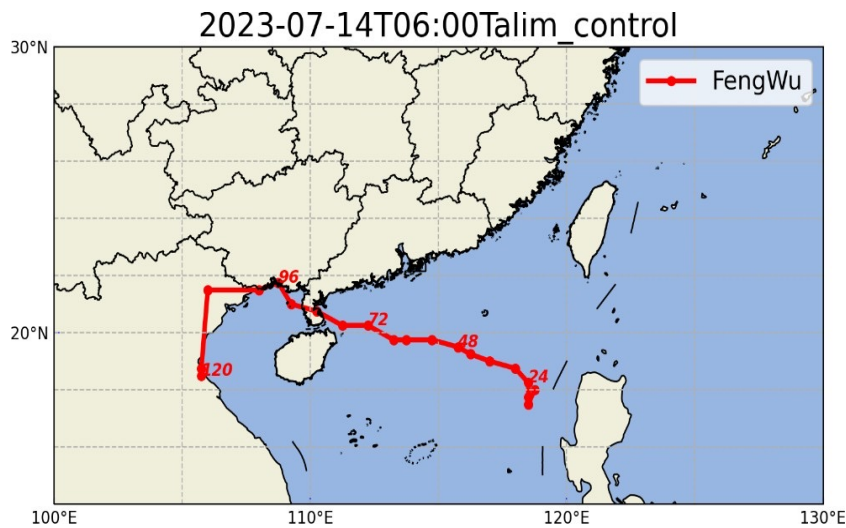
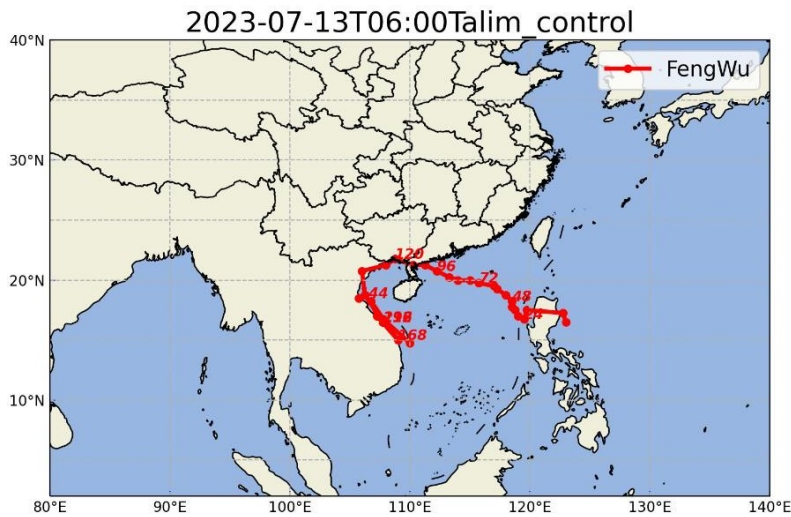
长三角-温比亚



西北太平洋-苏力



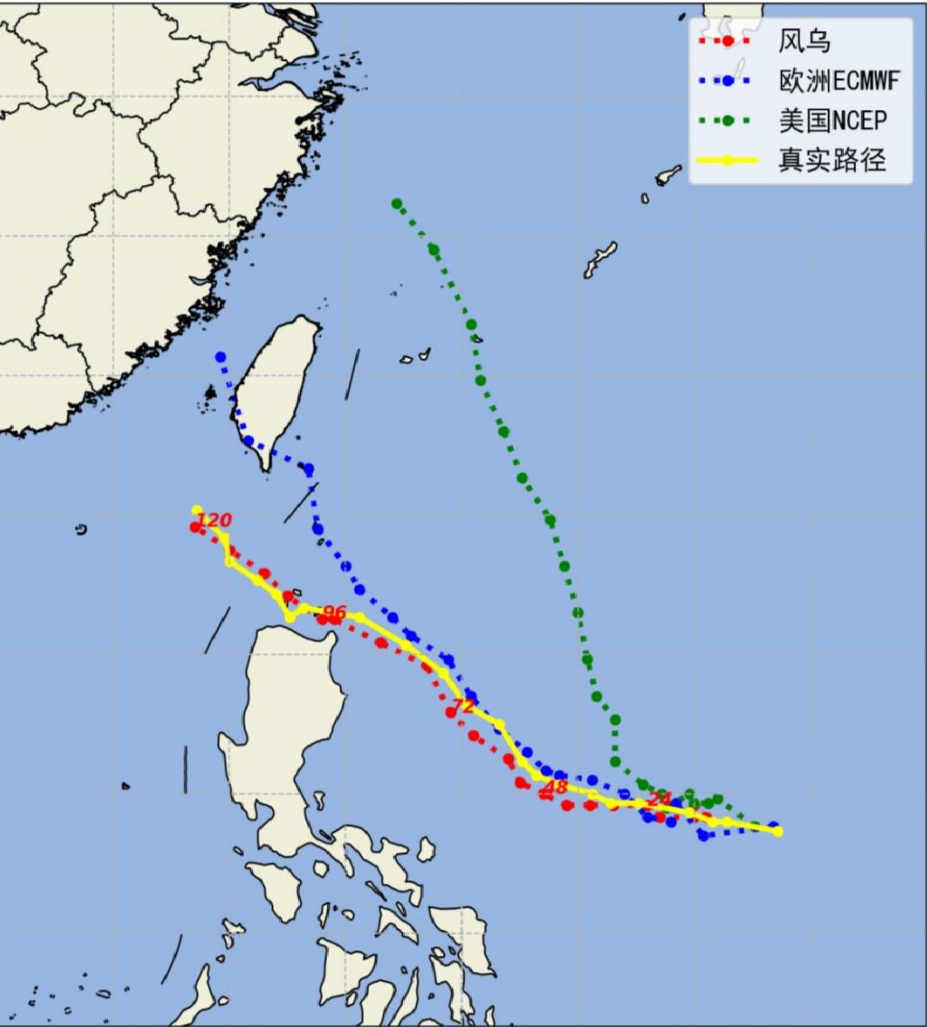
风乌可提前12小时预测台风泰利轨迹



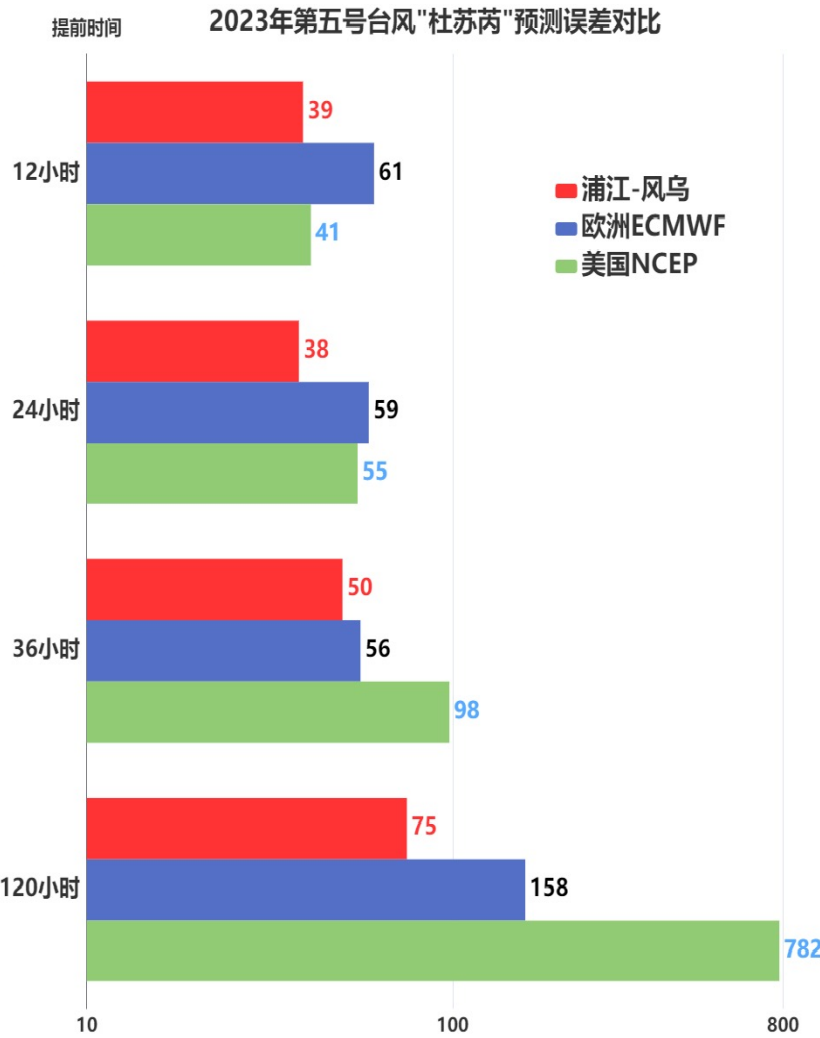
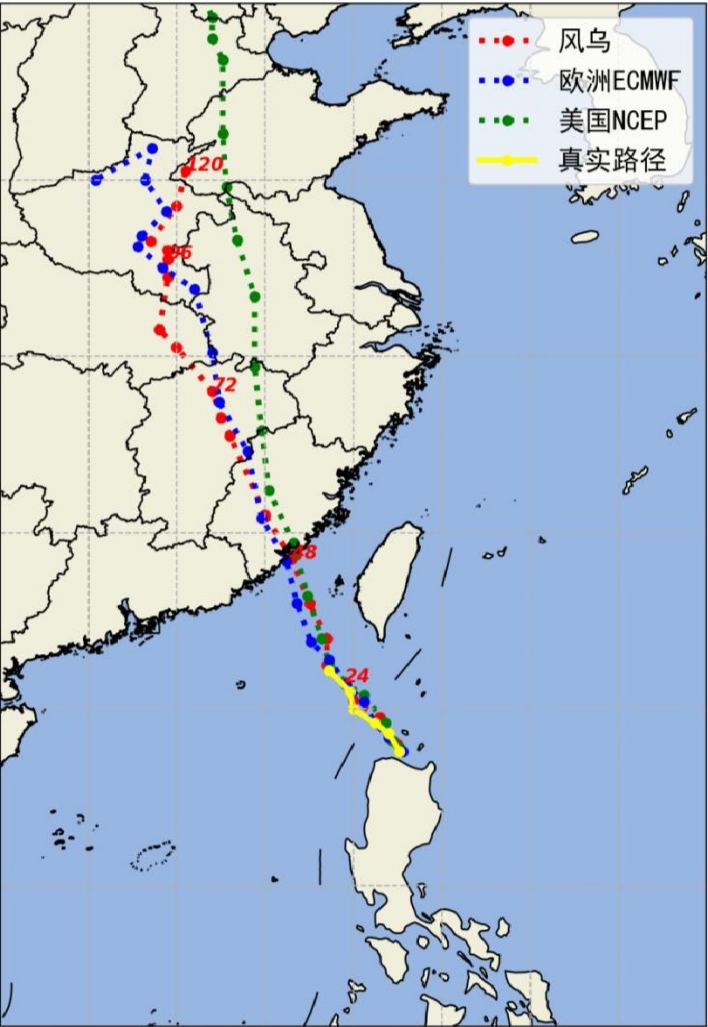
风乌可提前24小时预测台风杜苏芮



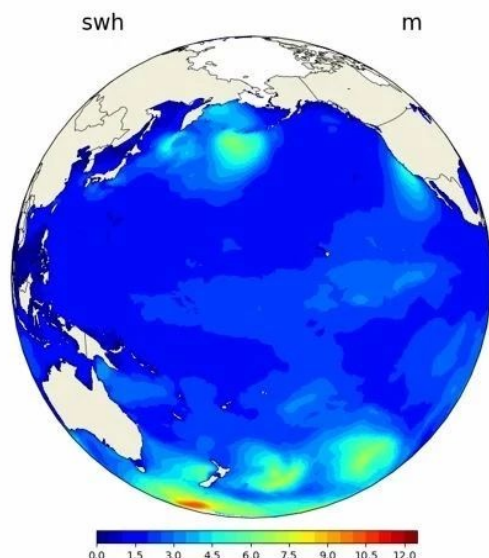
杜苏芮台风：2023年07月22日08点起报



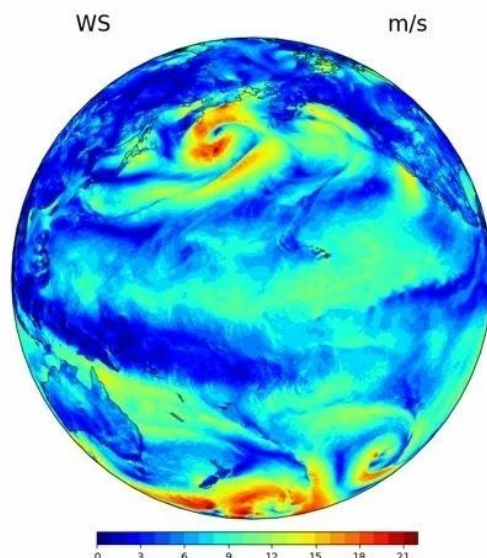
杜苏芮台风：2023年07月26日08点起报



- 2024年6月3日，复旦大学和上海科学智能研究院联合发布全新升级的伏羲系列气象大模型2.0，成立以“伏羲”2.0为核心的智能气象创新生态联盟。“伏羲”2.0在中期天气预报和次季节预测方面均有突破，且面向新能源、航空运输等行业展现出强大的应用潜力。



