

AlphaFold 原理、应用与展望

钟博子韬

上海交通大学

zbztzhz@gmail.com

目录 Content

I. 蛋白质结构预测

II. AlphaFold模型架构

III. Multimer的改进

IV. 预测结果

V. 高通量结构预测

VI. 应用实例

VII. 如何正确使用

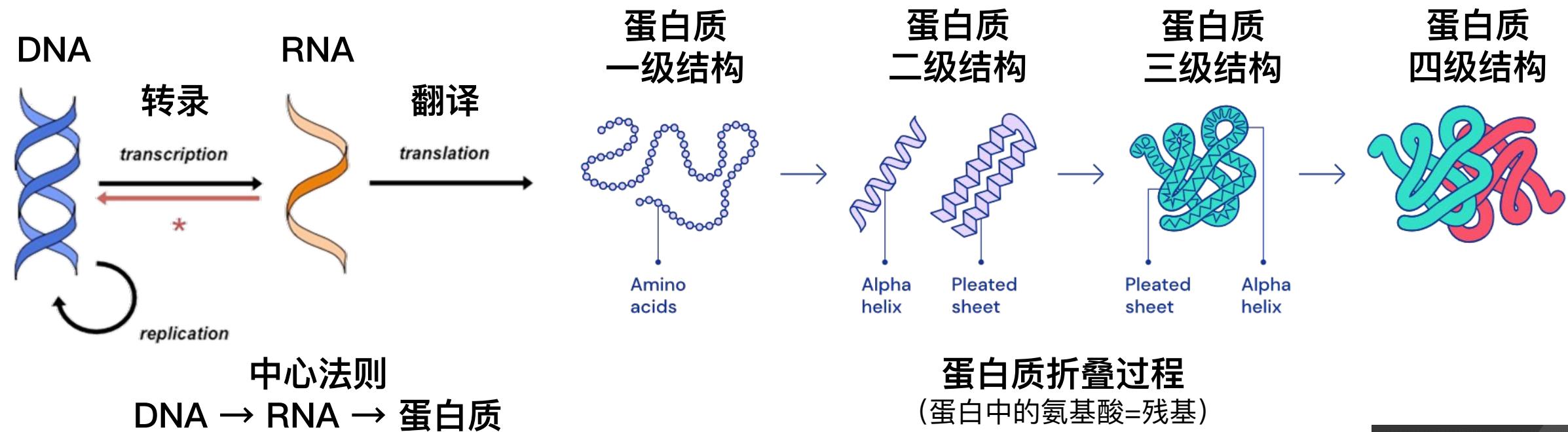
我希望实现这样的效果.....

- 低门槛，更多面向生物背景
- 讲模型是为了解释其为何能表现优异
- 能让大家更好地使用，提出更新颖的想法
- 明确AlphaFold能做什么，不能做什么
- 建立共识，扫除误解

相比于半年前的报告增加了大量新内容.....

- 高通量AlphaFold: ParaFold
- 各种应用实例
- 什么能做，什么不能做

蛋白质的折叠过程



Anfinsen's Dogma

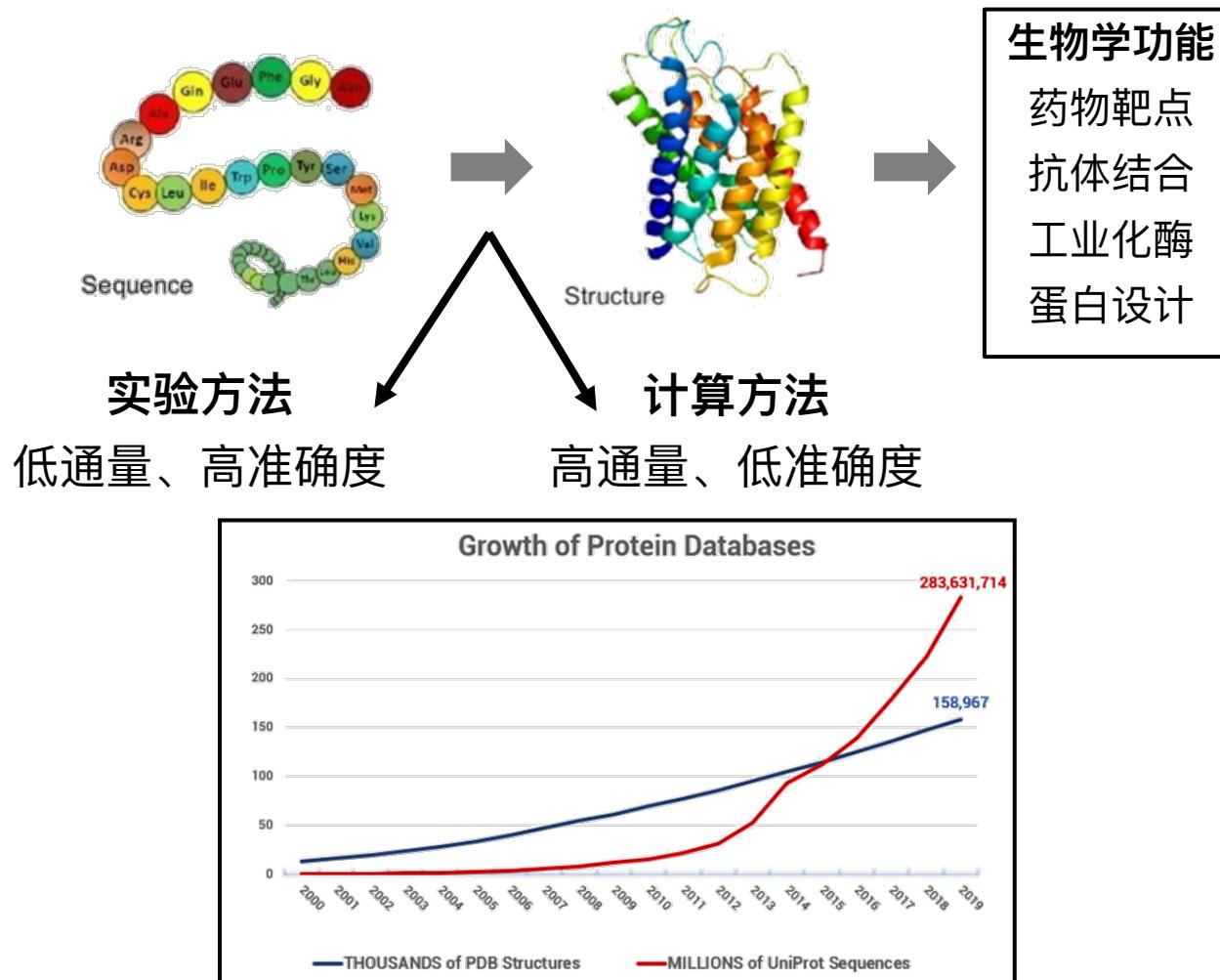
- 蛋白质折叠成原始结构所需的信息都已被编码在氨基酸序列中
- 蛋白质折叠到最小能量状态
- 大多数蛋白质会折叠成一个独特的构象

Christian B. Anfinsen
1972 Nobel Prize in Chemistry

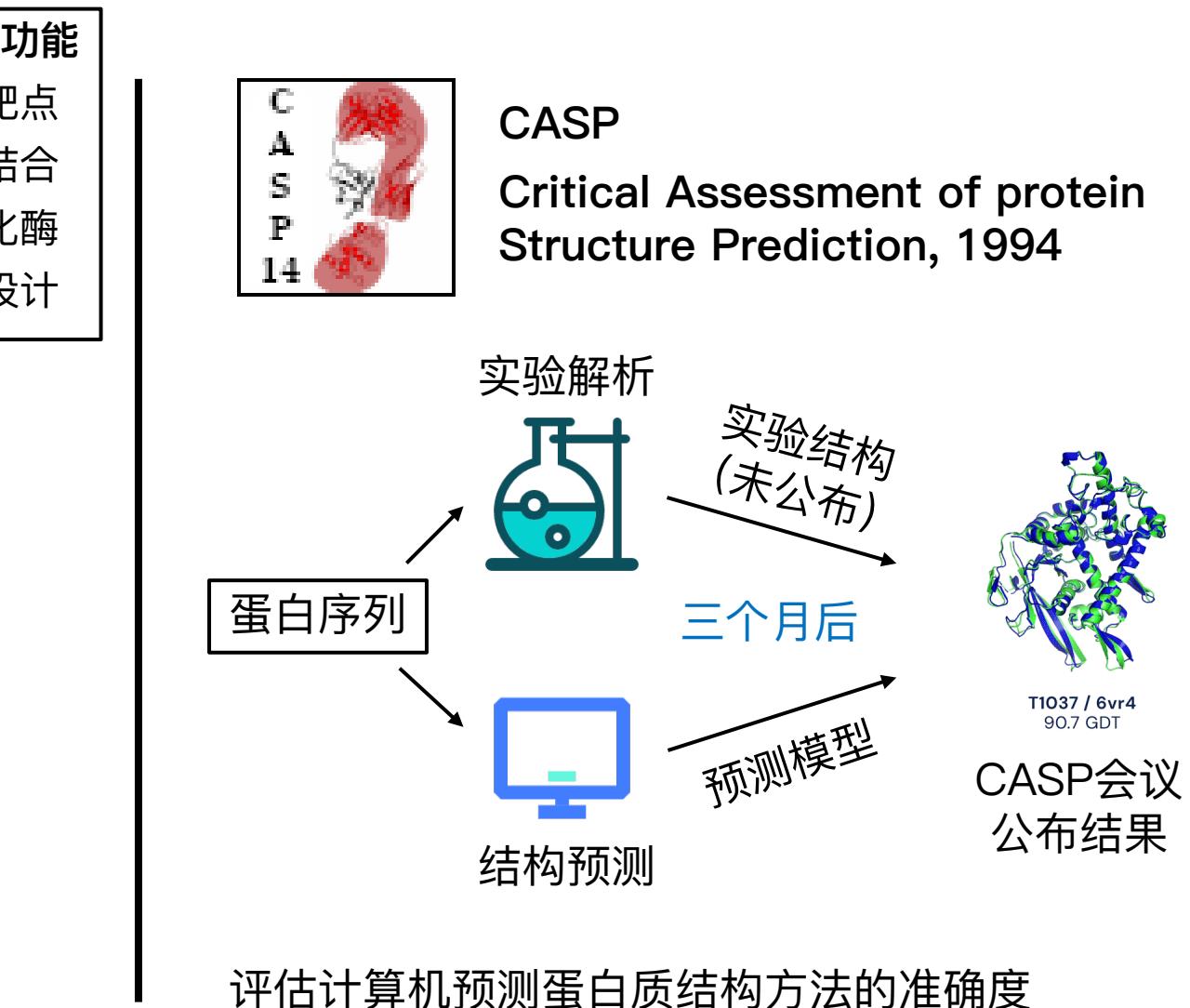


Picture from DeepMind
Anfinsen C B. Science, 1973, 181(4096): 223–230.