全球有效浪高预测



全球10m风速预测

- 目前全球所使用的AI气象大模型主要有华为云的盘古气象大模型、复旦大学和上海人工智能实验室联合开发的伏羲、上海人工智能实验室开发的风乌、阿里巴巴的SwinVRNN、谷歌Deep Mind开发的GraphCast、清华大学和中国气象局开发的NowcastNet、英伟达Nvidia开发的FourCastNet、微软、华盛顿大学开发的DLWPI以及ECMWF开发的CNN等。

- 2023年9月1日上午，复旦大学2023级新生开学典礼在光华楼前大草坪举行，这是学校近20年来第一次在户外、本研一起举行开学典礼。



- “伏羲”说：“今天，上海不下雨。”
- 户外典礼背后的高科技：AI4S精准知天气。

AI气象预报：关键挑战与未来方向

- **极端天气预报不足**

- MSE损失函数使模型预测模糊；热带气旋强度被系统低估
- 对策：概率生成模型 + RAG历史天气检索

- **过度依赖ERA5再分析数据**

- ERA5降水估算偏差；台风强度偏弱
- 对策：NWP合成高分辨率数据；HR-Extreme数据集 (3km)

- **缺乏物理约束**

- 纯数据驱动可产生负湿度等非物理输出
- NeuralGCM (Nature 2024)：可微分动力学核心+深度学习

- **未来：地球系统建模**

- 耦合海洋-陆地-冰冻圈，捕捉厄尔尼诺等跨系统现象



目录

1 AI For Science介绍

2 AI应用于微观科学

3 AI应用于气象学

4 **AI应用于生物医学**

5 AI应用于其他领域

AI+蛋白质结构：什么是蛋白质？



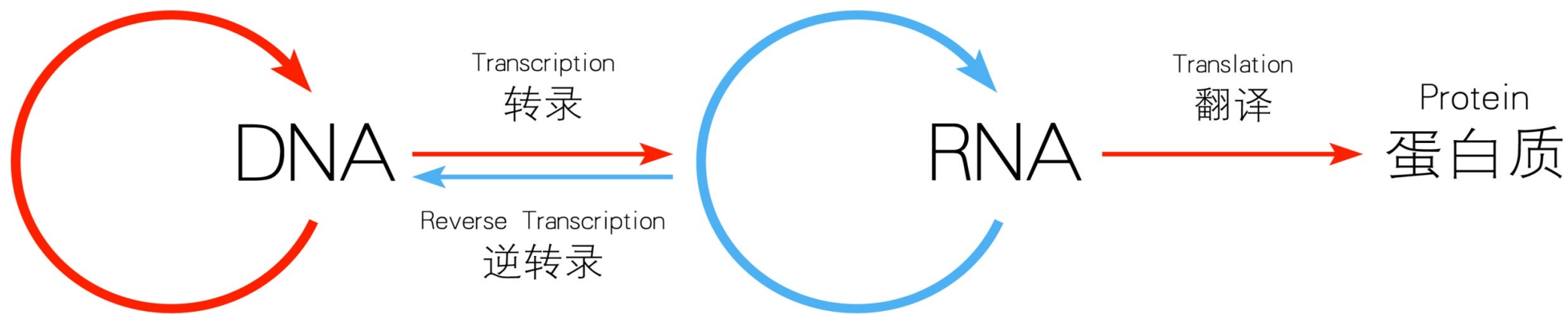
氨基酸是同时含有氨基和羧（suǒ）基的有机化合物，是组成蛋白质的基本单元。蛋白质氨基酸大约20种，结合在一起产生了大量的肽和蛋白质。

蛋白质是20种氨基酸不同组合链接而成的高分子链。

氨基酸的种类、数量、排列不同，肽链盘区折叠方式及所形成的蛋白质空间结构千变万化，共同导致了蛋白质种类的多样性。



蛋白质结构预测—中心法则



弗朗西斯·克里克于1958年提出中心法则

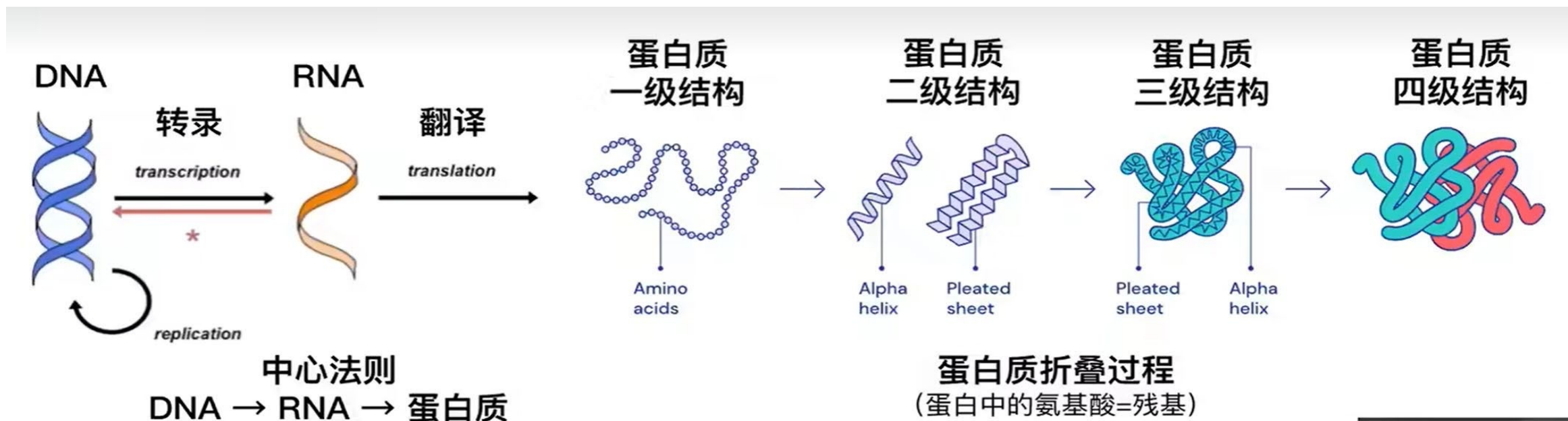
- **转录 (Transcription)** : 在细胞核中, DNA的一部分被复制成为RNA分子的过程
- **翻译 (Translation)** : 翻译是指mRNA分子通过核糖体 (ribosome) 的作用, 将其上的信息翻译成蛋白质的过程



AI+蛋白质结构：蛋白质结构决定功能



研究蛋白质结构，有助于了解蛋白质的作用，了解蛋白质如何行使其生物功能，认识蛋白质与蛋白质（或其它分子）之间的相互作用。例如：结构蛋白是细胞和生物体结构的重要物质，而功能蛋白与催化、运输、调节、免疫等有关。



安芬森法则（蛋白质折叠原理）

1. 蛋白质的三级结构是由其氨基酸序列所决定的；
2. 蛋白质折叠过程是一个自由能最小化的过程，即蛋白质会寻找能量最低的稳定状态；
3. 蛋白质会自发地折叠成其独特的三维结构。

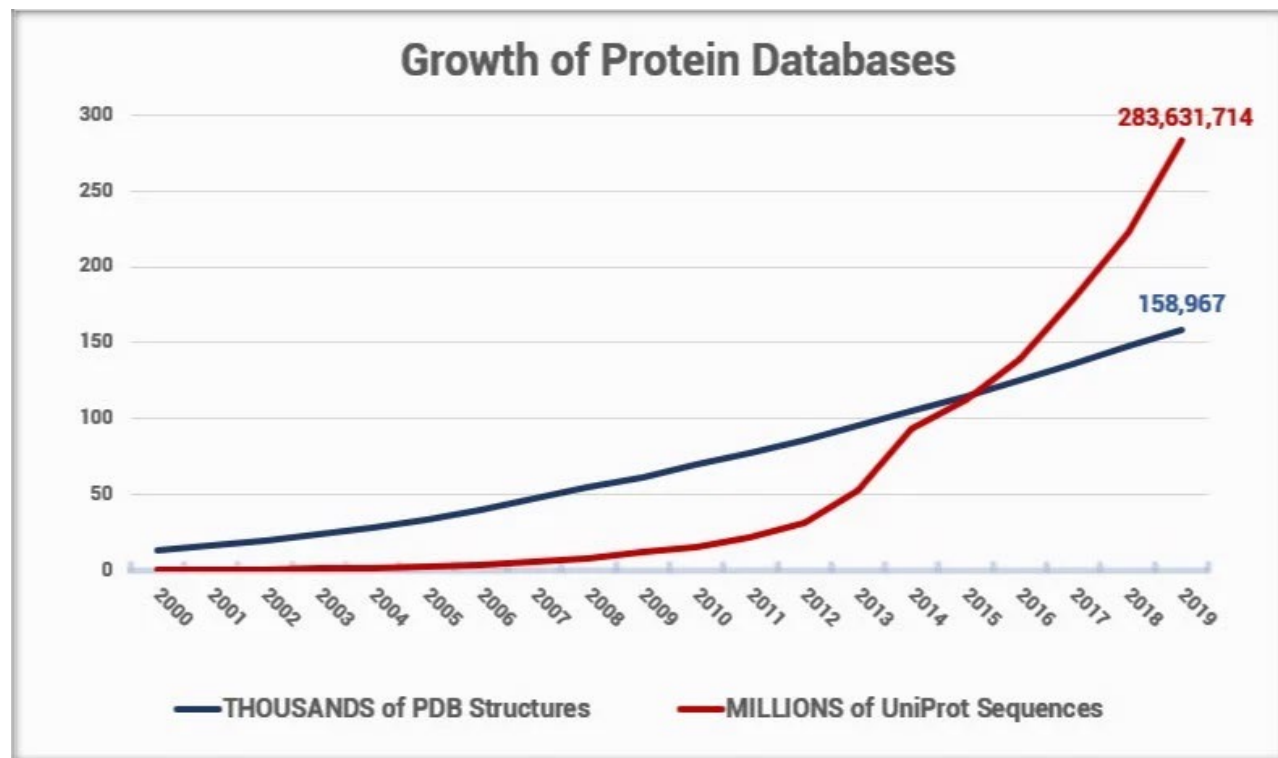


1972 Nobel Prize
in Chemistry

克里斯蒂安·B·安芬森

揭示了蛋白质折叠的基本原理，为蛋白质结构预测、蛋白质工程和药物设计提供了理论基础。

蛋白质结构预测



海量的序列信息，少量的结构信息，需要高精度高通量发现蛋白质结构的方法



CASP

Critical Assessment of protein Structure Prediction, 1994



评估计算机预测蛋白质结构方法的准确度



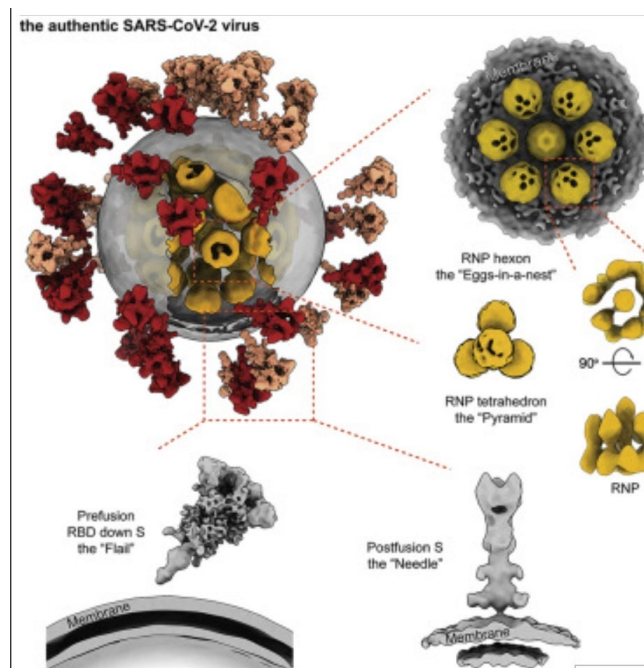
冷冻电子显微镜法 (Cryo-electron microscopy)



- 可以获得蛋白质的动态信息，如构象变化、与配体的相互作用等
- 适用于各种大小的蛋白质及蛋白质复合物研究，且具有较高分辨率，能够解析原子级别的细节
- 冷冻电镜技术在解析病毒结构、推测其侵染人体细胞的路径等传播原理发挥了重要作用



2017年10月4日，诺贝尔化学奖



国际首次解析新冠全病毒分子结构
2020年9月15发表于Cell



1. 同源建模法 (homology modeling)

- 相似的序列具有相似的蛋白质结构。根据相似序列结构构建最终模型

2. 穿线法 (threading)