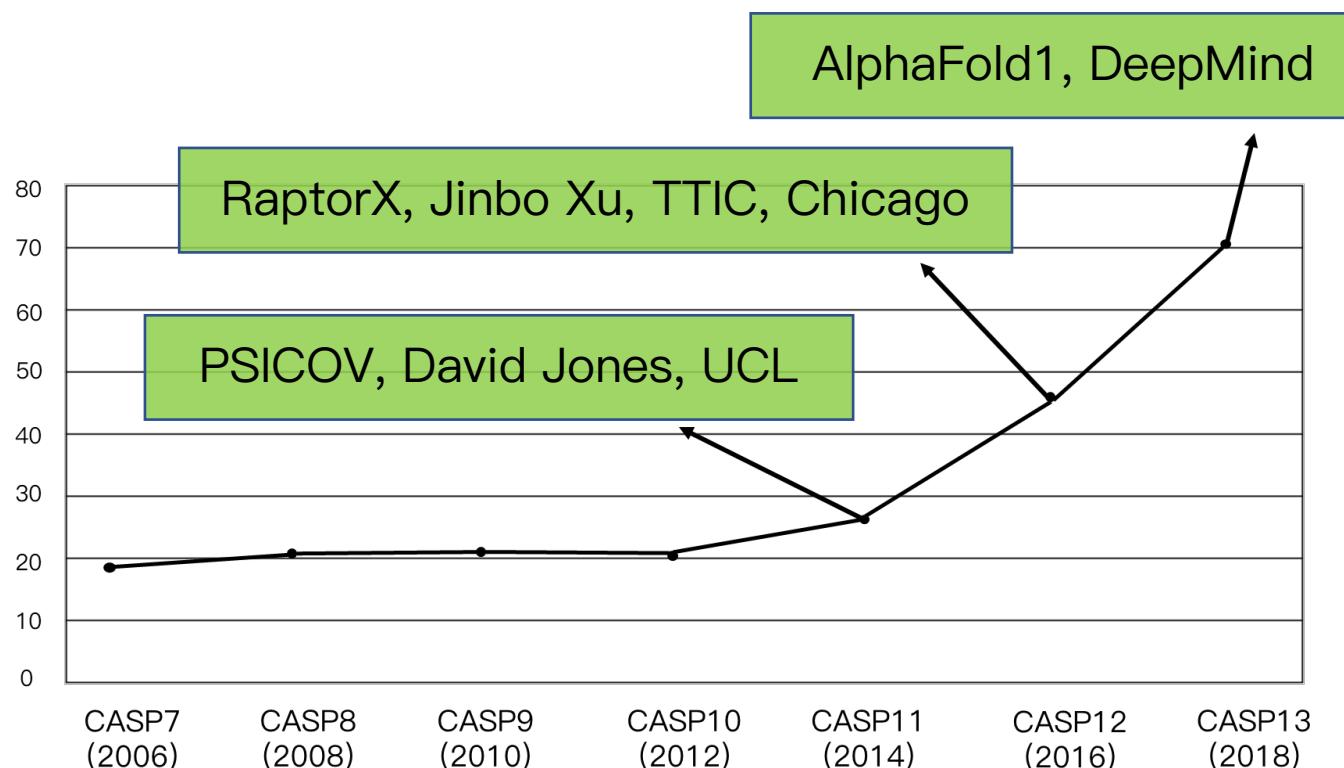
蛋白质结构预测：半个世纪的难题



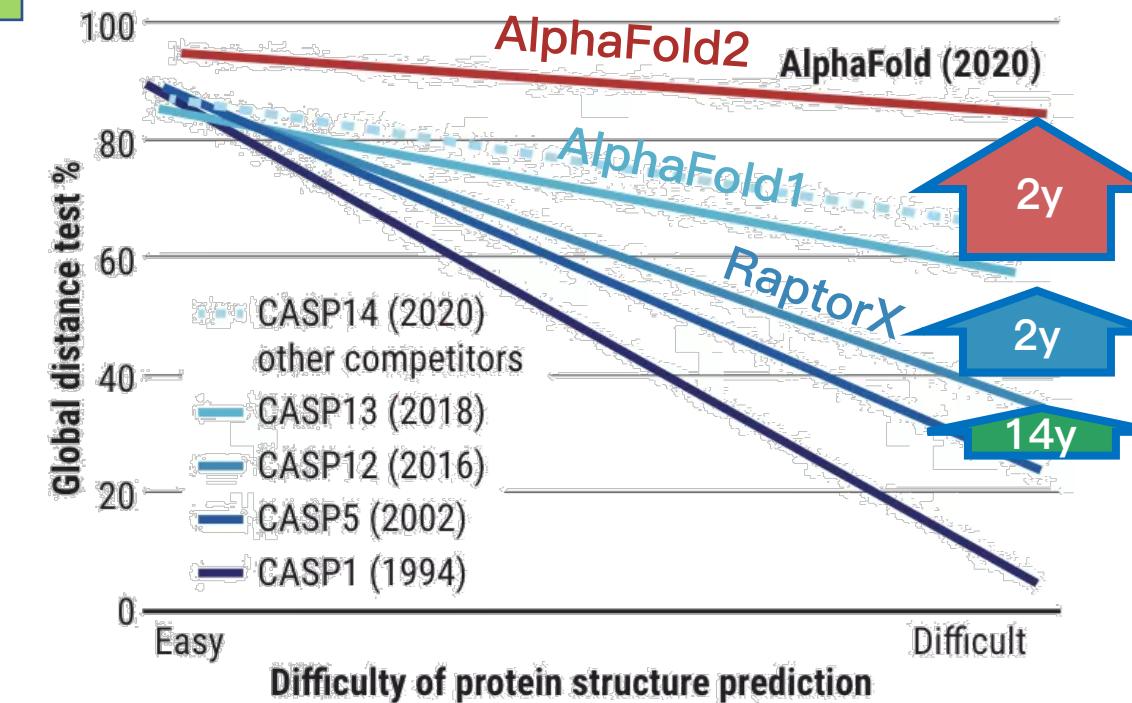
海量的序列信息，少量的结构信息
需要高精度高通量发现蛋白质结构的方法



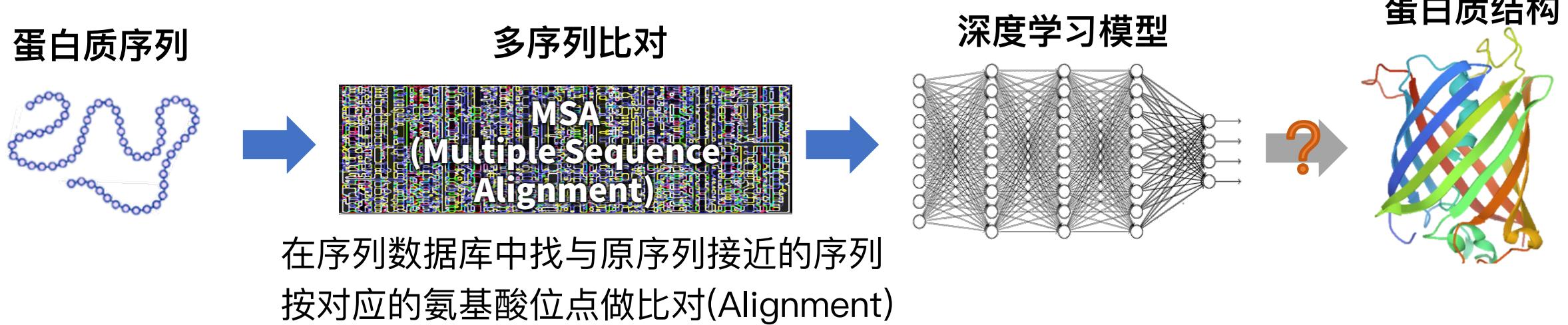
CASP中的关键提升



- 2014 (MSA): PSICOV, David Jones, UCL
- 2016 (Deep Learning/ResNet): RaptorX–Contact, Jinbo Xu, TTIC, Chicago
- 2020 (Transformer): AlphaFold2, DeepMind

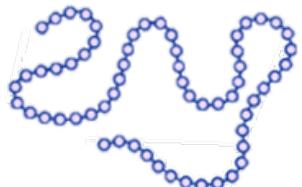


预测蛋白质的Contact Map: 间接预测蛋白质结构



预测蛋白质的Contact Map: 间接预测蛋白质结构

蛋白质序列

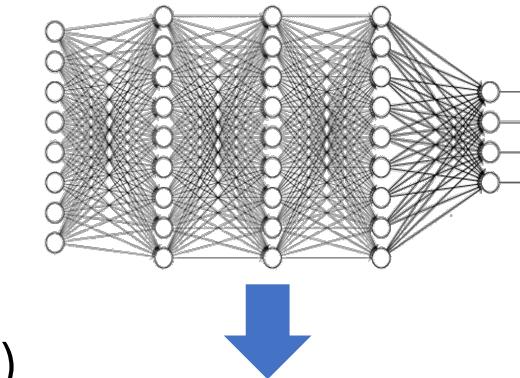


多序列比对

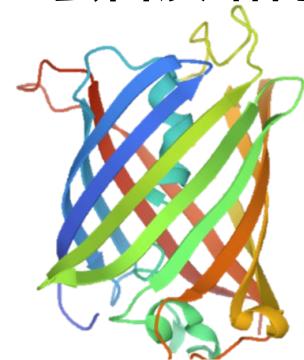


在序列数据库中找与原序列接近的序列
按对应的氨基酸位点做比对(Alignment)

深度学习模型



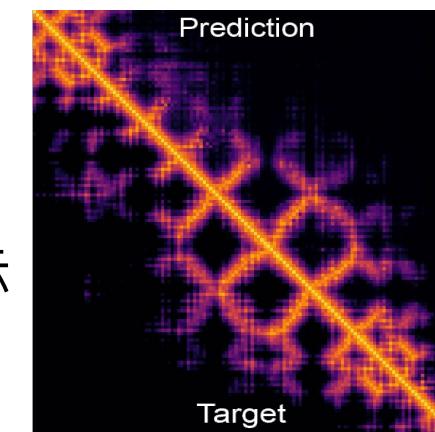
蛋白质结构



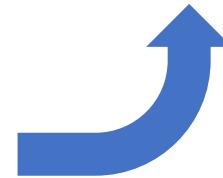
为什么选用这种方法：

- 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)

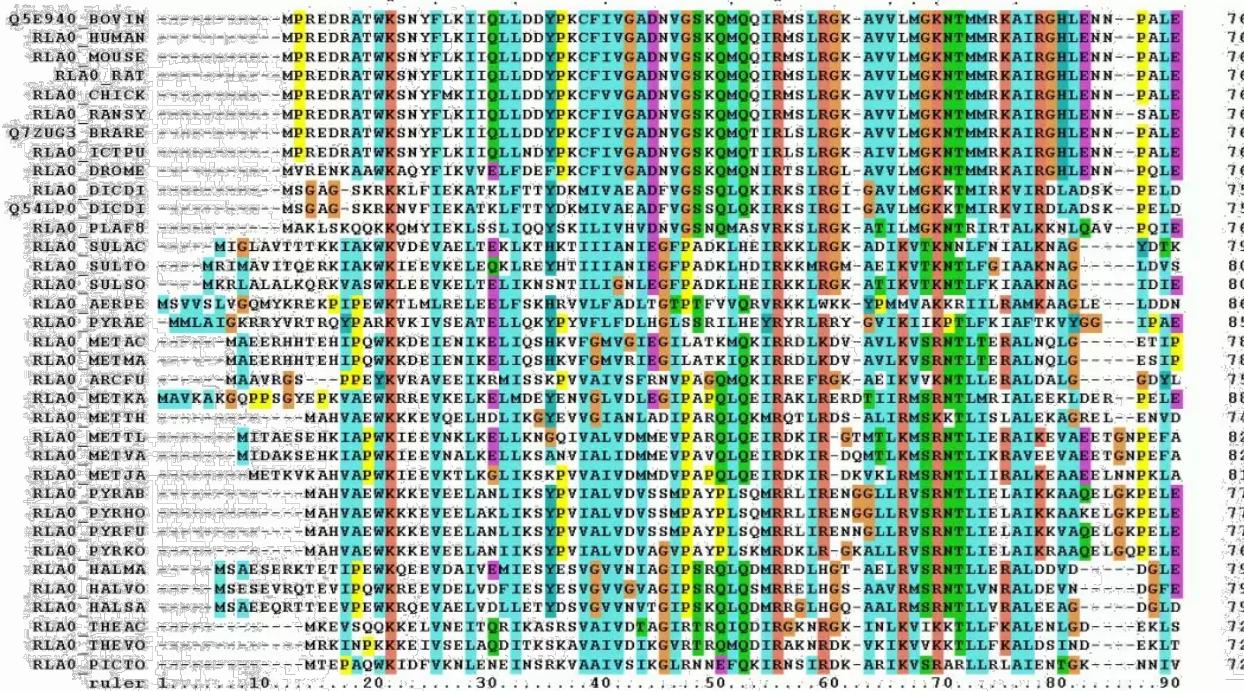


蛋白质中两两氨基酸的距离形成的map
Contact是指氨基酸之间距离小于某个阈值

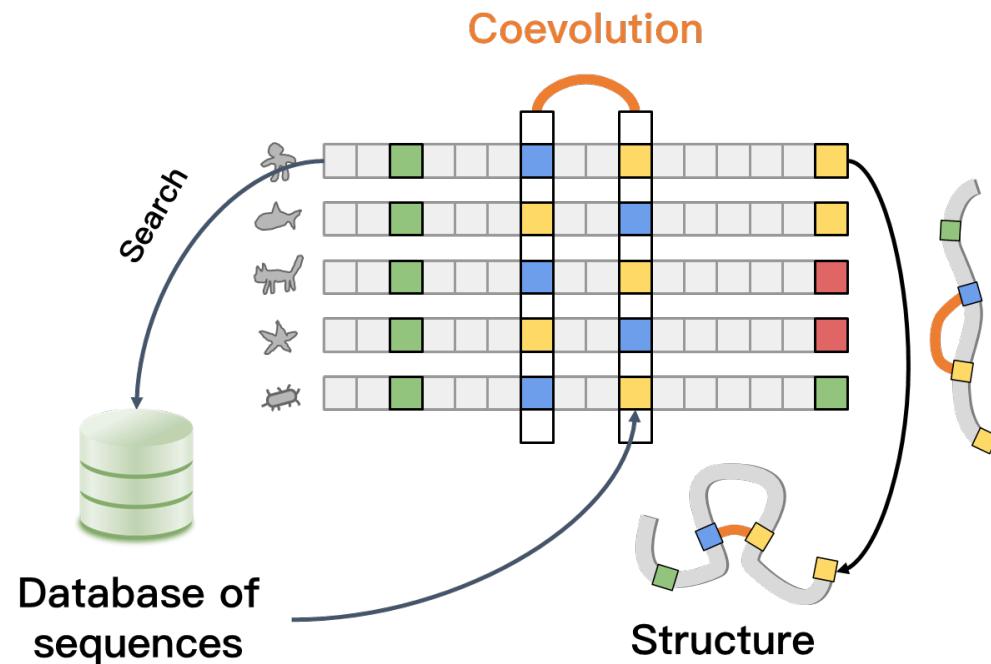


预测蛋白质的Contact Map: 理论基础

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

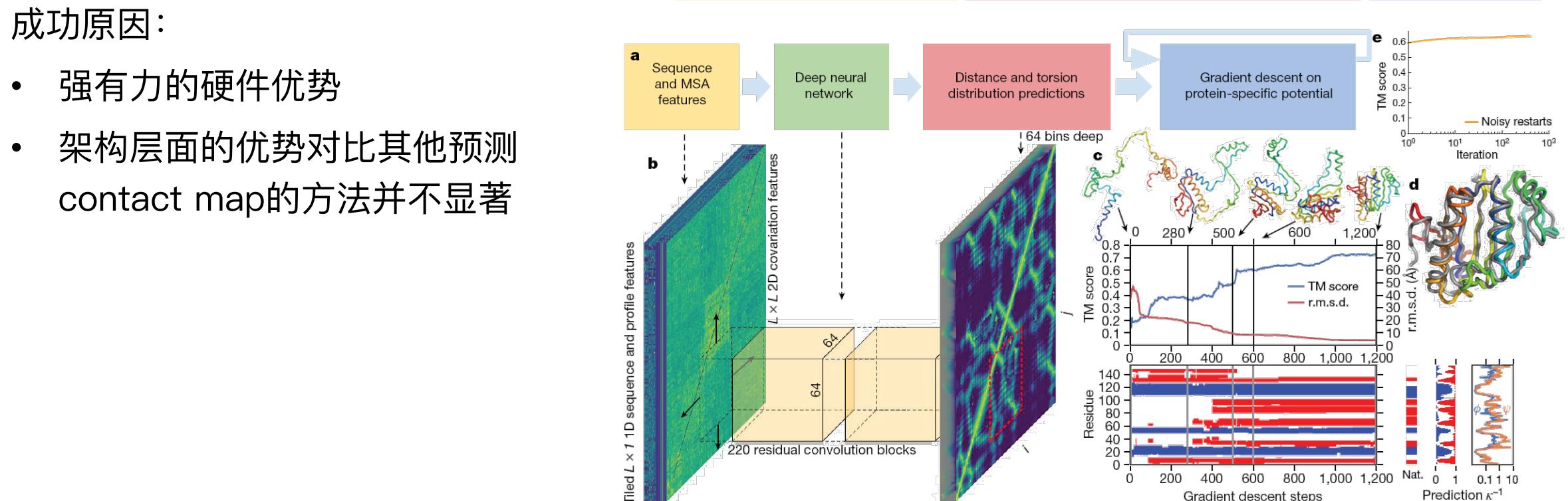
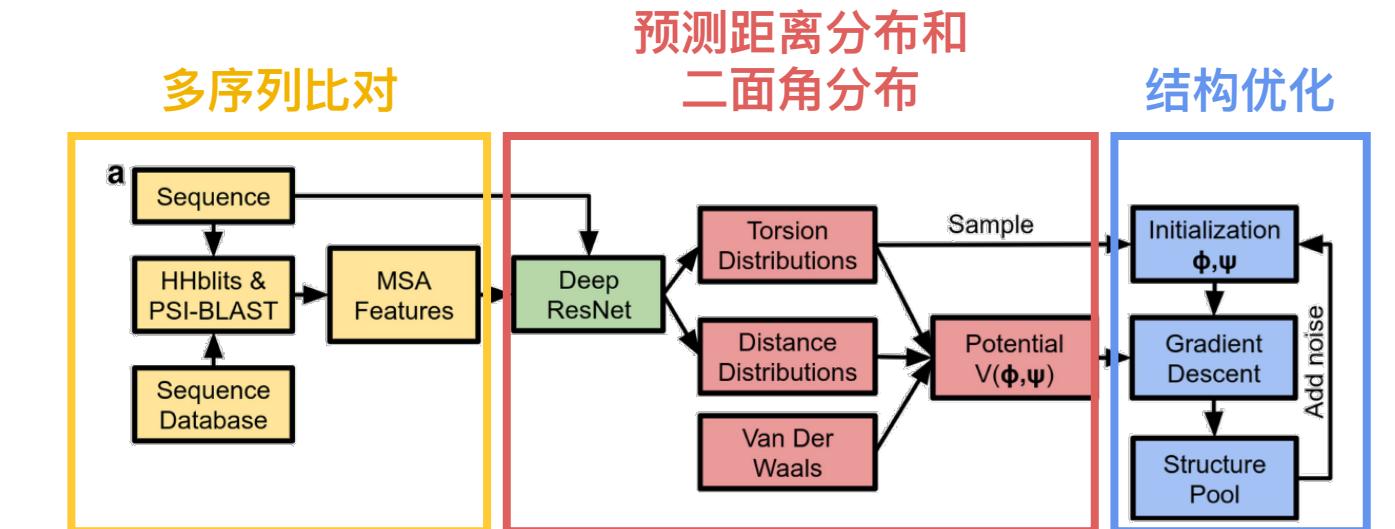
AlphaFold 2018

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：

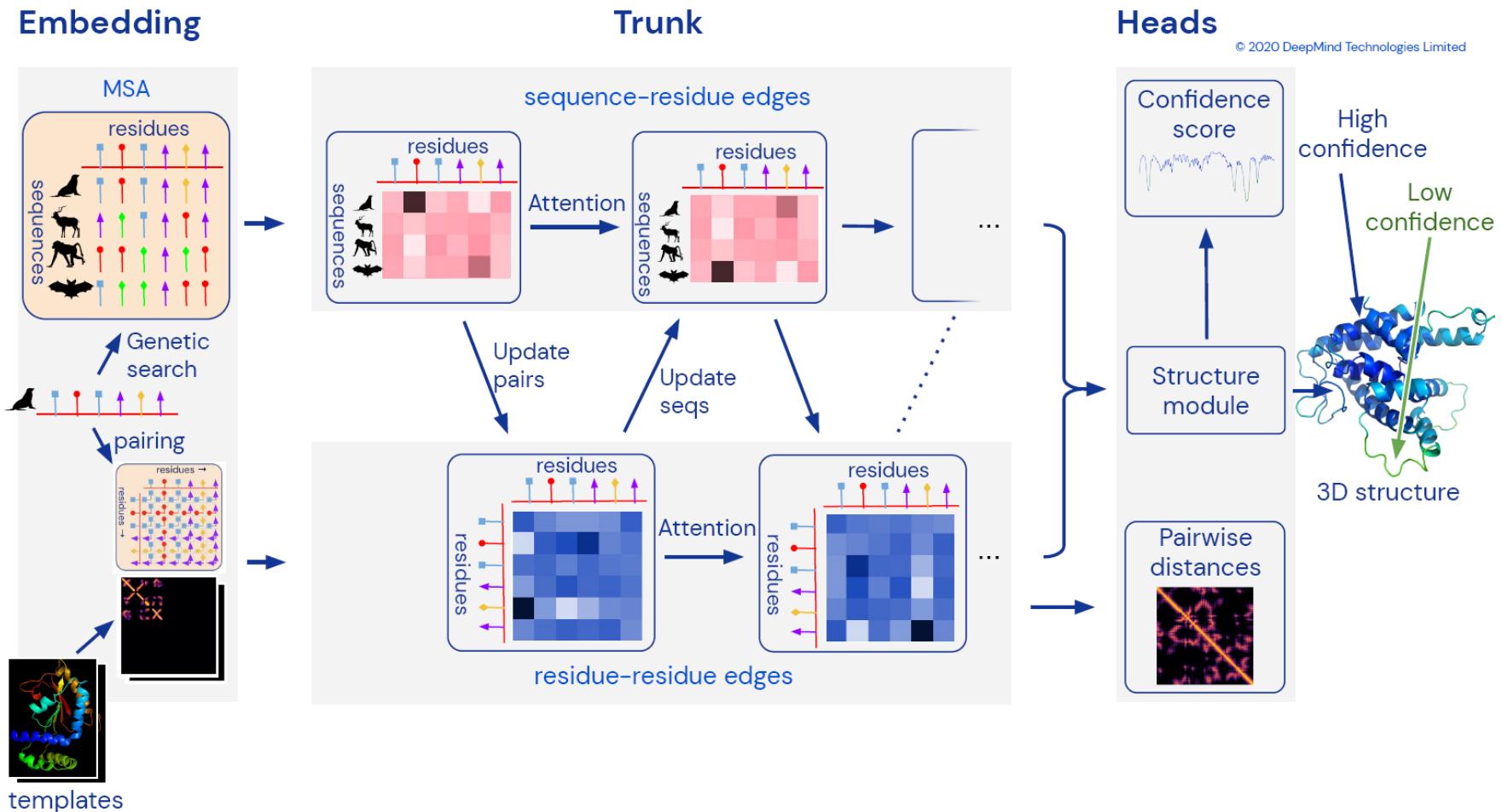
- 强有力硬件优势
- 架构层面的优势对比其他预测 contact map的方法并不显著



AlphaFold2的整体架构

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



整体架构的精彩之一： 模型输入——更强大的MSA & Templates

序列数据库

- UniRef90 (JackHMMER) 来自UniProt
- BFD (HHblits) 自建数据库
- MGnify clusters (JackHMMER) 宏基因组

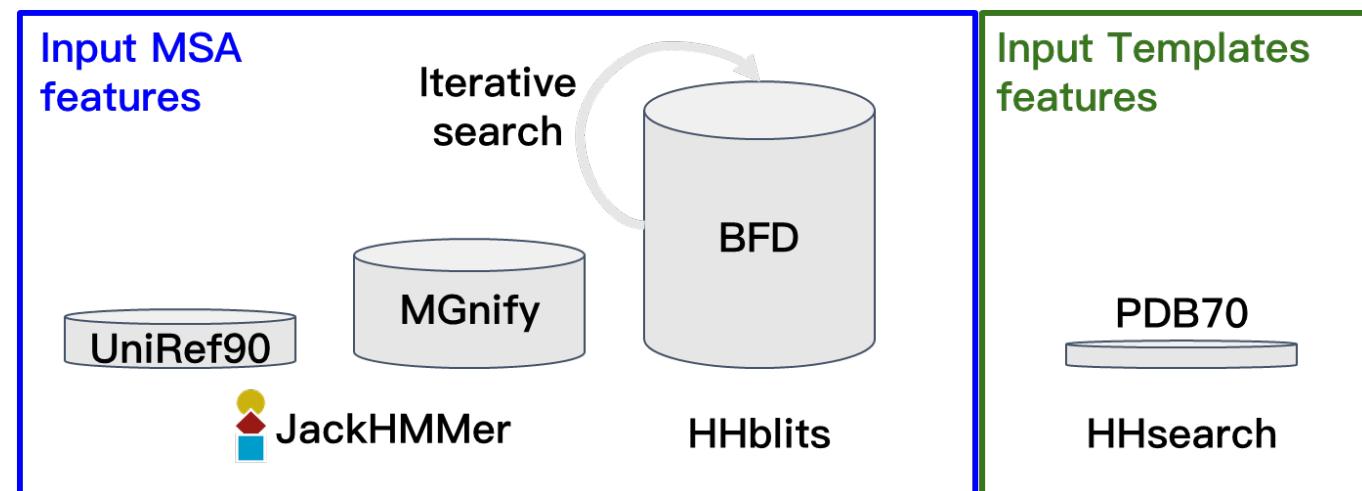
结构数据库

- PDB (用于训练)
- PDB70聚类 (HHsearch)