- 不相似的氨基酸序列 ($< 30\%$) 也可以对应着相似的蛋白质结构。将目标序列放到现有的结构里, 找出能量最低的结构构建最终模型。

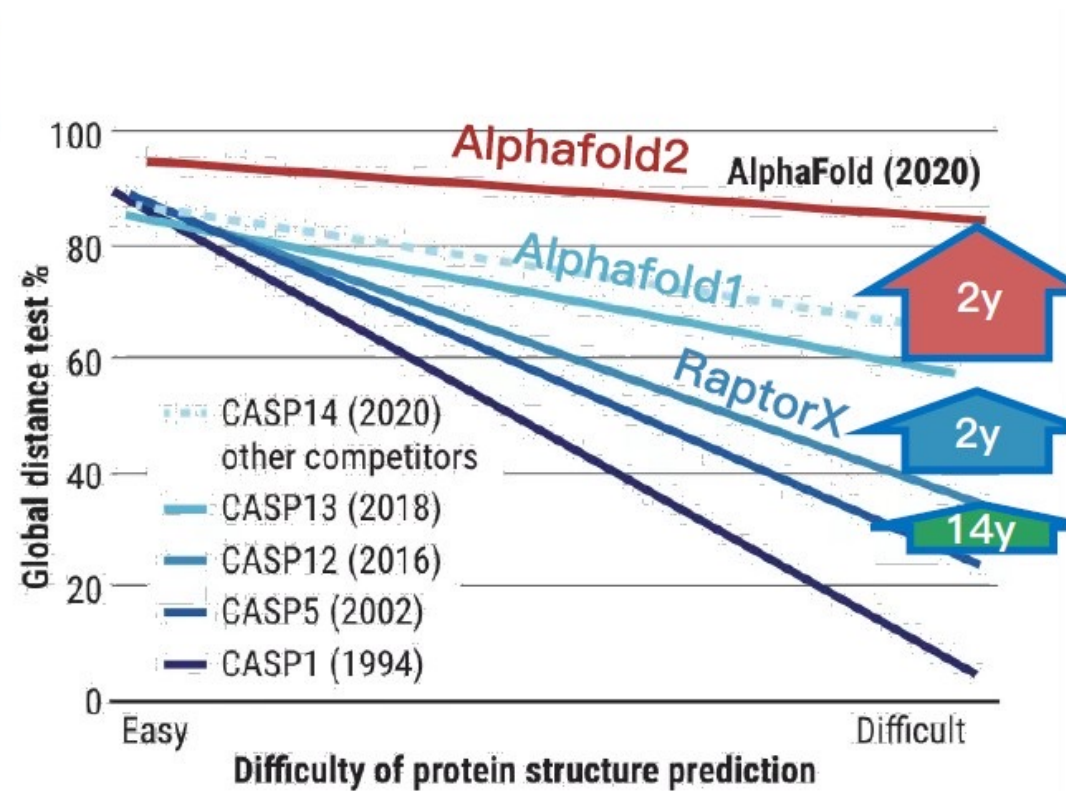
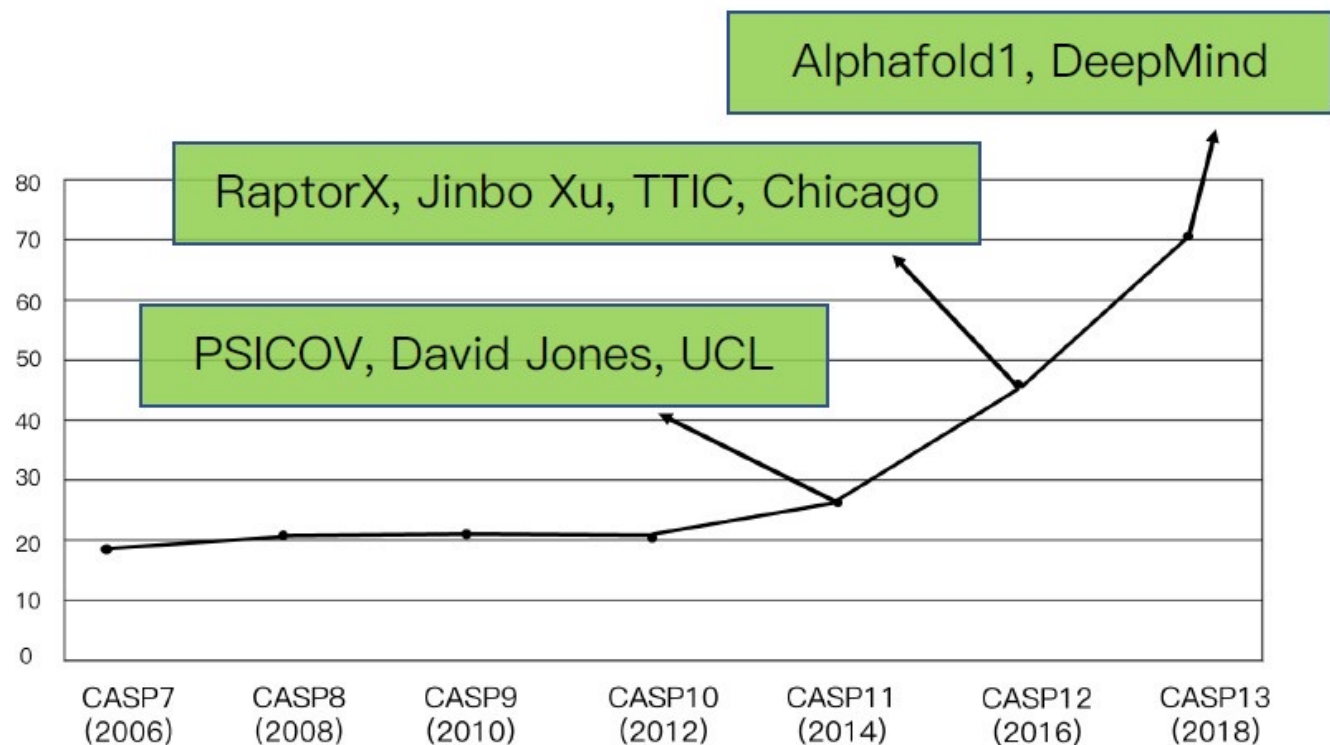
3. 从头计算法 (ab initio)

- 蛋白质的三维结构决定于自身氨基酸序列, 并且处于最低自由能状态。利用MSA获取序列特征, 根据溶剂可及性 (solvent accessibility)、二面角 (torsion angle)、距离特征 (distance profile) 等信息优化能量

4. 组合法 (ensemble method)

基于深度学习的蛋白质结构预测

• CASP比赛中的关键提升



- 2014 (MSA): PSICOV, David Jones, UCL
- 2016 (Deep Learning/ResNet): RaptorX-Contact, Jinbo Xu, TTIC, Chicago
- 2020 (Transformer): AlphaFold 2, DeepMind

AlphaFold 1 预测蛋白质的Contact Map 理论基础



蛋白质序列

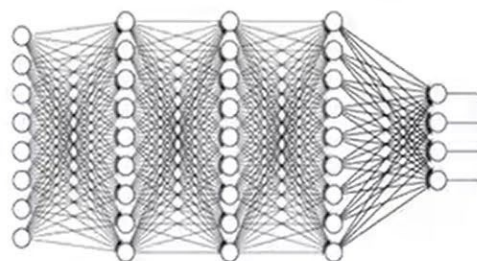


多序列比对



在序列数据库中找到与原序列接近的序列
按对应的氨基酸位点做比对(Alignment)

深度学习模型



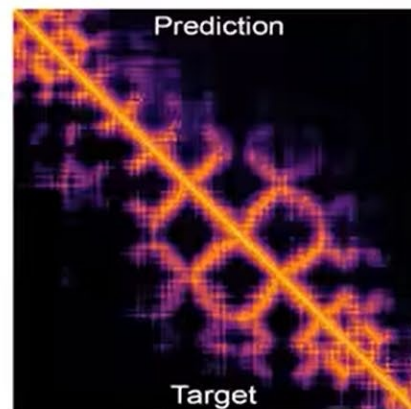
蛋白质结构



为什么选用这种方法：

- 直接预测蛋白质结构3D坐标比较困难
- 先预测蛋白质的Contact map，然后作为限制来优化蛋白质折叠，相对来说更简单

蛋白质结构的二维表示
Contact Map
(Distance Map)

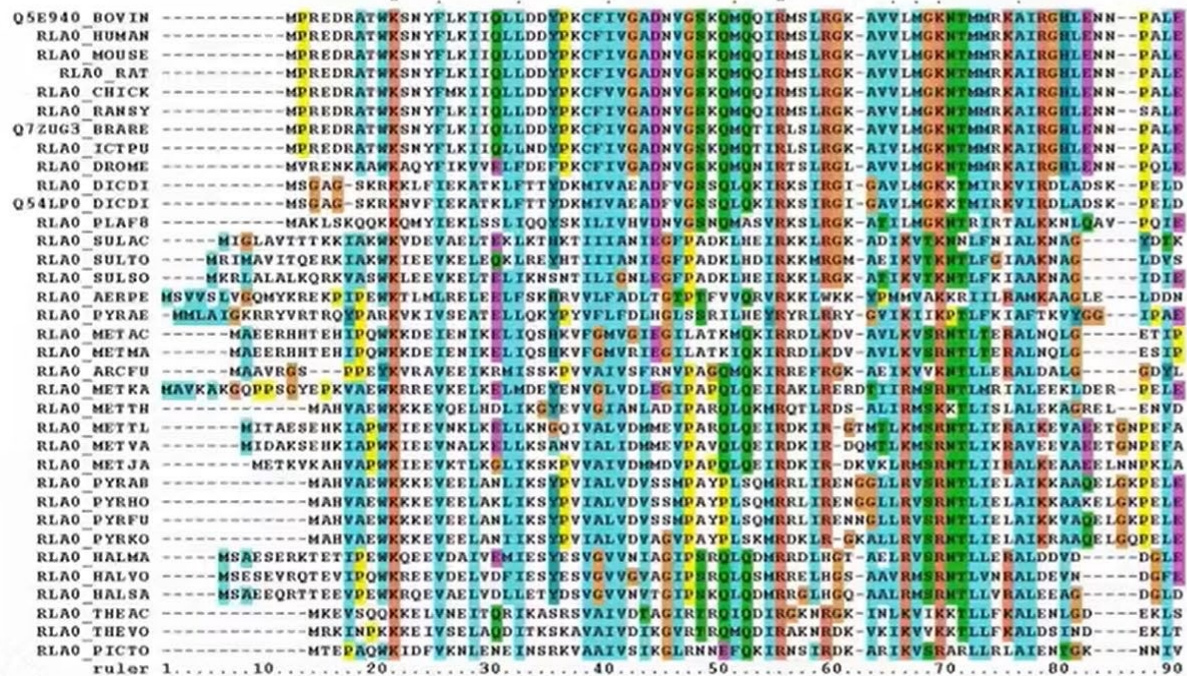


蛋白质中两两氨基酸的距离形成的map
Contact是指氨基酸之间距离小于某个阈值

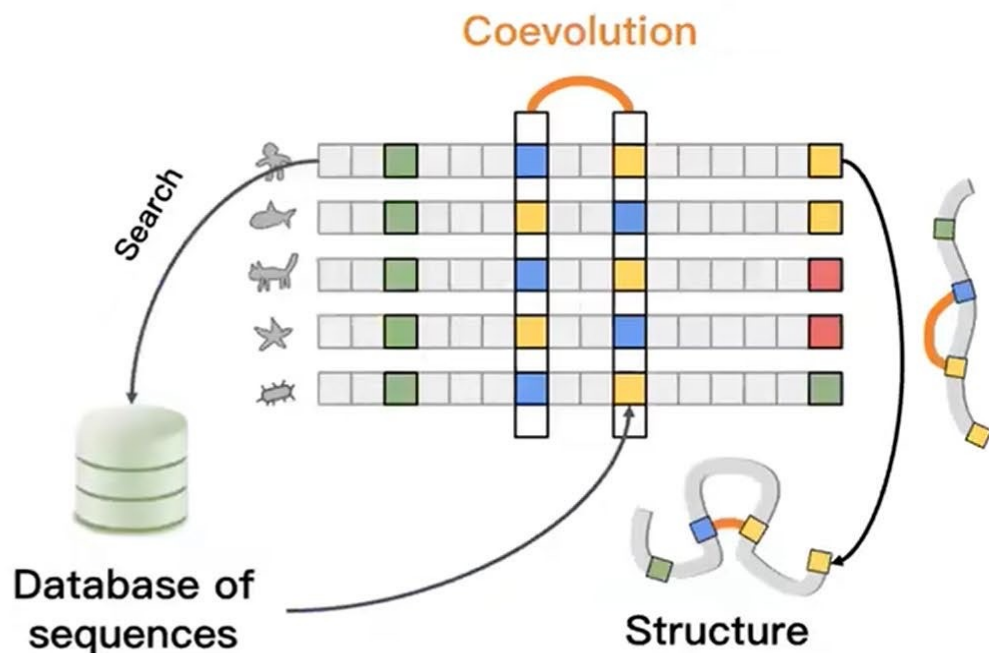
AlphaFold 1预测蛋白质的Contact Map: 理论基础



多序列比对 (MSA)



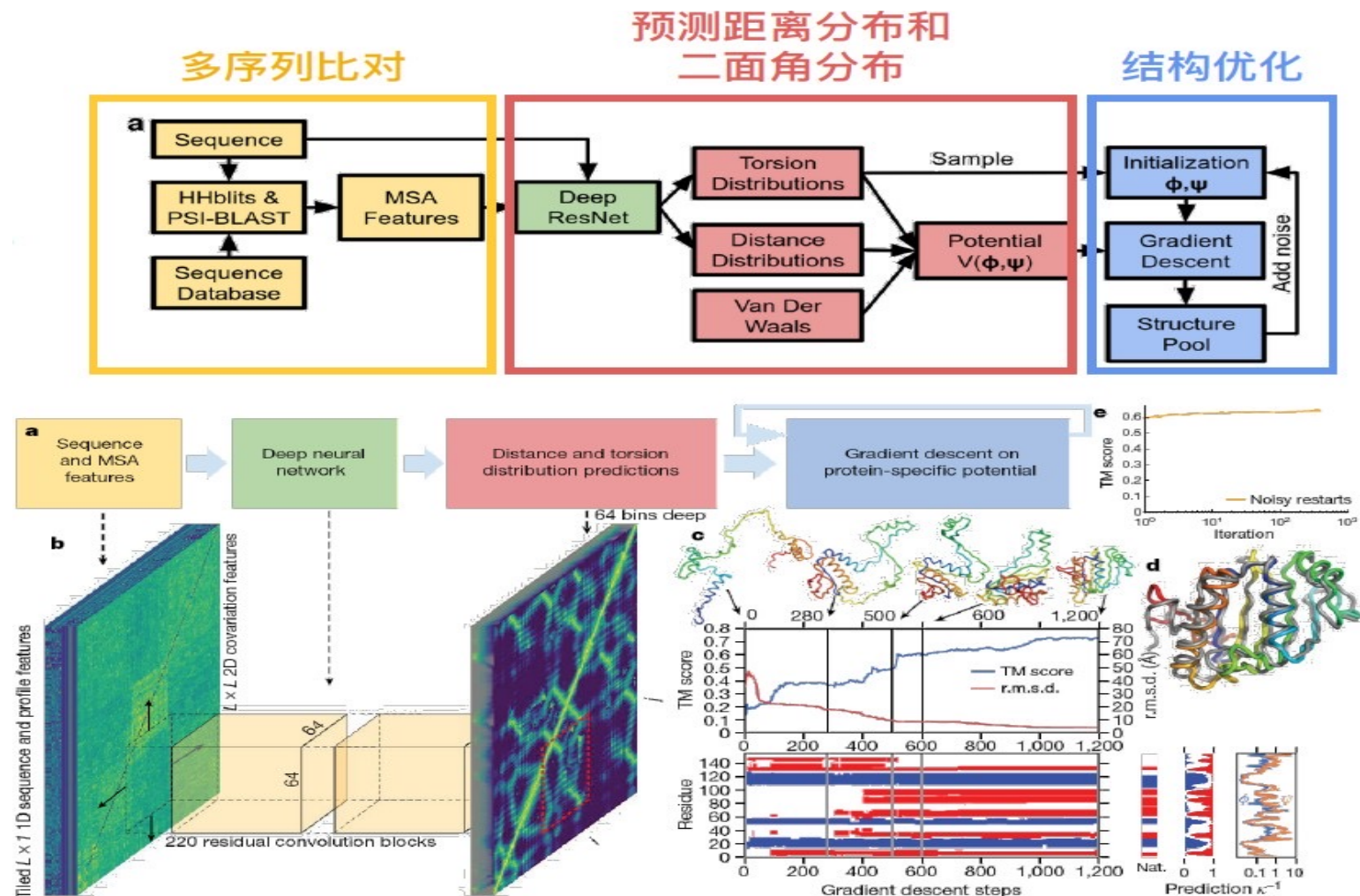
共进化信息预测蛋白质Contact



序列保守性信息：在不同物种中同一个位置的氨基酸不变
序列共进化信息：两个不同位置的氨基酸同步变化



AlphaFold 1 (Nature, 2020)



主要特点:

- 用ResNet从MSA预测Distance Map
- 同时预测了蛋白质中的二面角

成功原因:

- 强有力的硬件优势
- 架构层面的优势对比其他预测Contact Map的方法并不显著

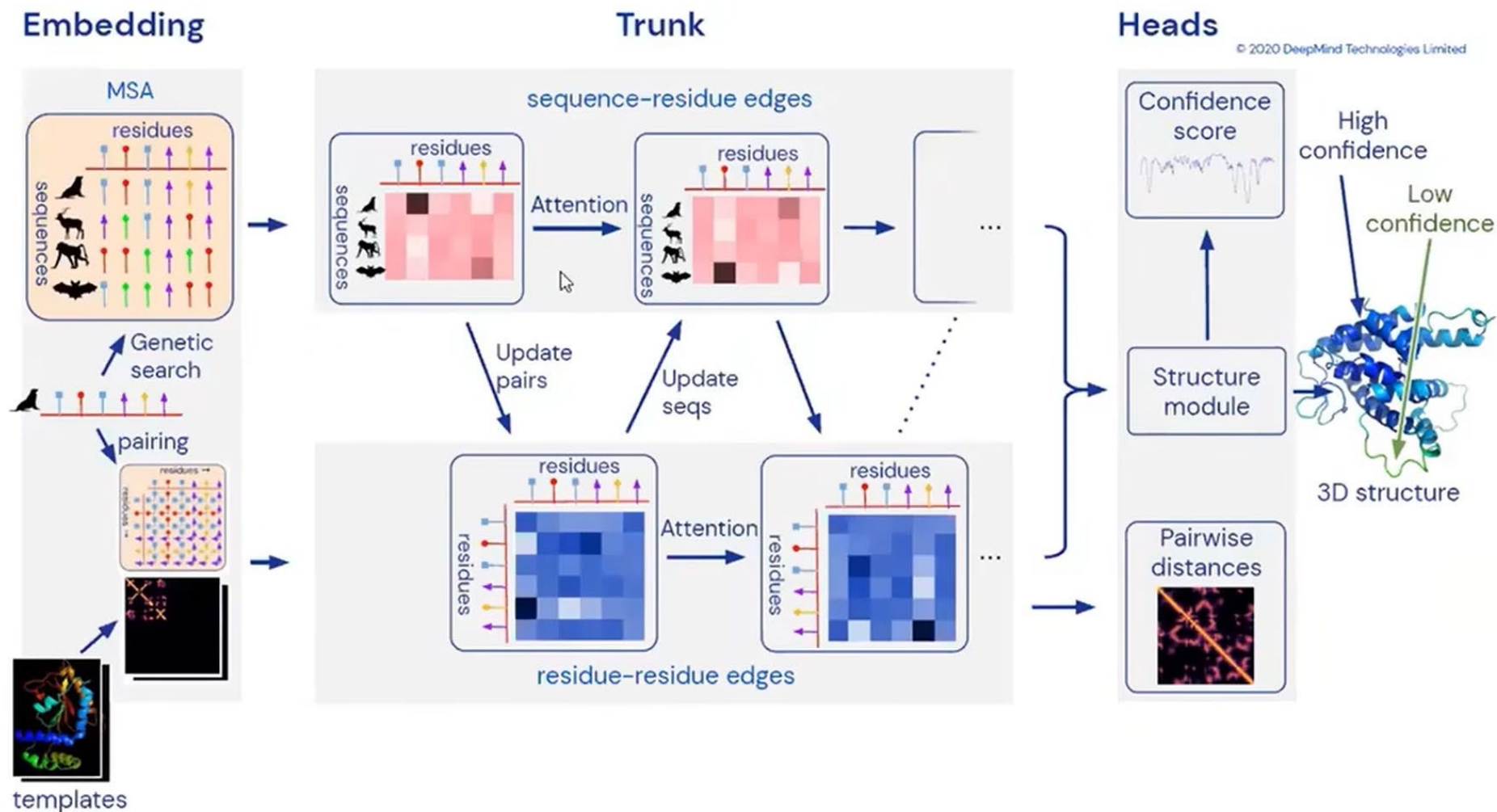


AlphaFold 2 (Nature, 2021)



主要特点:

- End-to-end架构
- 1D与2D信息之间使用了Attention
- 3D Equivariant (等变) Structure Module



Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021.



AlphaFold 2 (Nature 2021)



- 模型输入——更强大的MSA & Templates

序列数据库

- UniRef90 (JackHMMER) 来自Uniprot
- BFD (HHblits) 自建
- MGnify clusters (JackHMMER) 宏基因组

结构数据库

- PDB (用于训练)
- PDB70聚类 (hhsearch)

所有数据均来自公开数据库

丰富的MSA数据库加上优秀的MSA搜索比对方法，得到高质量的MSA结果用于训练



Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021.



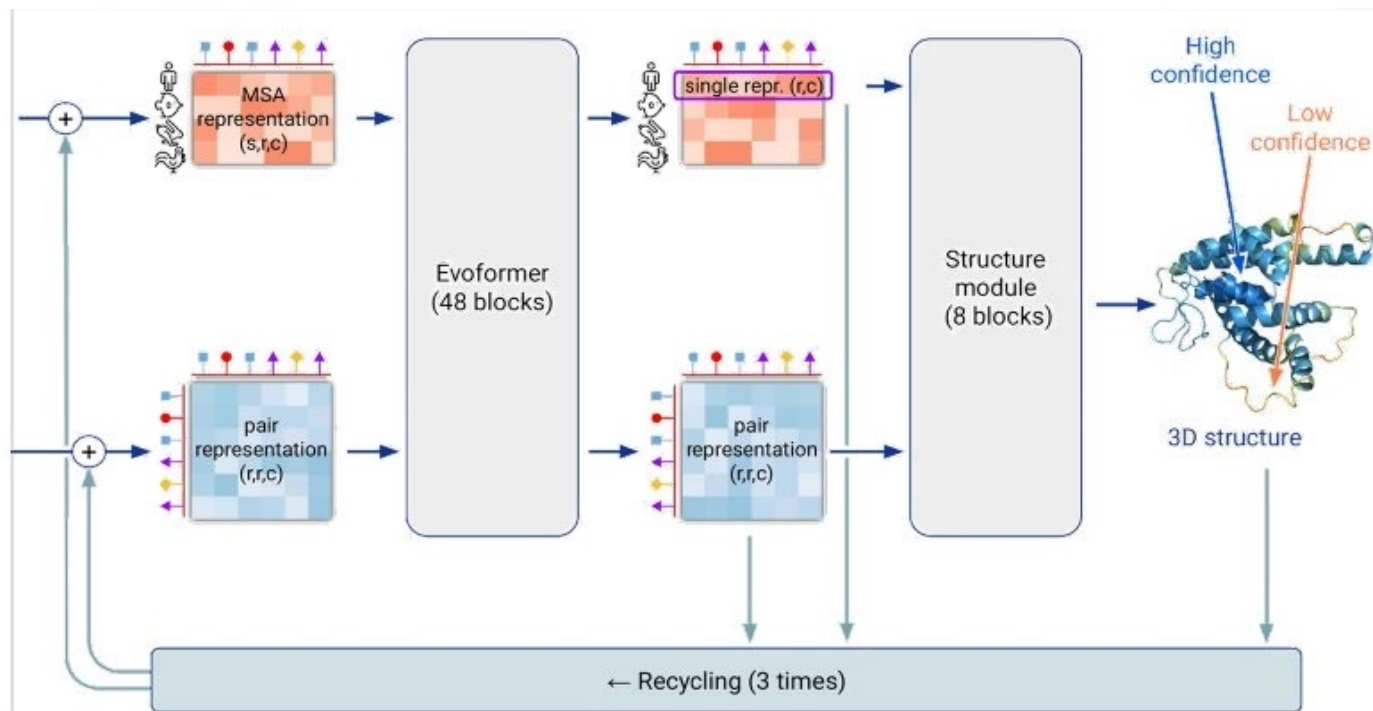
AlphaFold 2 (Nature 2021)



- 使用Recycling进行多轮迭代训练和测试

Recycling最初用于计算机视觉中的姿态估计 (post estimation)问题，将训练的结果返回输入继续迭代训练

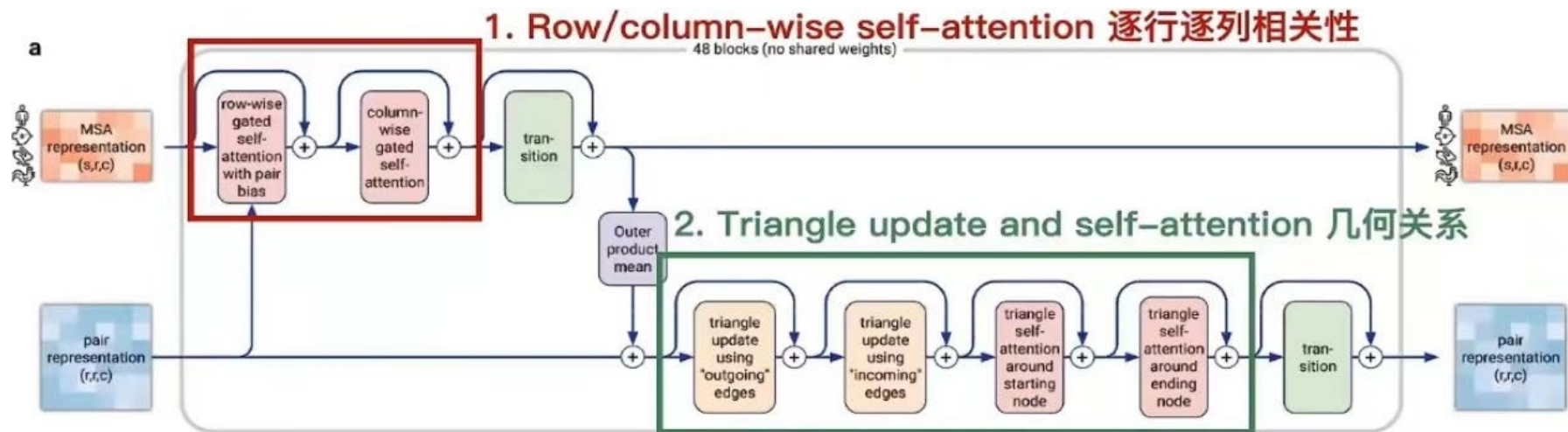
“We find it helpful to execute the network multiple times, each time embedding the previous outputs as additional inputs.”