所有序列数据均来自公开数据库

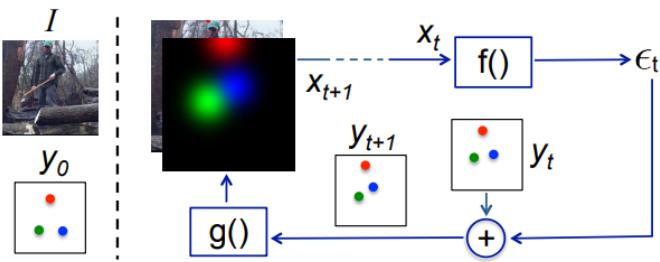
MSA决定结构预测准确度的上限

AlphaFold的序列数据库足够大，MSA结果有一定的保证



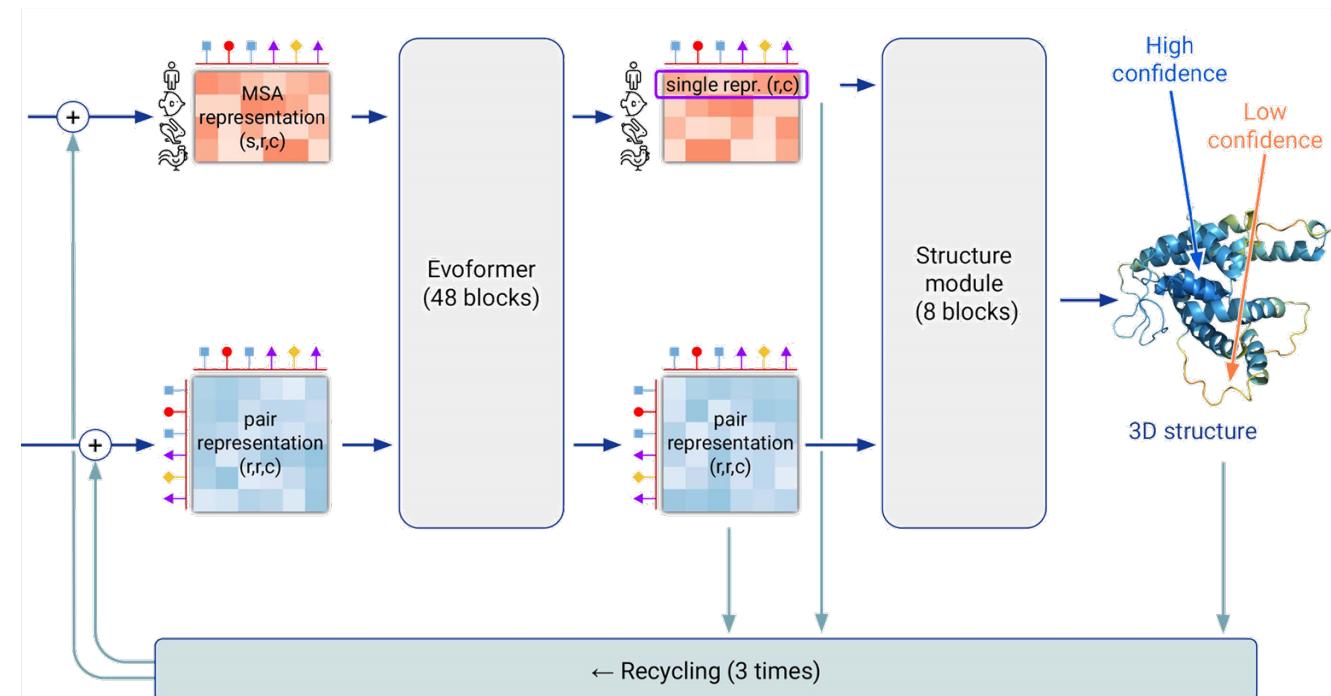
整体架构的精彩之二：

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

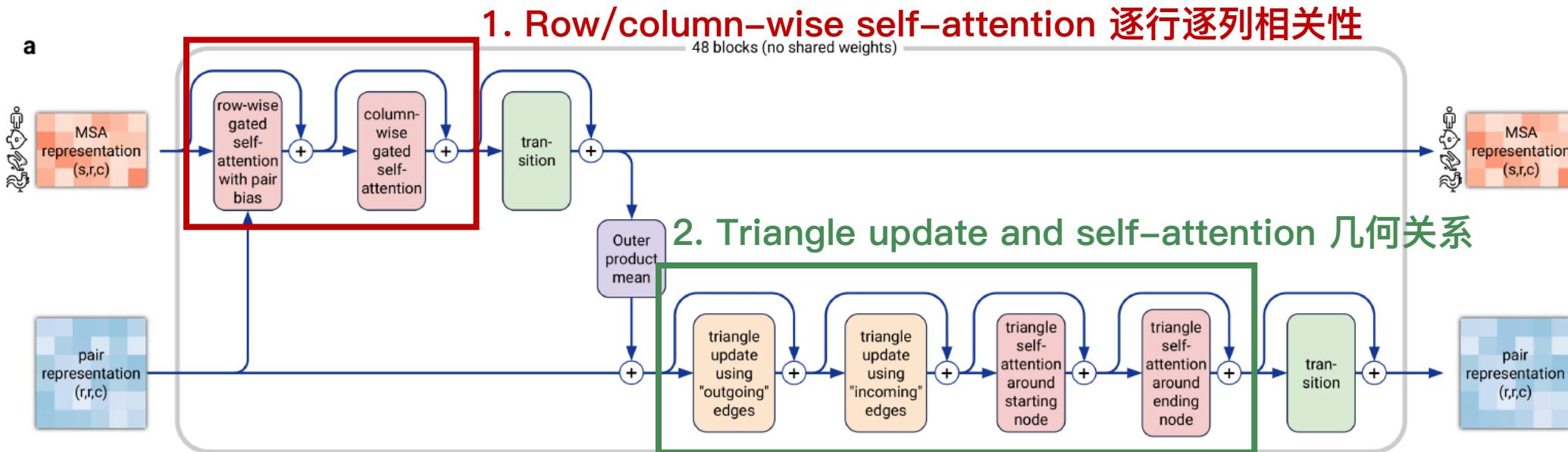
“We find it helpful to execute the network multiple times, each time embedding the previous outputs as additional inputs.”



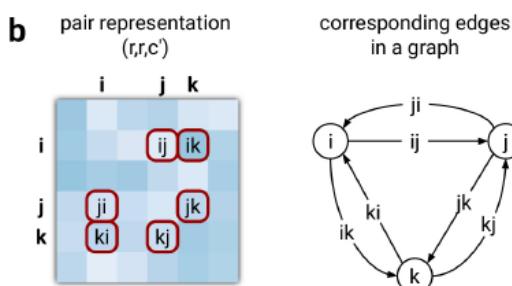
多轮迭代让模型的深度更深，能够通过迭代让结构预测更精确，预测到更复杂的结构

整体架构的精彩之三：

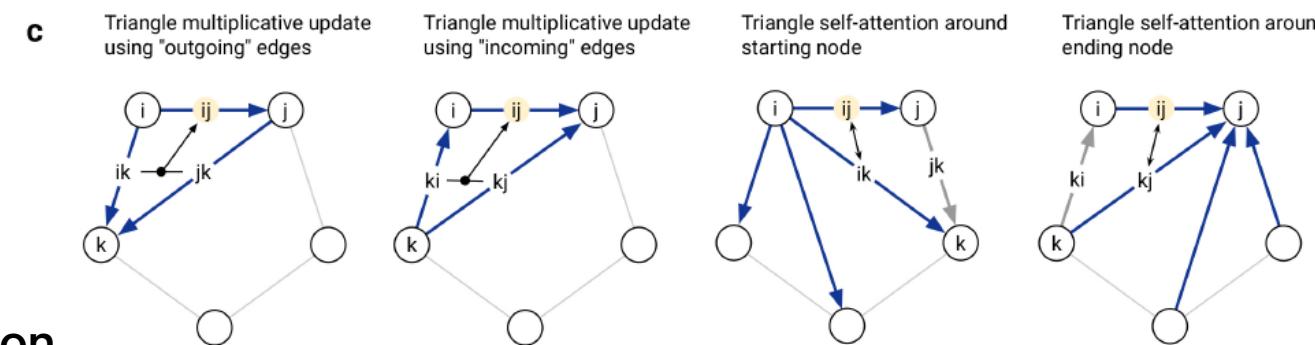
Evoformer: 基于Attention提取进化信息



Evoformer的整体架构



Triangle update and self-attention
利用边之间的三角形关系中互相推断



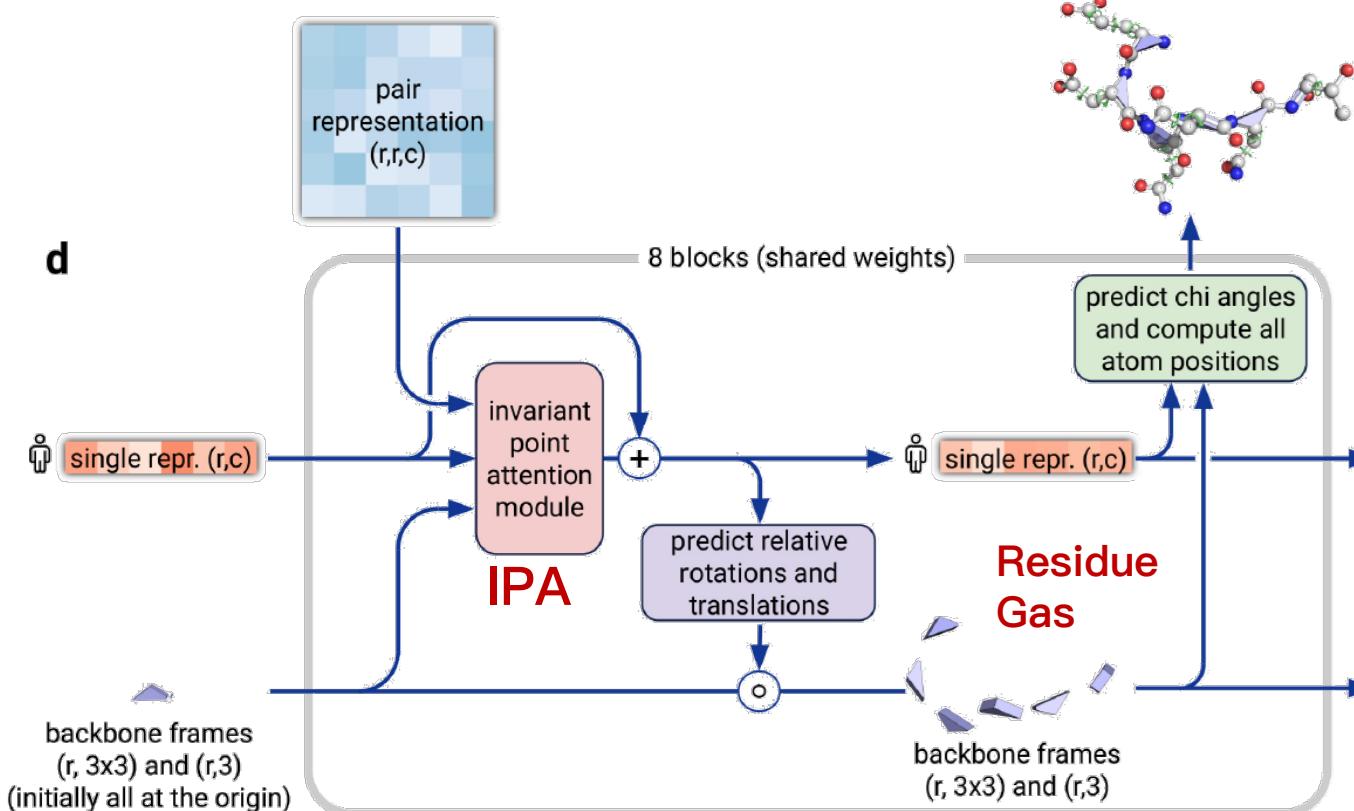
用ik, jk推断ij的信息

用ik推ij的信息，是否接受更新取决于jk边

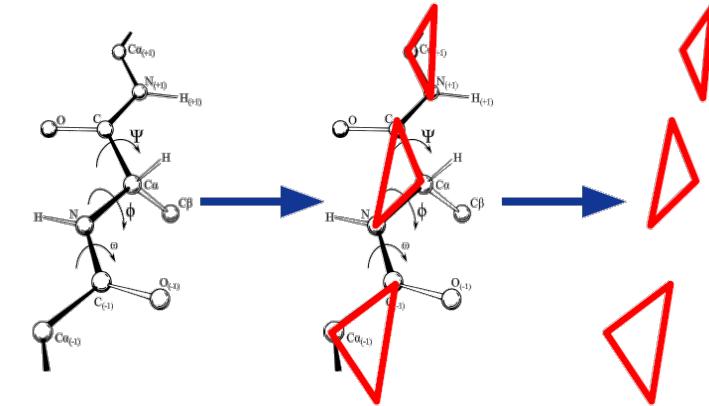
整体架构的精彩之四：

Structure Module的关键——Equivariant

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



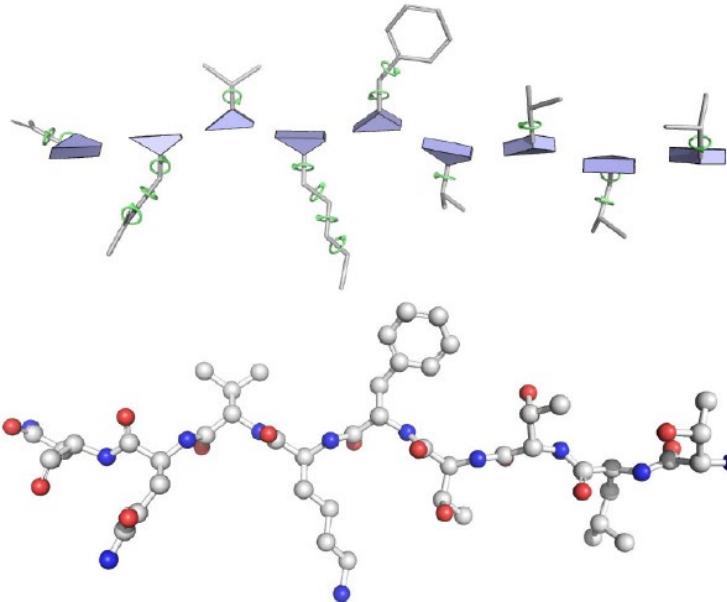
Protein backbone = gas of 3-D rigid bodies