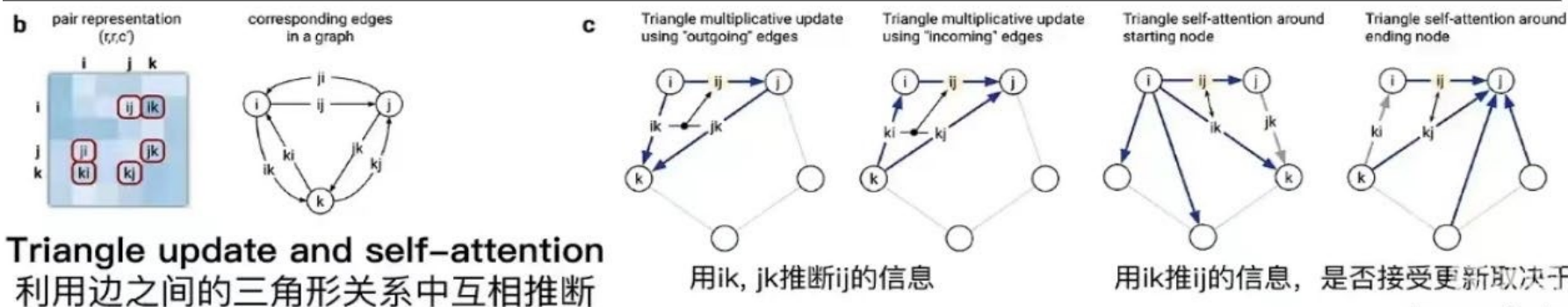
多轮迭代让模型的深度更深，能够通过迭代让结构预测更精确，预测到更精确的结构



- Evoformer: 用于结构预测的Attention架构



Evoformer的整体架构



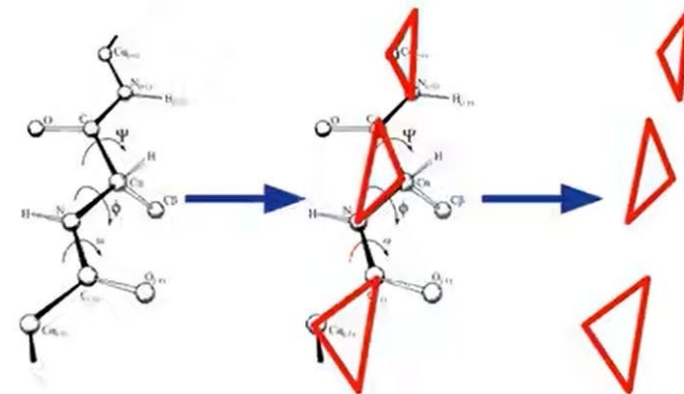
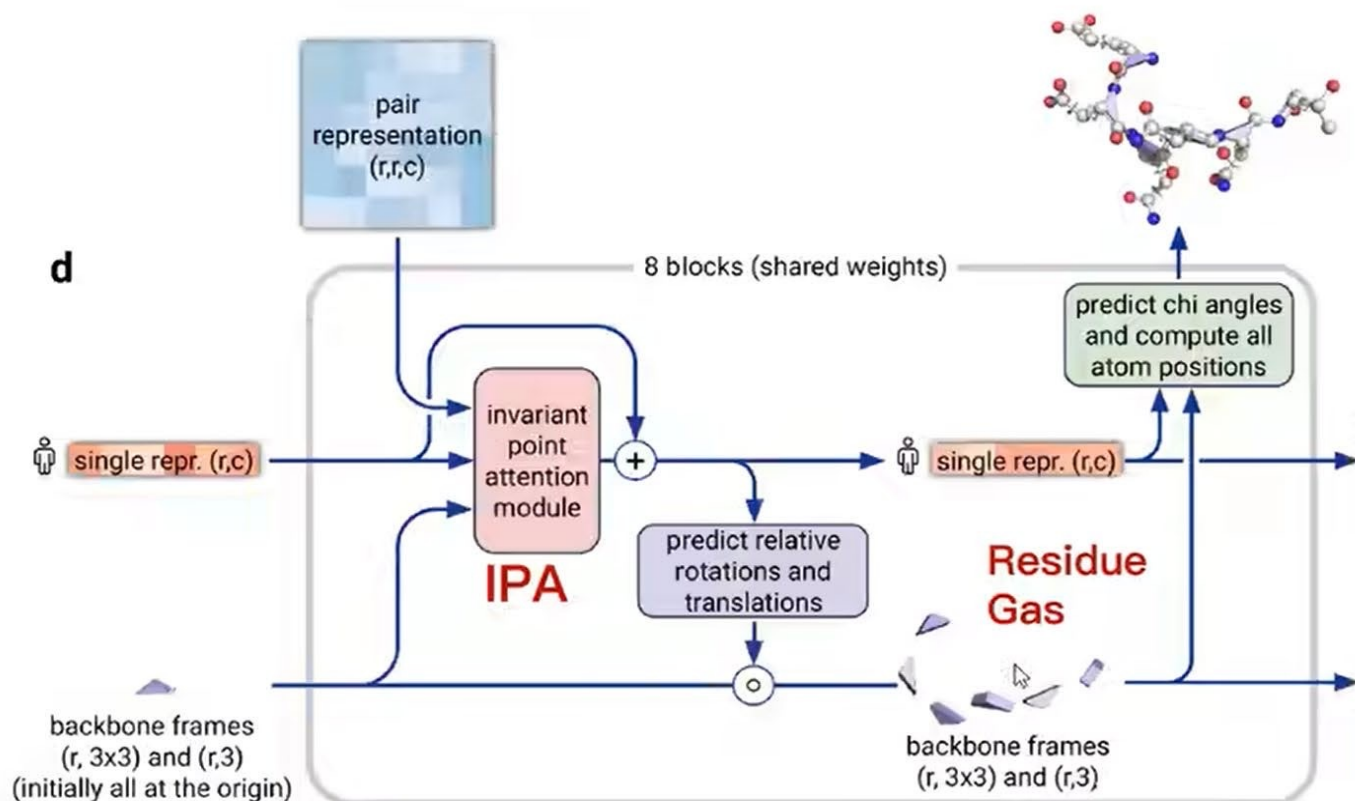
Jumper, J, et al. Nature, 2021

AlphaFold 2 (Nature 2021)



- Structure Module中的关键--Equivariant

重要架构：IPA (Invariant Point Attention) 和 Residue Gas

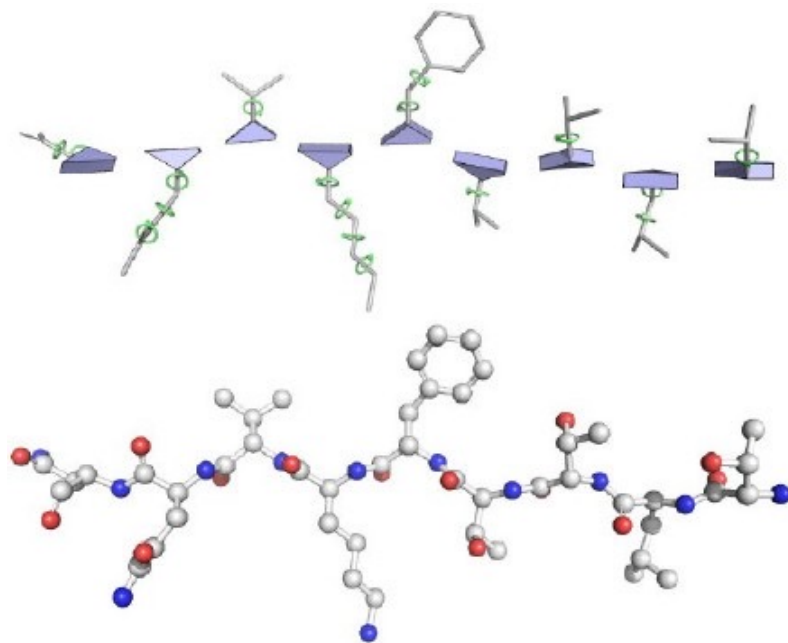


- IPA用于实现3D Equivariant (平移旋转等变性)
- Residue Gas用于表示蛋白质结构
- 输入：
 - 序列信息 (目标蛋白)
 - Distance Map信息
 - 蛋白质骨架初始Residue Gas
- 输出：
 - 全原子的位置坐标
 - IDDT-Ca (评估建模精度)

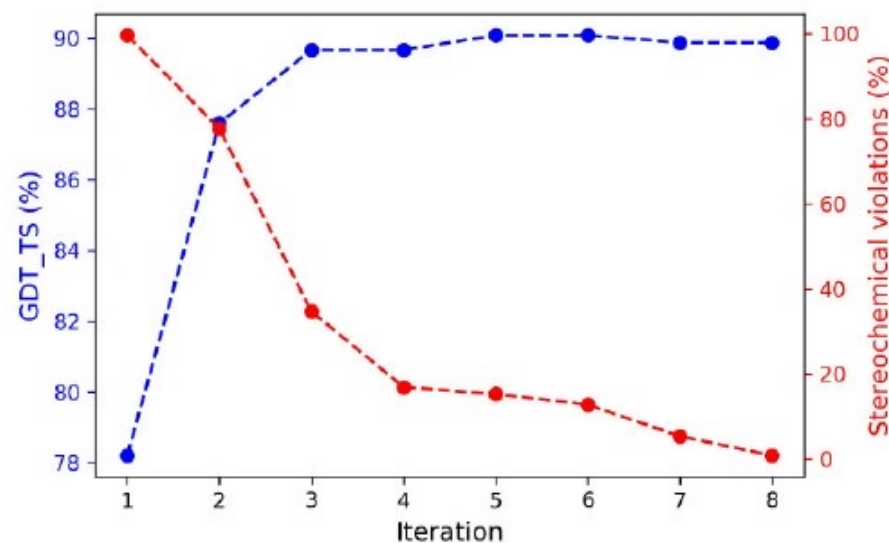
Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021.



- Structure Module中的优化过程——原子水平的优化



同时对主链结构和支链结构的优化，实现了原子水平的end-to-end三维结构预测和优化。

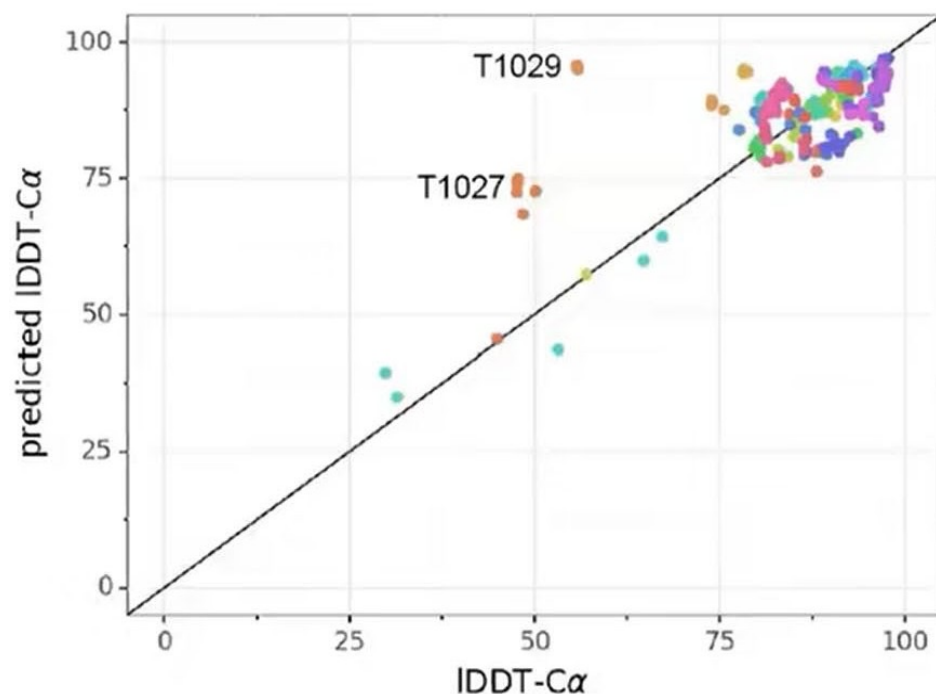


Target: T1041

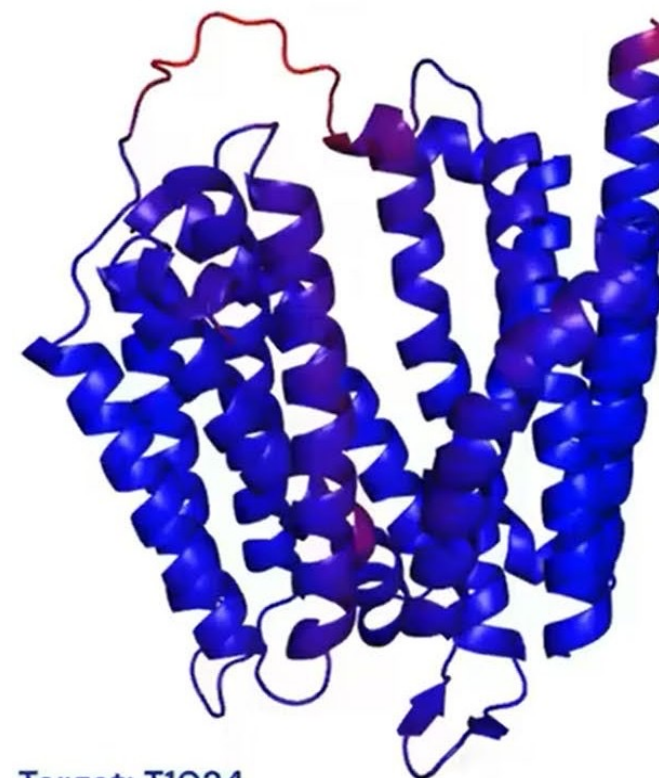
在建模准确性提升的同时，结构中的不合理成分也逐步降低

- 多输出—如何知道预测结构的精确度

模型预测的IDDT- $C\alpha$ 与实际值十分接近
MAE=3.3



预测的IDDT- $C\alpha$ 能反应预测结果的准确度
(蓝色：准确度高，红色：准确度低)