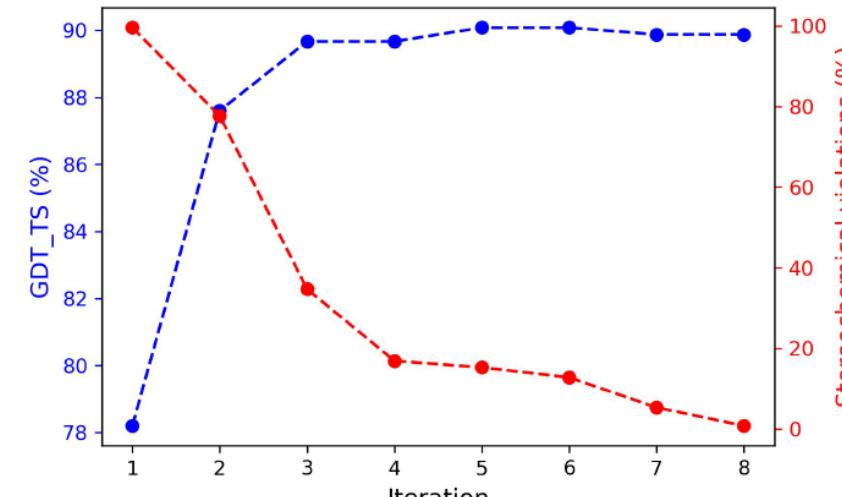
- IPA用于实现3D Equivariant（平移旋转等变性）
- Residue Gas用于表示蛋白质结构
- 输入：
 - 序列信息（目标蛋白）
 - Distance Map信息
 - 蛋白质骨架初始Residue Gas
- 输出：
 - 全原子的位置坐标
 - IDDT-Ca（评估建模精度）

整体架构的精彩之四：

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。

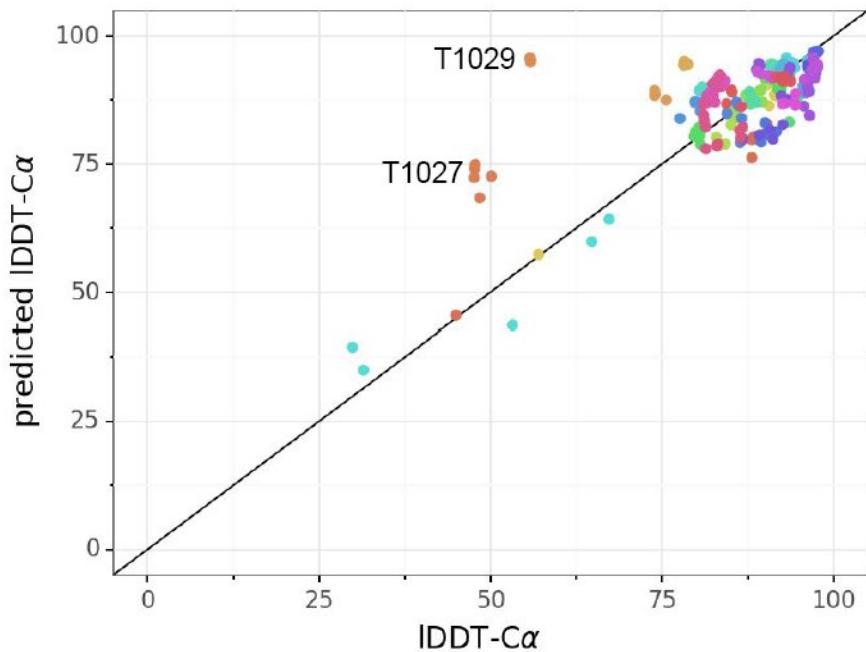


Target: T1O41

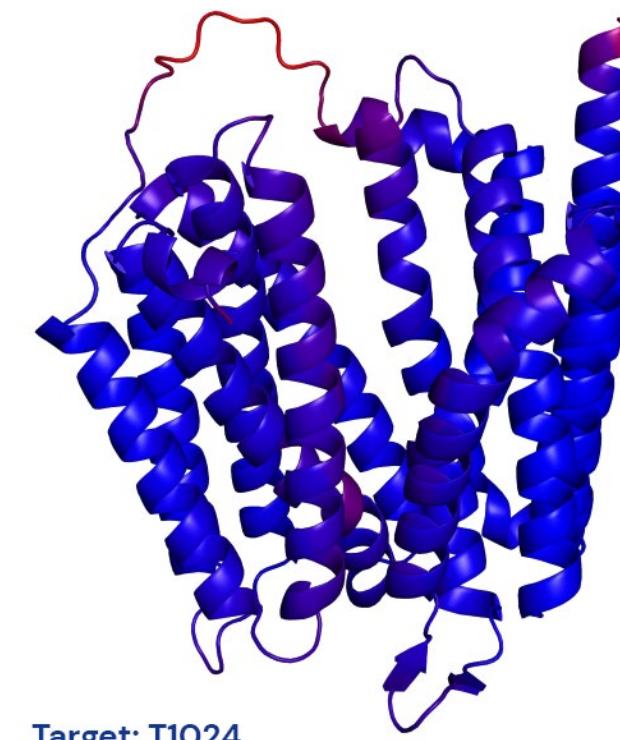
在建模准确性提升的同时，结构中的不合理成分也逐步降低

整体架构的精彩之五： 多输出——如何知道预测的结构的精确度

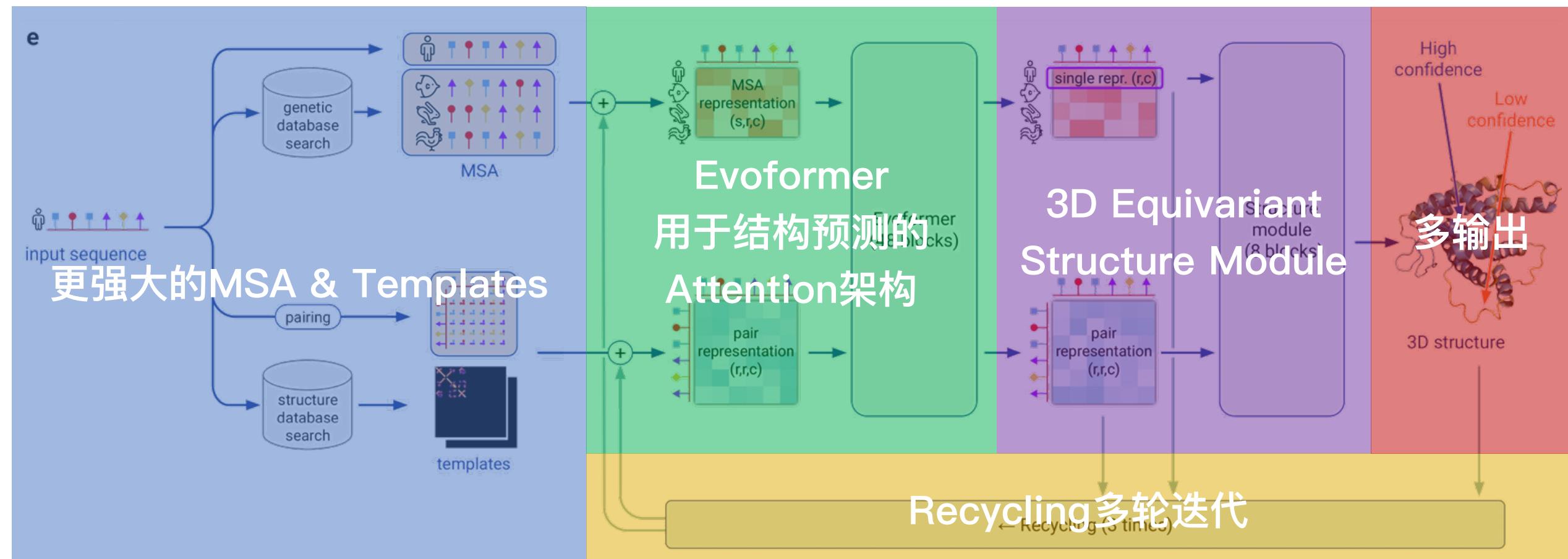
模型预测的IDDT-C α 与实际值十分接近
MAE=3.3



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



再回顾AlphaFold2的整体架构

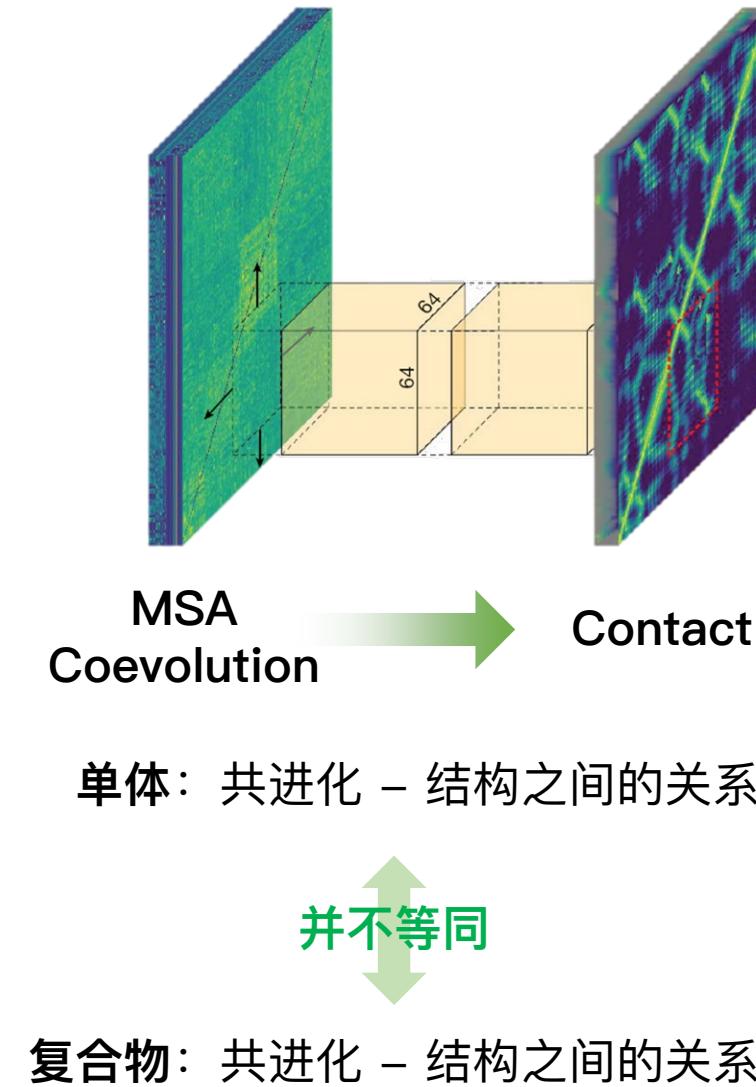
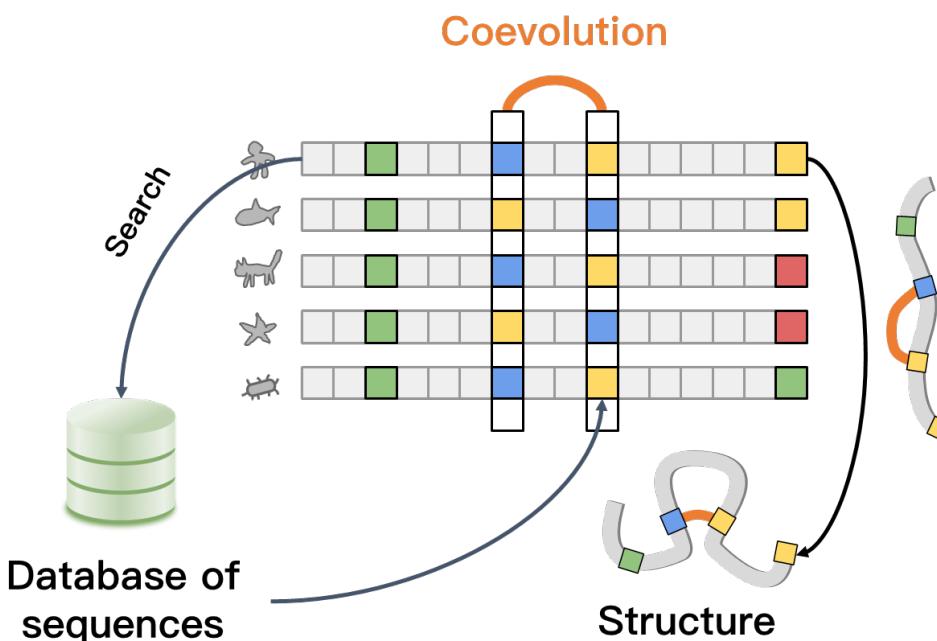


AlphaFold 优点总结和补充

- 基于**recycling**的迭代优化。这一点在很多领域已经得到过应用，比如计算机视觉中的姿态估计 (post estimation)
- 广泛应用的**Attention**架构。将二维的表横着做Attention、再竖着做Attention，对于图可以在局部做Attention，不断精化了Embedding过程；Structure module中也继续用到了Attention
- 实现了端到端(**end-to-end**)架构。完整建立了用于蛋白质结构预测的端到端架构，让模型能够在提升准确度的同时，融合结构的优化步骤。
- 半监督学习拓展训练集 (**Self Distillation**)。用带标签的数据先训练一遍，再用无标签的数据预测一遍形成新的数据集，然后再混合继续训练。这种方法曾经在Google Brain的noisy student使用过，在这里再次得到了应用
- 类似BERT的mask结构。Mask对各种输入添加噪音以增加模型的鲁棒性，这在BERT类模型中非常的常见

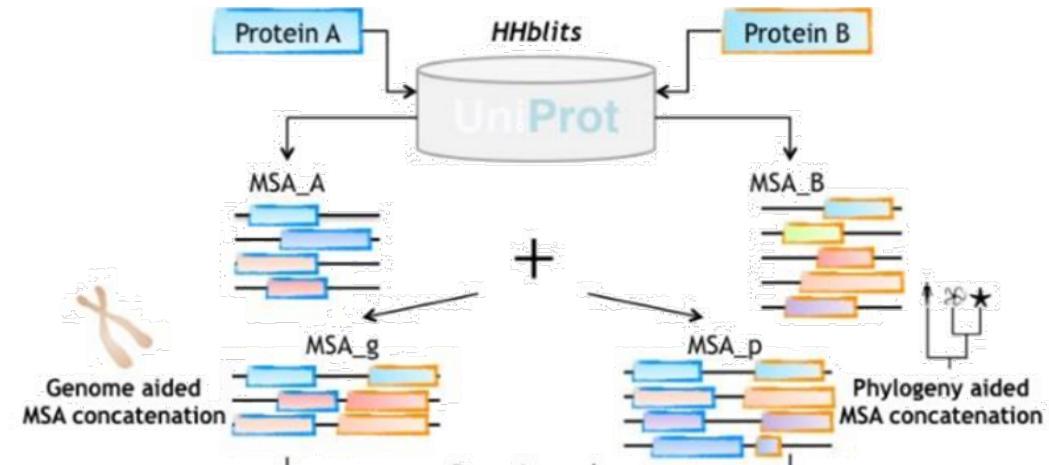
结构预测的本质：从Coevolution到Contact

- 蛋白质结构预测的本质：
 - 从共进化信息推断蛋白质结构的contact
 - 共进化信息并不是物理的作用关系



Multimer模型对输入进行改进：Cross–Chain Genetics

- 主要借鉴了Zhou et al. 文中用于构建MSA的方法
- 使用对应蛋白的物种标签（来自UniProt）



Prokaryotic 原核生物

使用smallest genetic distance配对

Smallest genetic distance:

- 使用UniProt中的accession ID估算

Eukaryotic 真核生物

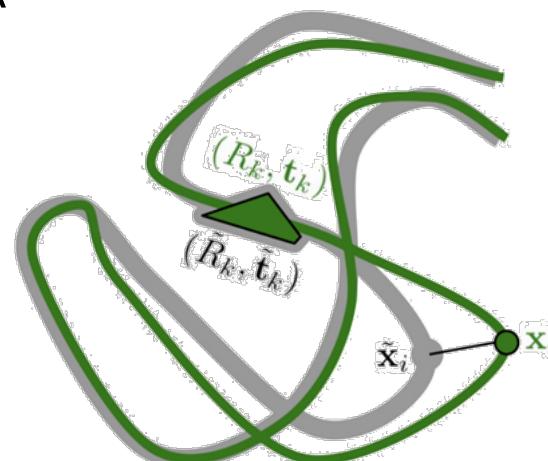
用与目标蛋白的序列相似性排序
相同排序的拼接在一起

Multimer模型对损失函数和判断标准的改进

FAPE loss

在原本使用的FAPE loss上做了修改：

- FAPE: 预测结构和实际结构对每个Residue Gas做叠合，计算其他每对原子的距离(<阈值)
- 原本的计算中有10 Å的阈值
- 考虑到复合物计算的特性：
 - 链内仍为10 Å
 - 链间不设阈值



Model Confidence

原有判据：预测的TM-score (pTM)

增加判据：预测的interface TM-score (ipTM)

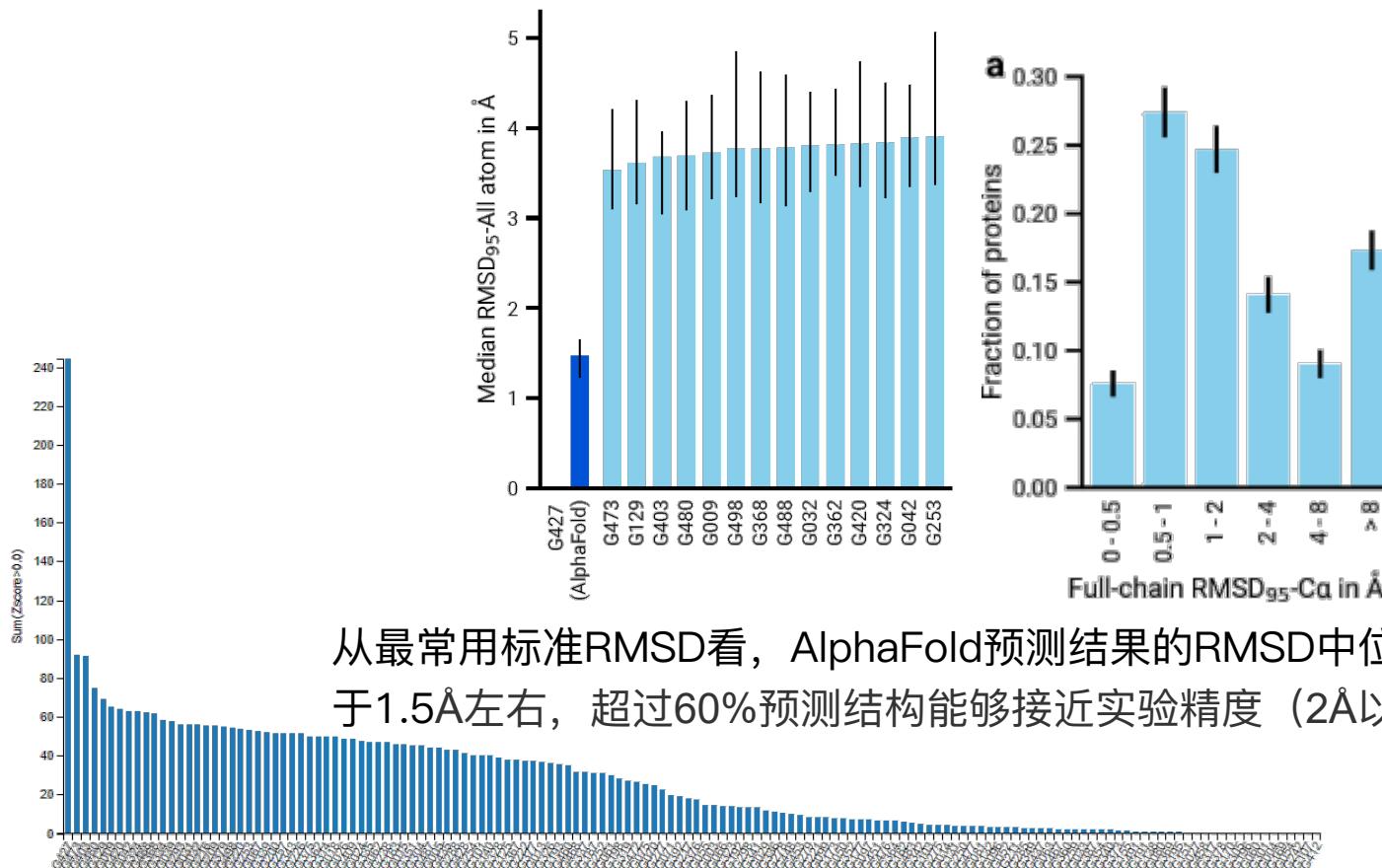
$$pTM(\mathcal{D}) = \max_{i \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \mathbb{E} \frac{1}{1 + \left(\frac{e_{ij}}{d_0(|\mathcal{D}|)} \right)^2}$$

$$ipTM = \max_i \frac{1}{|\mathcal{D}_{-chain(i)}|} \sum_{j \in \mathcal{D}_{-chain(i)}} \mathbb{E} \frac{1}{1 + \left(\frac{e_{ij}}{d_0(|\mathcal{D}_{-chain(i)}|)} \right)^2}$$

$\mathcal{D}_{-chain(i)}$ 所有除了Residue i 所属链以外的链

$$\text{model confidence} = 0.8 \cdot ipTM + 0.2 \cdot pTM$$

CASP14中AlphaFold2的碾压性优势

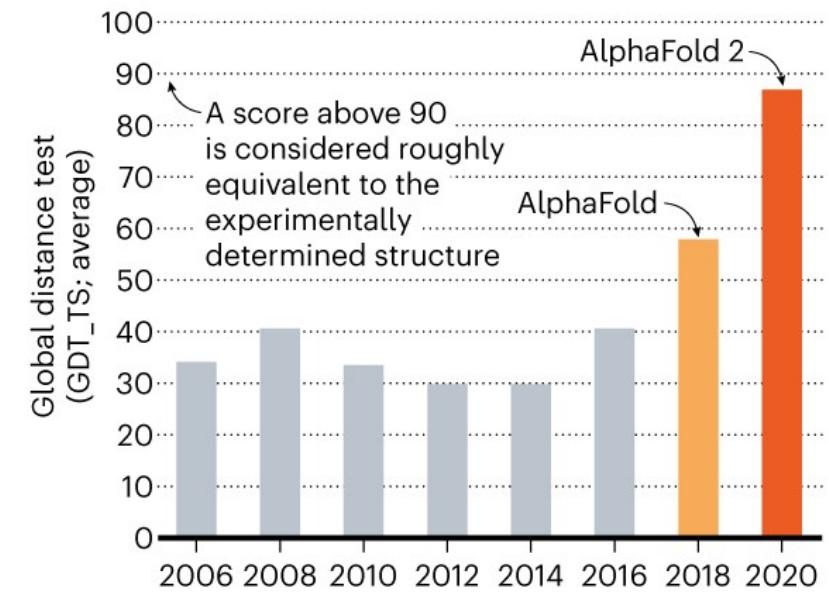


本届CASP14比赛比赛中，Alphafold的GDT-TS总分取得了碾压性的优势，远超第二名的BAKER团队

精确度的显著提升让AlphaFold2成为了突破性的成果，让科学家有足够的信心接受其预测结果

STRUCTURE SOLVER

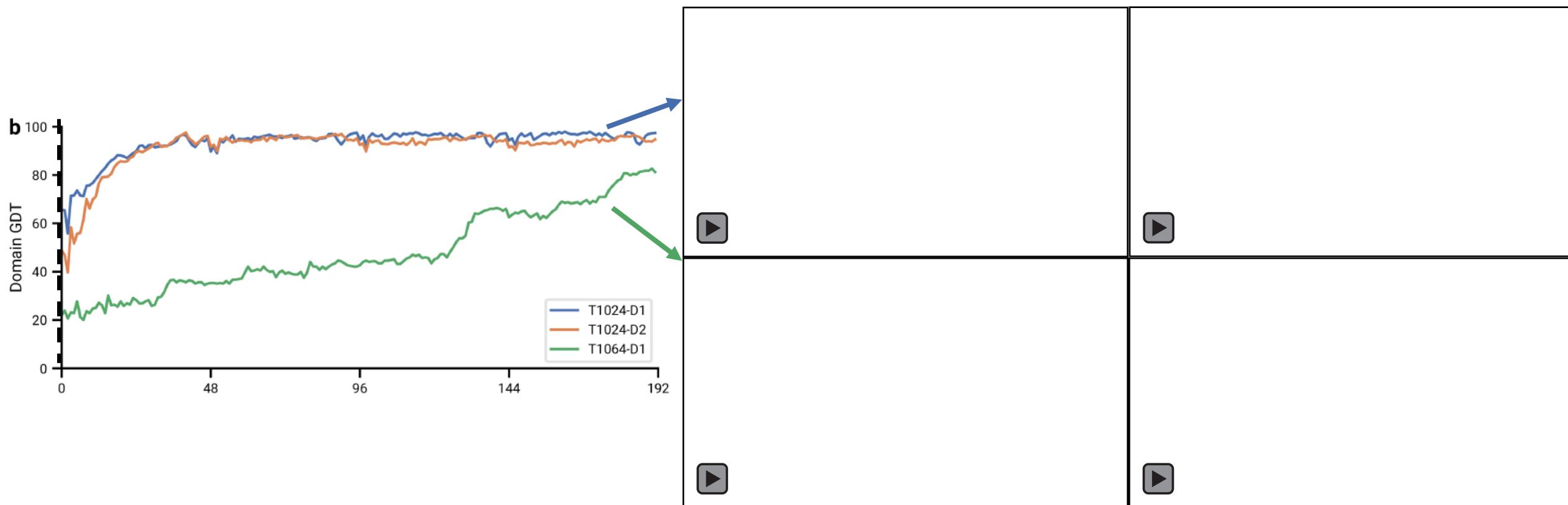
DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature