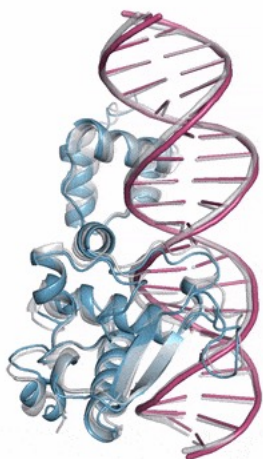


Target: T1024

- 在医药创新领域存在着“双十定律”，即生物医药企业研发一种创新药的完整周期至少需要10年，花费10亿美元。
- 约90%的药物会在临床试验阶段失败。
- AI技术能够颠覆传统药物研发进程，快速识别药物靶点，从数据库中匹配合适分子，设计、合成化合物并预测药物代谢性质和理化性质，可大大缩短药物研发时间、降低研发成本并提高成功率。
- 2022年7月，DeepMind团队发布的 AlphaFold 2 能够预测所有已分类蛋白质的3D结构。

- 2024年5月，**AlphaFold 3** 横空出世。初代 AlphaFold 诞生之后，一直停留在预测蛋白质的宇宙中。升级后的AlphaFold 3能够以前所未有的精度，预测出所有生物分子的结构和相互作用。

7R6R



Ground truth shown in gray

7R6R

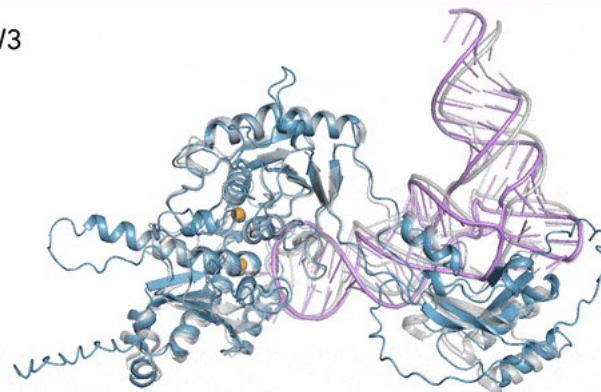


7R6R

Ground truth s

7R6R

8AW3



Ground truth shown in gray

8AW3



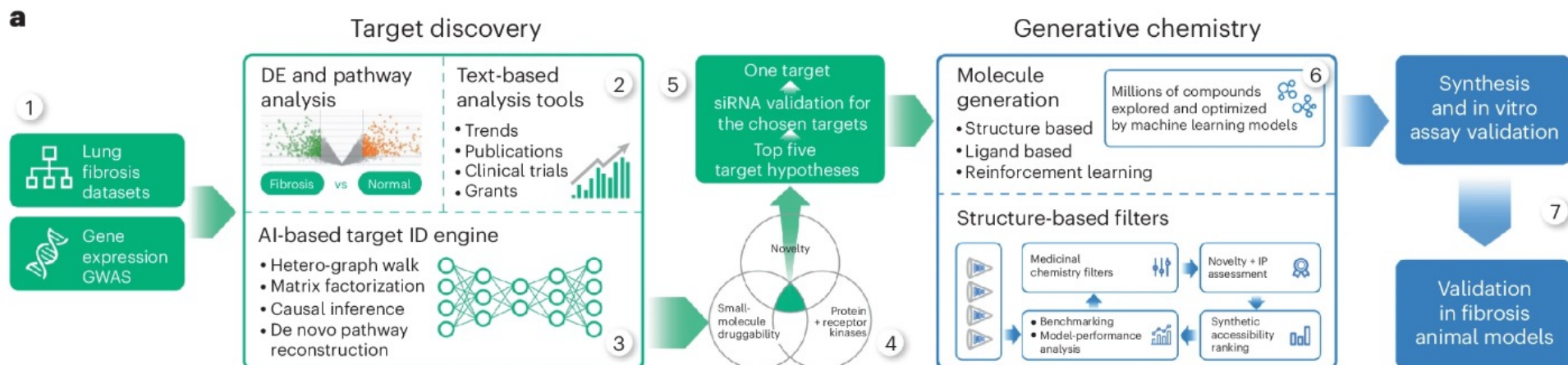
8AW3

Ground truth shown in gray

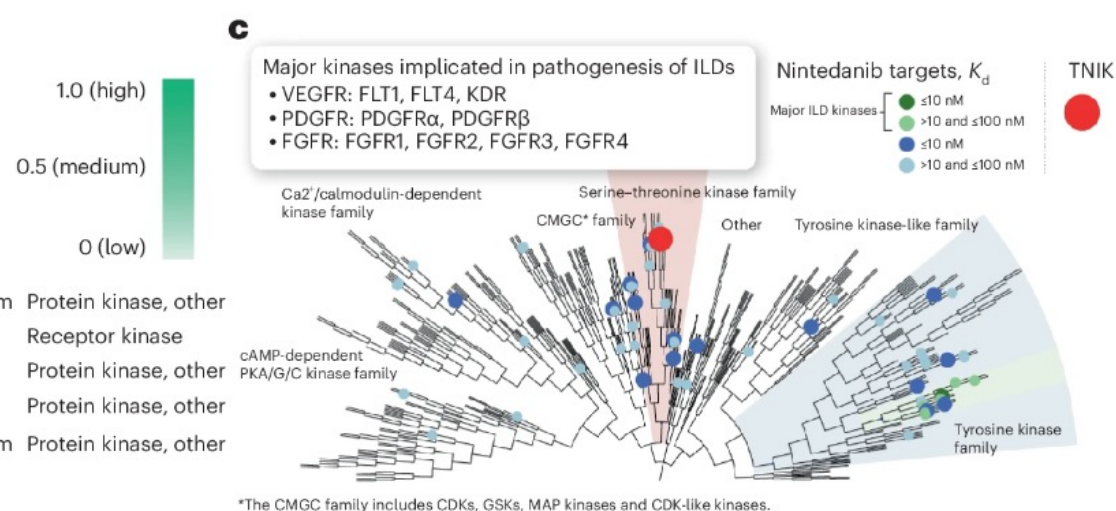
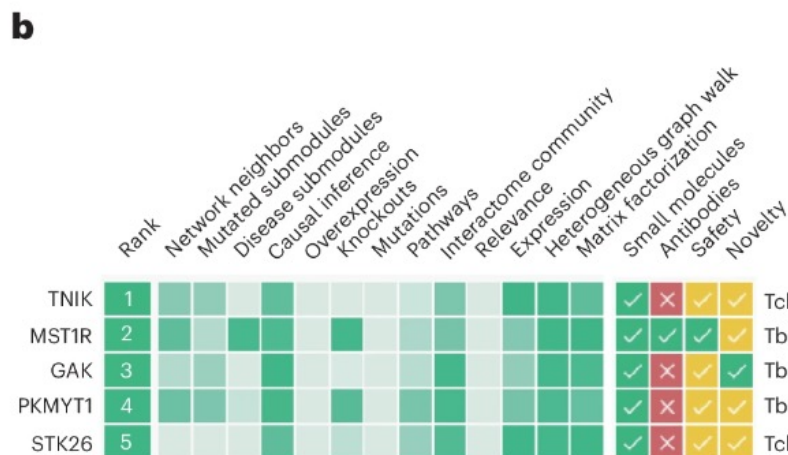
8AW3

7R6R-DNA结合蛋白：AlphaFold 3对一个分子复合体的预测展现了蛋白质（蓝色）与DNA双螺旋（粉色）的结合，其预测结果与通过繁复实验得到的真实分子结构（灰色）高度吻合

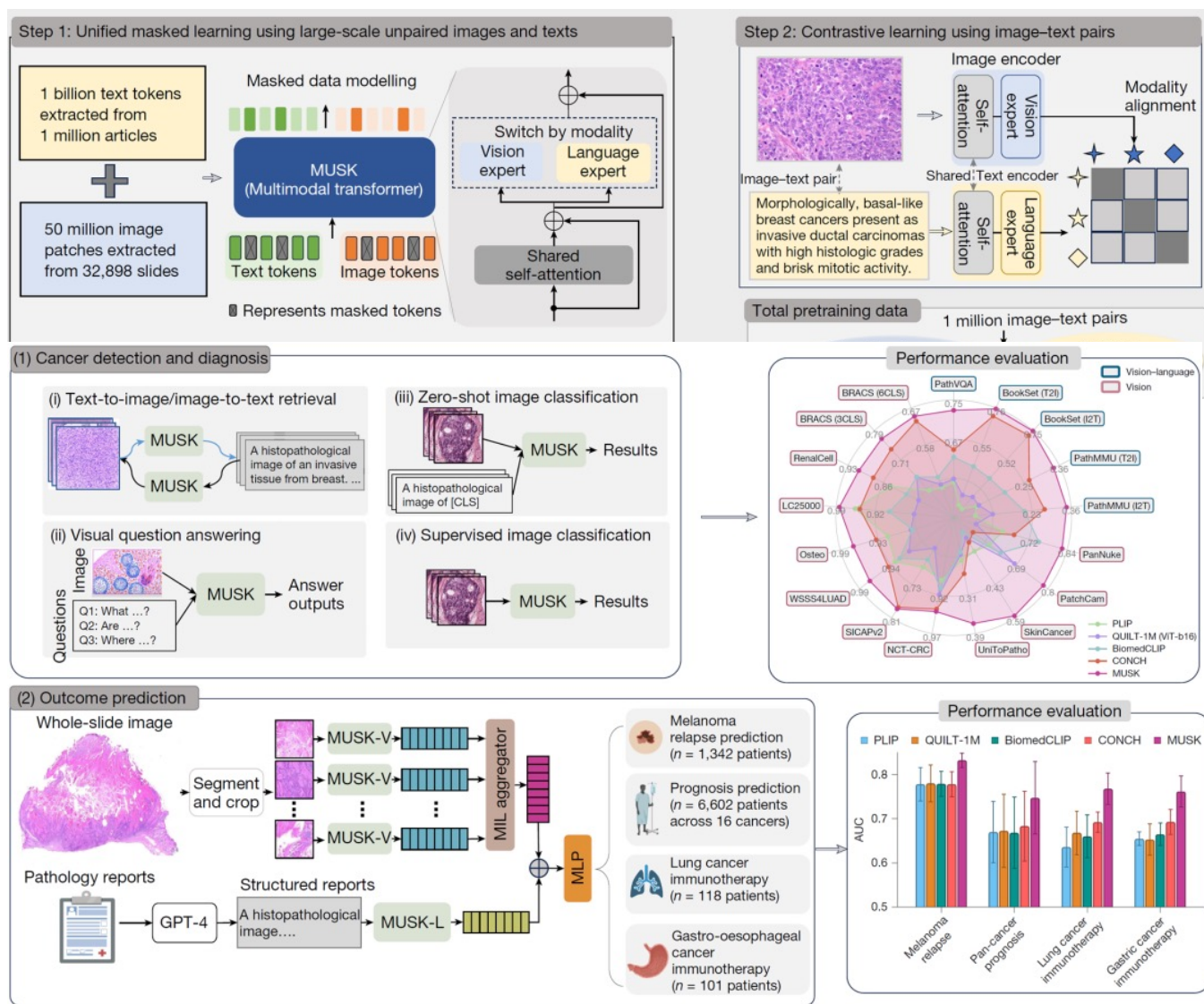
8AW3-RNA修饰蛋白：AlphaFold 3对一个包含蛋白质（蓝色）、一条RNA（紫色）和两个离子（黄色）的分子复合体的预测与真实结构（灰色）非常吻合



- 2023年, AI制药公司 Insilico Medicine 宣布, 其通过生成式AI筛选靶点并设计的抗特发性肺纤维化小分子候选药物 **INS018_055** 已完成II期临床试验, 首例患者给药。

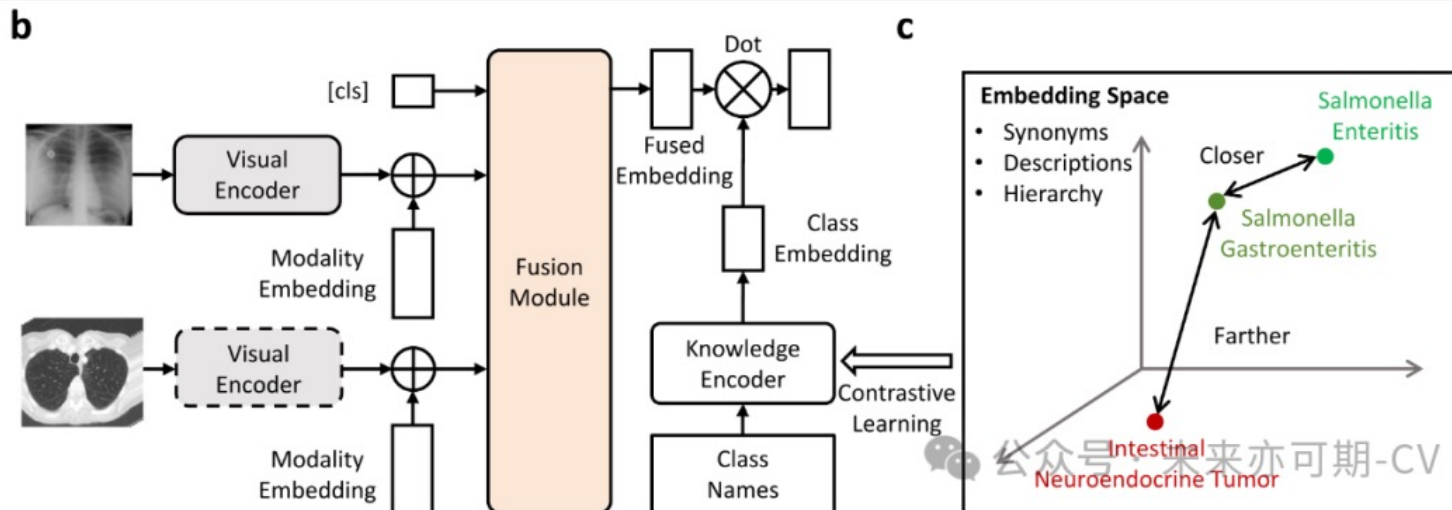
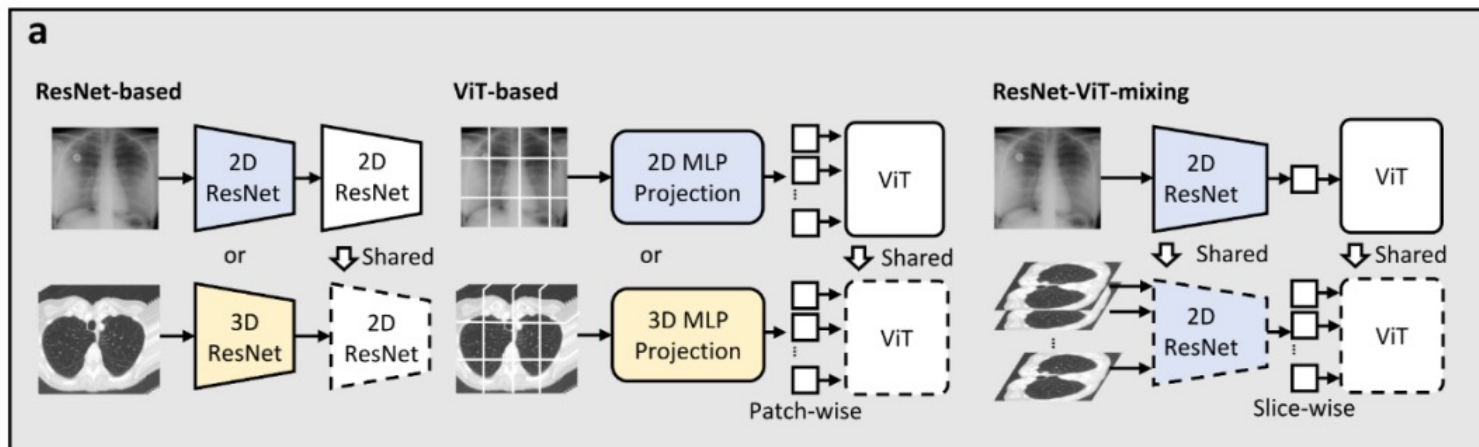