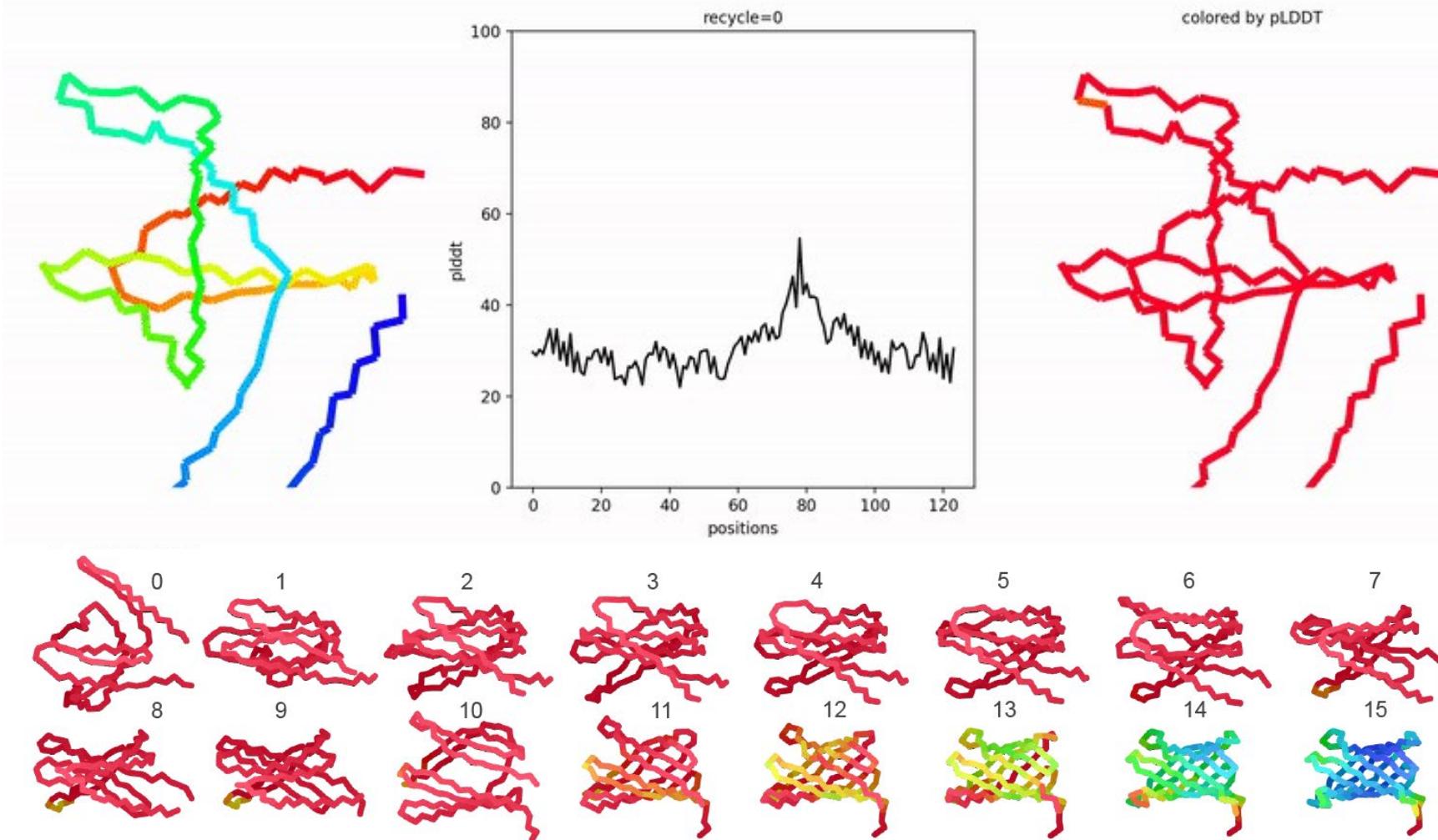
对比历届CASP的结果，AlphaFold的提升是非常显著的

Recycling的必要性：有的蛋白很快就能够折叠，有的蛋白很慢



多轮迭代优化有一定的必要性，较为复杂的蛋白可能在优化流程最后
(4轮优化) 才能折叠到正确的结构

更多轮的Recycling迭代：ColabFold中的实验



更多轮的迭代能够最大程度优化蛋白质结构到最稳定的构象

MSA深度和模板的选择

我们需要何种质量的MSA?

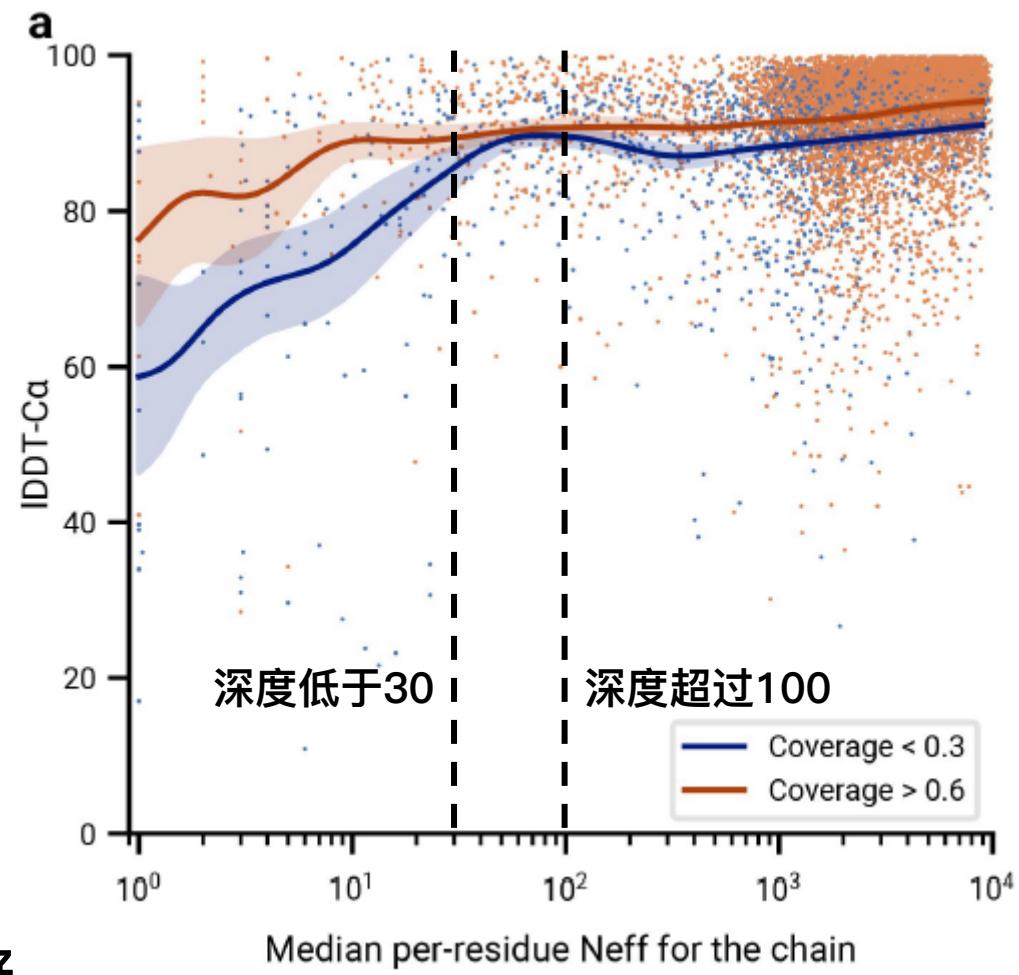
对于MSA的深度

- MSA平均深度需超过30才能取得较好的预测效果
- MSA深度超过100的则提升并不显著

对于Template的覆盖率

- MSA深度低于30时，模板的相似度才会有比较大的对准确度的影响（高相似度模板优于低相似度模板）

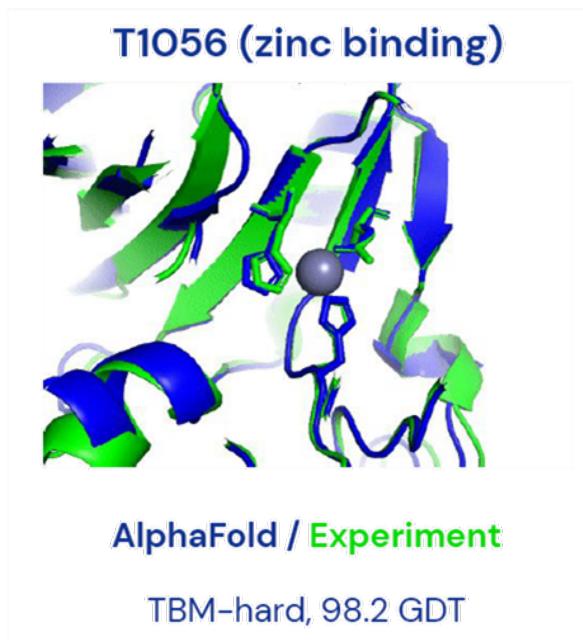
MSA做的够好的话，没有很好的模板也能跑的很好



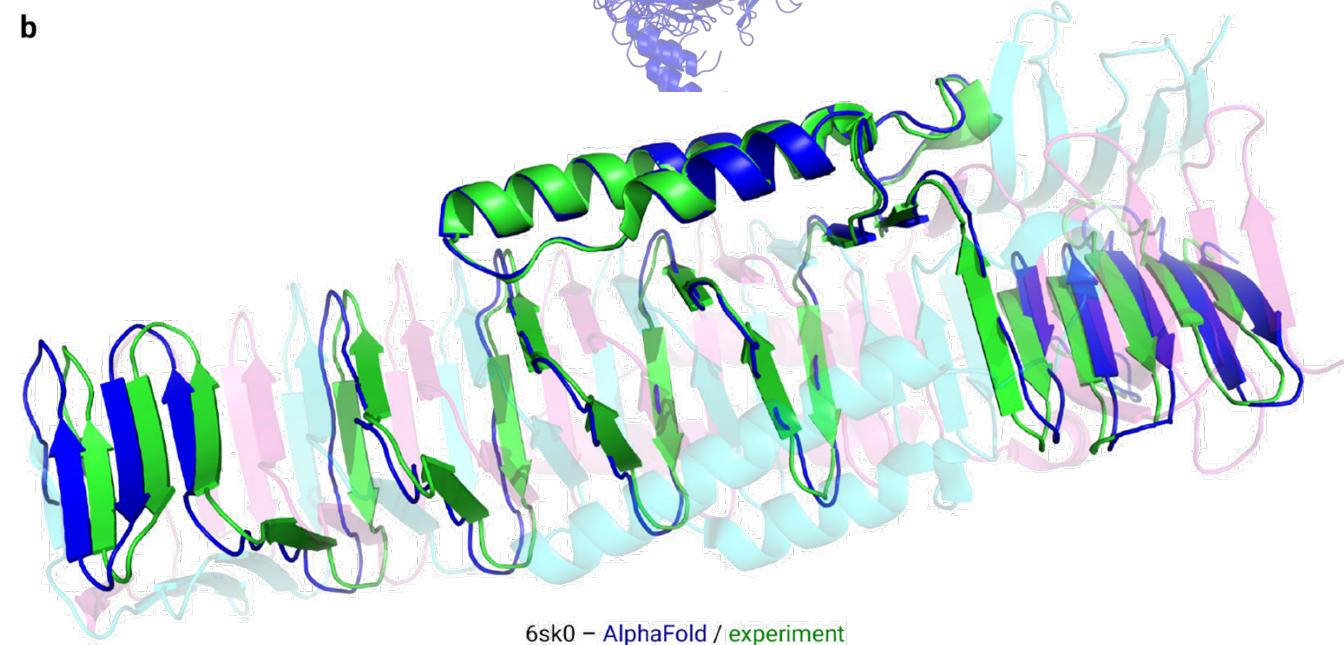
Neff: Normalized number of effective sequences

意外之喜：复杂的蛋白也是可以预测的

- 计算结构预测通常是不明确的
 - 低聚态，配体，DNA结合，实验条件，多种构象等情况
- 我们的网络使用了各种物理和进化信息，隐含地建模了缺失的部分，使预测结果仍然十分准确