- 2024年3月8日, Insilico Medicine 在 *Nature Biotechnology* 发文介绍了该药物的研发历程。



精准肿瘤学的视觉-语言基础模型

- 研究团队基于BeiT3的网络结构，开发了MUSK预训练基础模型
- MUSK模型在23个病理图像的 patch-level和slide-level基准数据集上进行了全面评估，并在超过8000名患者的多模态数据上验证了其有效性。
- 这些验证任务涵盖了多个关键的临床场景，包括黑色素瘤复发预测、16种癌症类型的预后预测，以及肺癌和胃食管癌中的免疫治疗反应预测。



放射学图像上的大规模长尾诊断

- 构建了一个大规模放射学影像数据集，包含4万+临床案例，5千+疾病种类，7种解剖部位和9种成像模态，据此训练了一个多模态、跨部位、多病种的病例级放射学综合诊断基础模型
- 提出了新的模型架构，可以处理来自不同成像模态的任意数量的输入扫描，该架构经过知识增强训练，高效地将医疗领域知识结合起来
- 为多解剖部位的多模态放射学影像长尾诊断建立了一个新的多维度全方位评测基准

“女娲” 系列生命科学大模型由复旦大学与上智院共同研发

- **生物结构大模型**依据不同数据特性，采用不同分子处理等级，统一至token等级，以获取生物分子复合结构的几何embedding信息。同时，该模型利用共进化信息与Template信息，建模复合物中各个原子之间的作用关系，并通过扩散模型实现原子级别的坐标建模，捕捉蛋白质和其它生物分子的相互关系，达到了Alphafold3的同等水平
- **基因导航大模型**能够对基因组非编码的“暗物质”区域进行功能注释，发现复杂疾病的机制和可能的疾病治疗靶点。目前该模型在不同细胞组织调控关系预测精度上表现优异，与实验验证结果相比，精度可达85%以上，在长距离调控关系预测精度上更是比现有最好模型高一倍以上，如在大脑皮层、大脑前额叶、骨髓等细胞组织上均有出色表现。
- **生命流体大模型**通过医学图像几何约束与物理学方程约束，在实际临床场景里针对复杂血管几何进行高精度血液流体动力学模拟时表现优异。以实际血管几何数据为例，它能够实现仅8%的速度物理场模拟误差，这一数据大幅超越了JMLR 2024上发表的方法，并且计算效率相较于传统计算流体动力学（CFD）提升了百倍之多，极大地提高了模拟的精准性与时效性。

目录

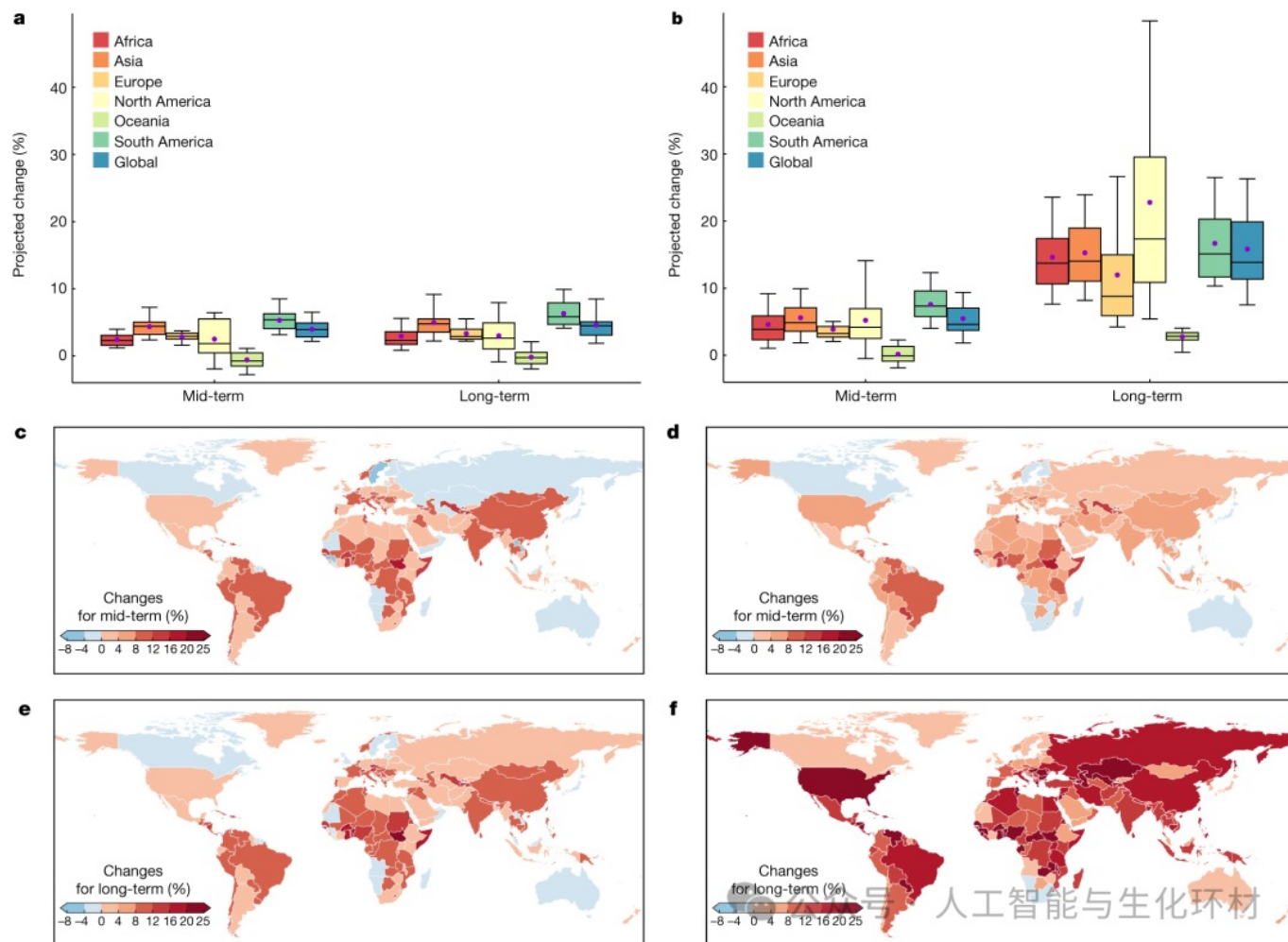
1 AI For Science介绍

2 AI应用于微观科学

3 AI应用于气象学

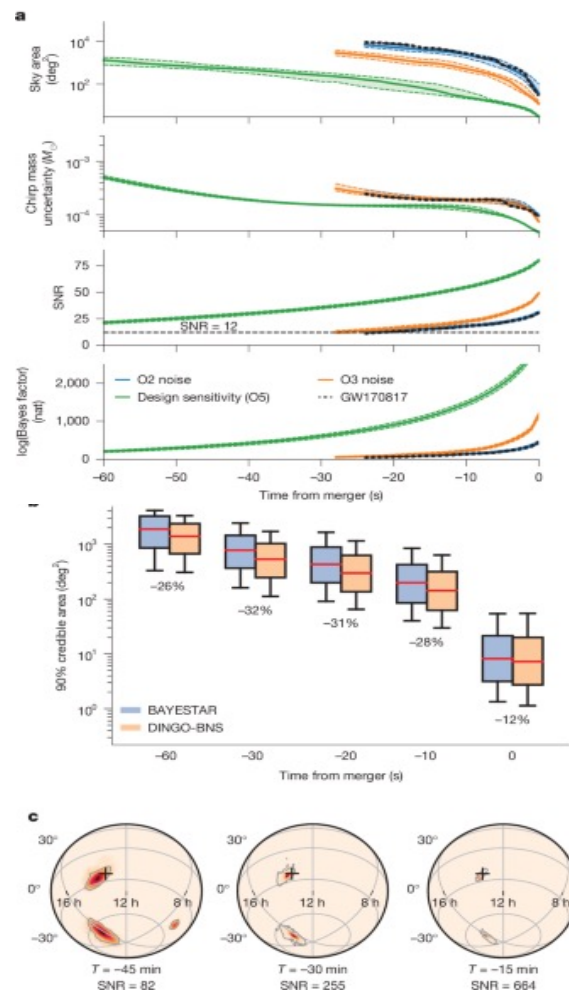
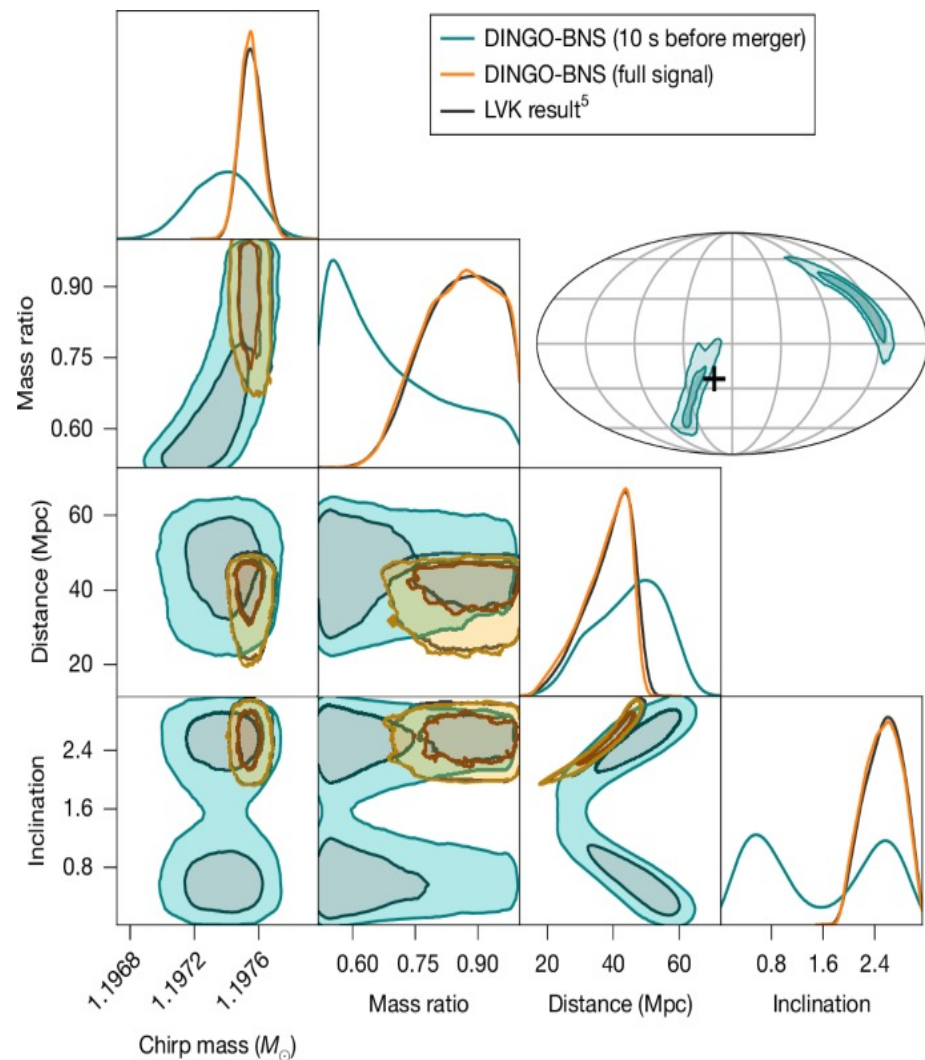
4 AI应用于生物医学

5 **AI应用于其他领域**



全球氨减排的肥料管理

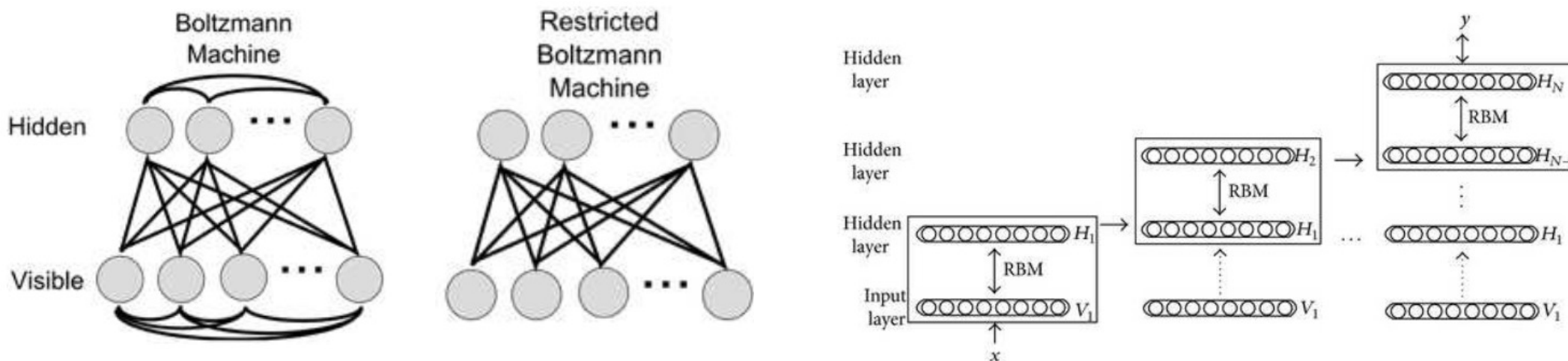
- 开发了一种机器学习模型，通过整合实地观测数据，生成作物特定的、空间分辨率为全球5弧分的NH₃排放因子
- 根据机器学习模型指导优化施肥管理，可在不改变总氮肥投入的情况下，将NH₃排放量减少约38%



基于机器学习对双中子星合并实时推理

- DINGO-BNS机器学习方法可在引力波信号抵达地球后1秒内完成双中子星合并的定位
- 甚至在合并之前就可以进行准确定位; 与近似低延迟方法相比, 定位精度提高了约 30%
- 该方法可以扩展到长达一小时的长信号, 从而作为下一代地面和天基探测器的数据分析蓝图

- 玻尔兹曼机是一种基于统计物理学的机器学习模型，由Geoffrey Hinton在Hopfield网络的基础上发展而来。核心思想源于统计物理学中的玻尔兹曼分布，通过模拟神经元之间的随机激活与协同作用，学习数据中的潜在分布规律。
- 其变体有受限玻尔兹曼机（RBM）、深度玻尔兹曼机（DBM）、量子玻尔兹曼机（QBM）等
- 相对论重离子碰撞实验中，QBM通过量子退火生成低维特征表示，用于快速分类末态粒子（如区分 π 介子与 ω 介子），减少后续分析的算力需求



Ackley D H. et al. A learning algorithm for Boltzmann machines. Cognitive science, 1985.

Hinton G E. et al. A practical guide to training restricted Boltzmann machines. Neural Networks, 2012.

AI + 物理科学

• 计算物理：新范式

- GNoME (DeepMind, Nature 2023): 预测超200万种新型稳定晶体结构
- 复旦大学: AI预测电子哈密顿量, 为材料科学提供新工具

• 大型实验装置智能控制

- 深度强化学习稳定托卡马克等离子体, 防止撕裂不稳定 (Nature 2024)
- AlphaQubit: 量子纠错解码器 (Nature 2024)

• 物理定律的符号发现

- MIT AI-Feynman: 神经网络进行物理公式识别与理论推导
- AI发现隐藏物理对称性, 拓展新物理定律的发现边界



• AI + 机器人化学家

- 自主移动机器人完成探索性化学合成 (Nature 2024)
- 自动化自优化光催化流动反应 (Science 2024)

• AI + 势能面 & 分子设计

- 神经网络势函数兼顾精度与速度，成功用于反应机理预测
- TamGen化学语言模型 (Nat Commun 2024)：
靶点感知药物分子生成
- 闭环知识迁移 (Nature 2024)：AI与实验交互，自主产生化学规律



AI + 材料科学

● AI驱动材料逆向设计

- MatterGen (Microsoft): 扩散模型生成稳定材料, 拓展搜索空间
- 图神经网络+多保真度学习, 电池/催化/聚合物快速筛选

● AI智能体与自主实验室

- MatChat / MatterChat: 材料科学大模型, 可溯源问答与结构预测
- A-Lab (UC Berkeley): 机器人+主动学习, 全自动材料合成实验室



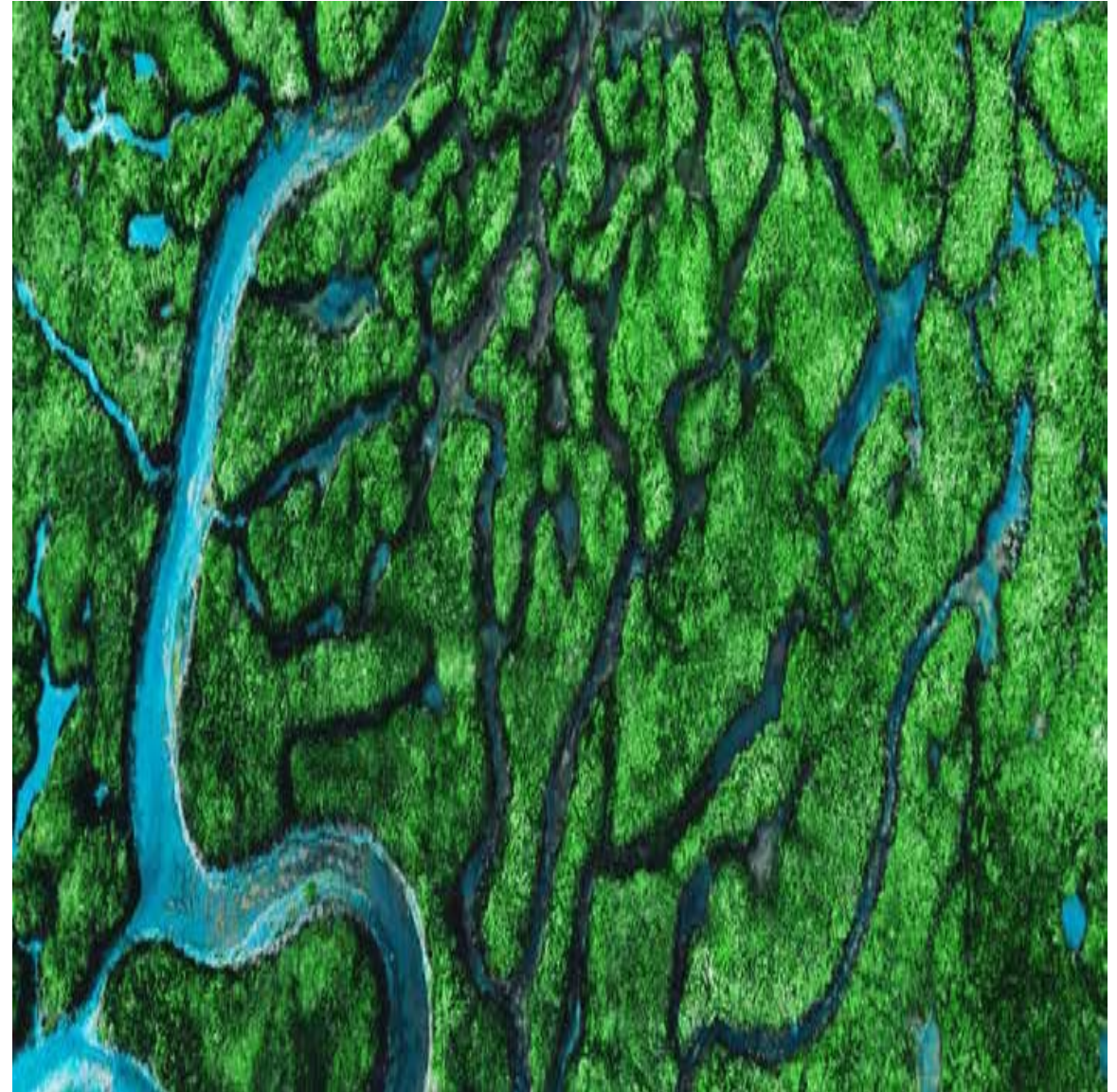
AI + 工程学与人文社科

● AI + 工程学

- 通信：AI优化6G网络调制识别、信道估计与干扰抑制
- 遥感：多模态深度学习融合SAR与光学图像，高精度地物分类
- 微电子：AI辅助EDA加速芯片版图规划，缩短IC设计周期

● AI + 人文社科

- 计算社会科学：NLP分析舆情；图神经网络建模信息传播
- 数字人文：AI辅助历史文献分析，文化遗产数字化保护
- AI伦理与治理：构建负责任AI框架，平衡创新与安全



AI4Science 展望

● 未来技术挑战

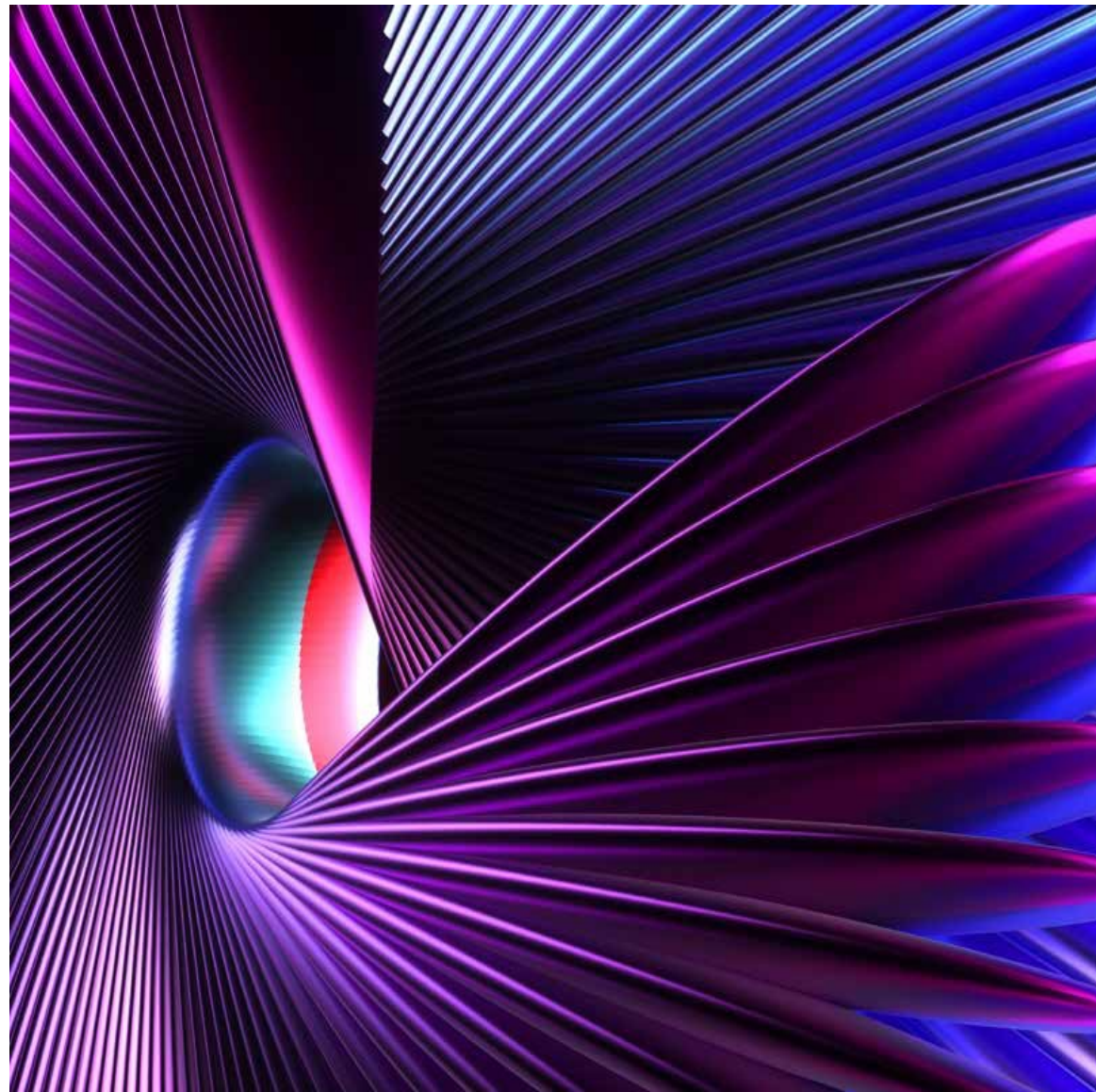
- 跨尺度建模：量子到宏观的统一多尺度框架
- 自主科学发现：AI+机器人实验，闭环假说生成与验证

● 政策建议 (AI for Science 2025白皮书)

- 建立开放科学数据共享平台，统一数据标准
- 加强AI+科学跨学科人才培养
- 构建负责任AI伦理框架；推动产学研协同转化

● 中国与复旦大学的战略定位

- 中国AI4S论文全球领跑；2024年物理科学AI出版物超7万篇
- 风乌/伏羲（气象）、女娲（生命科学）等系统性突破



- AI4Science有着和其他AI应用类似的问题（小样本、少标注）
- AI4Science将对自然科学带来深刻的影响
- AI4Science需要顶尖学者通力合作

谢谢！