含有金属离子的体系

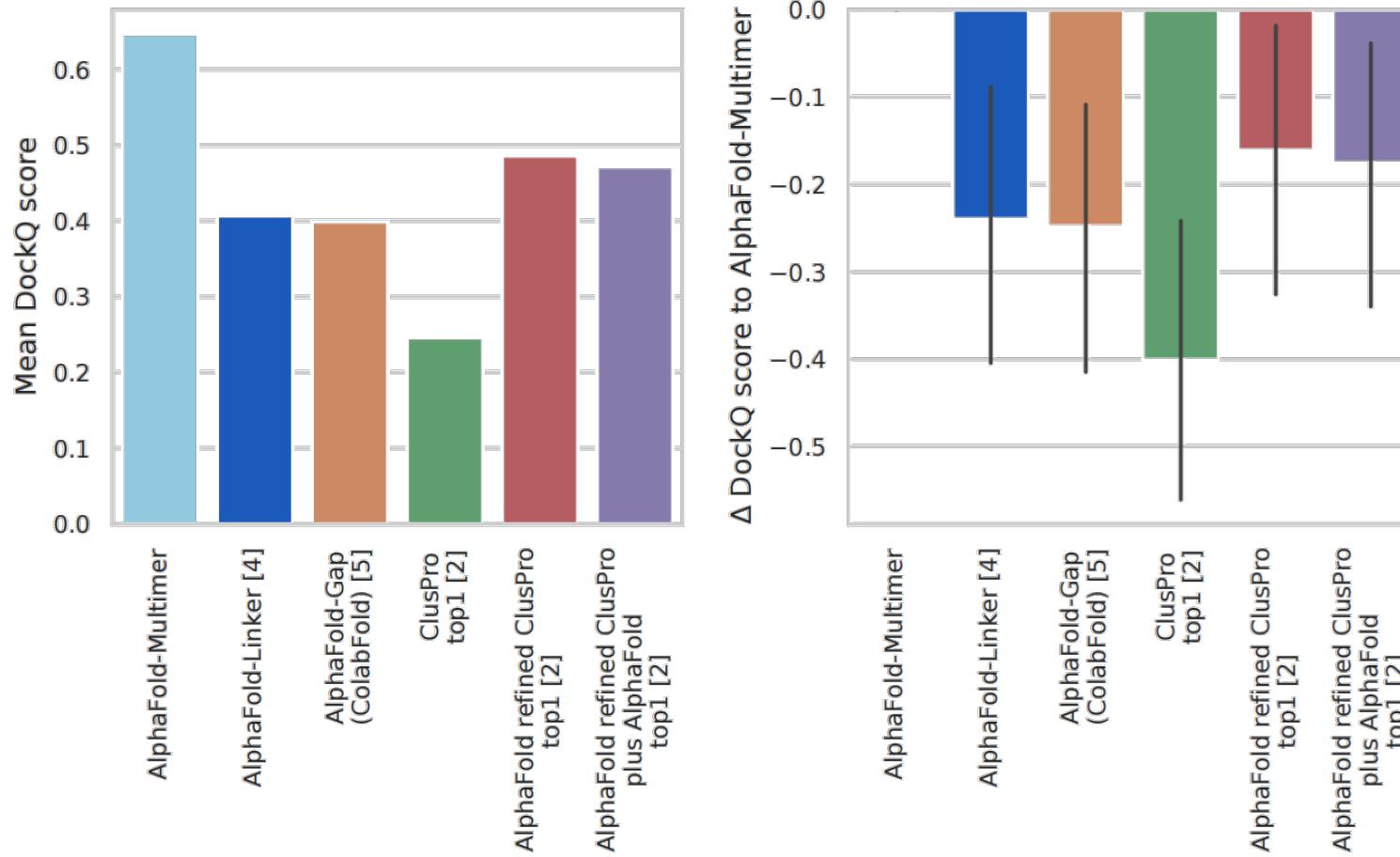


寡聚体蛋白质体系

Concern: AlphaFold学习了从序列到晶体结构的映射，而晶体结构并不能代表真正的蛋白质构象

Multimer模型的结果：DockQ 分数情况

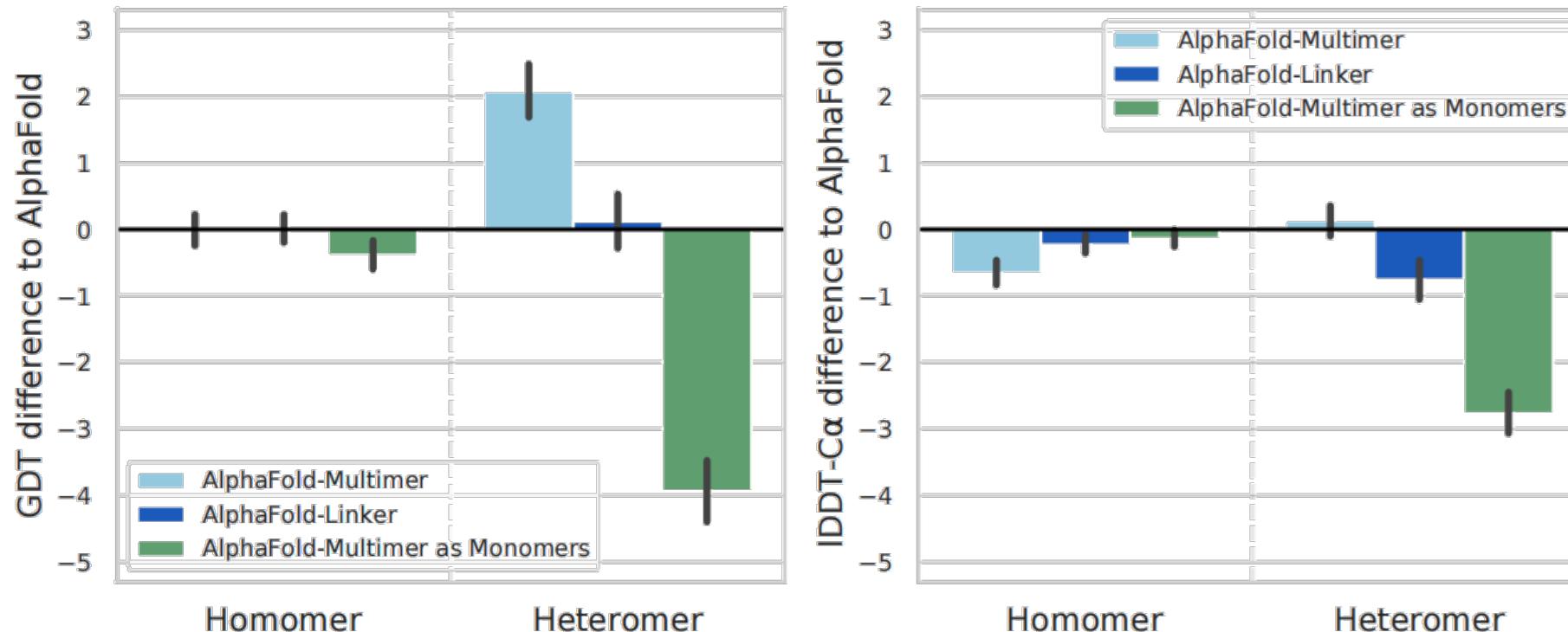
DockQ评估蛋白质界面上的预测质量，位于0~1之间，>0.8意为质量较高，<0.23为预测错误



DockQ评估结果显示AlphaFold–Multimer优于Linker或者Gap的方式

能把Multimer模型用来预测Monomer吗？

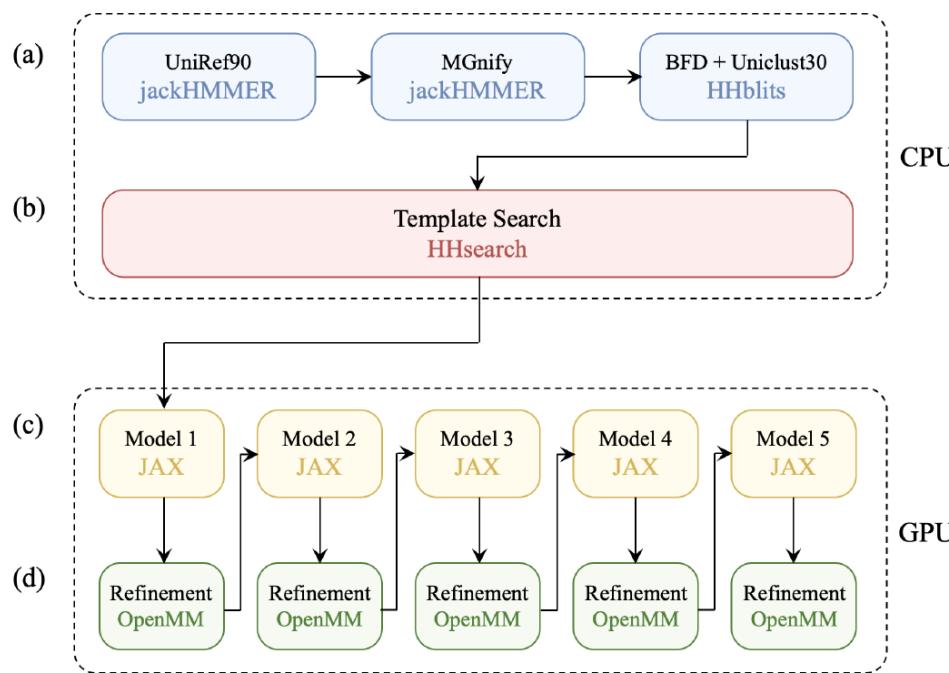
Multimer模型会比原来的AlphaFold更准吗？并不会



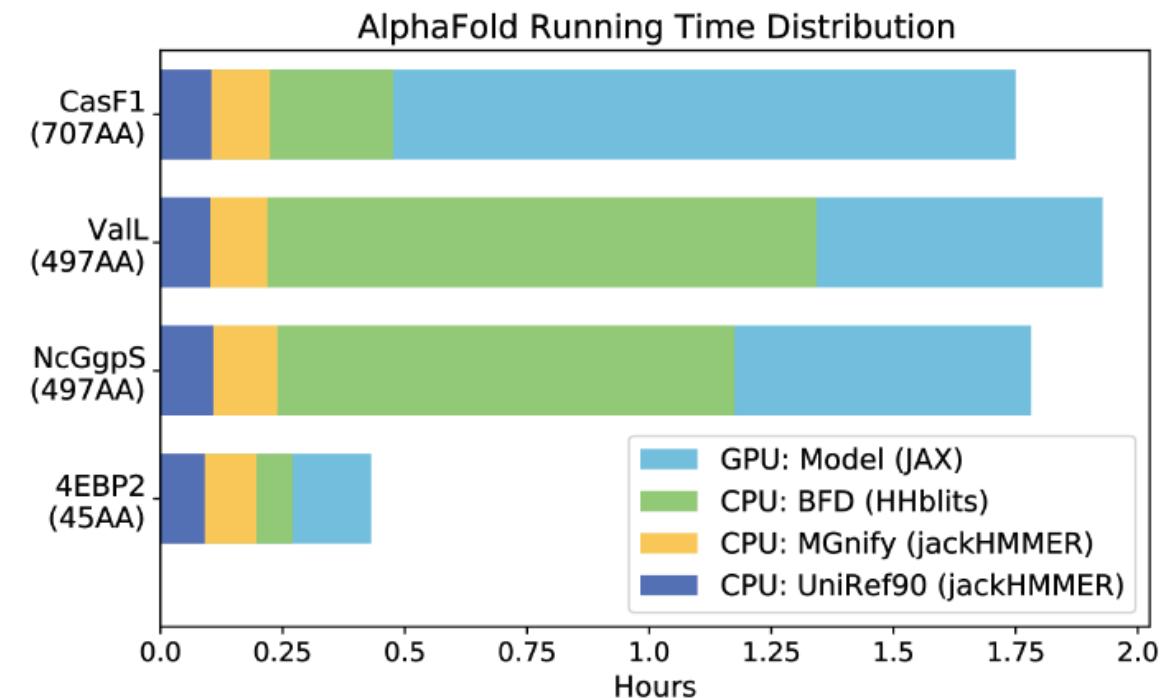
- AlphaFold–Multimer/AlphaFold–Linker: 先预测Complex结构，取其中的一个单链
- AlphaFold–Multimer_as_Monomer/AlphaFold: 输入就只有单链，得到单链的结构

AlphaFold流程：子任务串行执行

串行流程延缓预测速度，在GPU结点完成全部计算浪费GPU资源



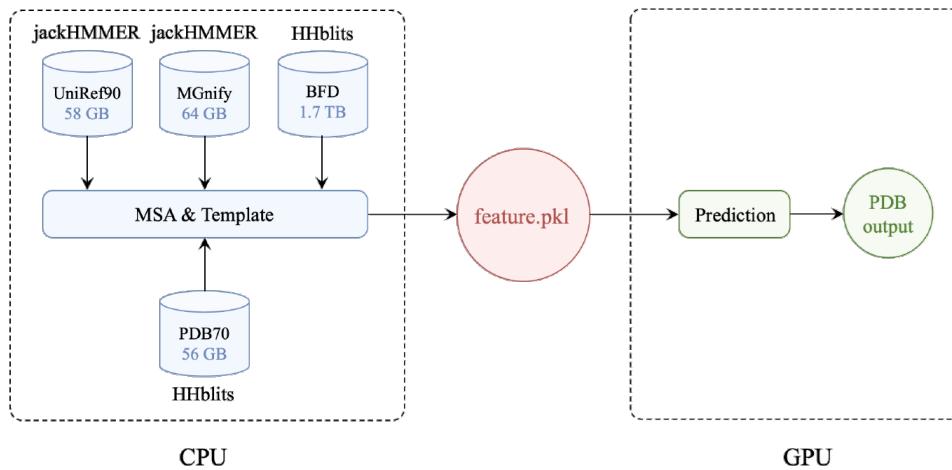
AlphaFold结构预测采用串行流程



大部分预测所用时间浪费在CPU计算上

CPU每个蛋白算1~2小时， GPU一个模型5~10分钟

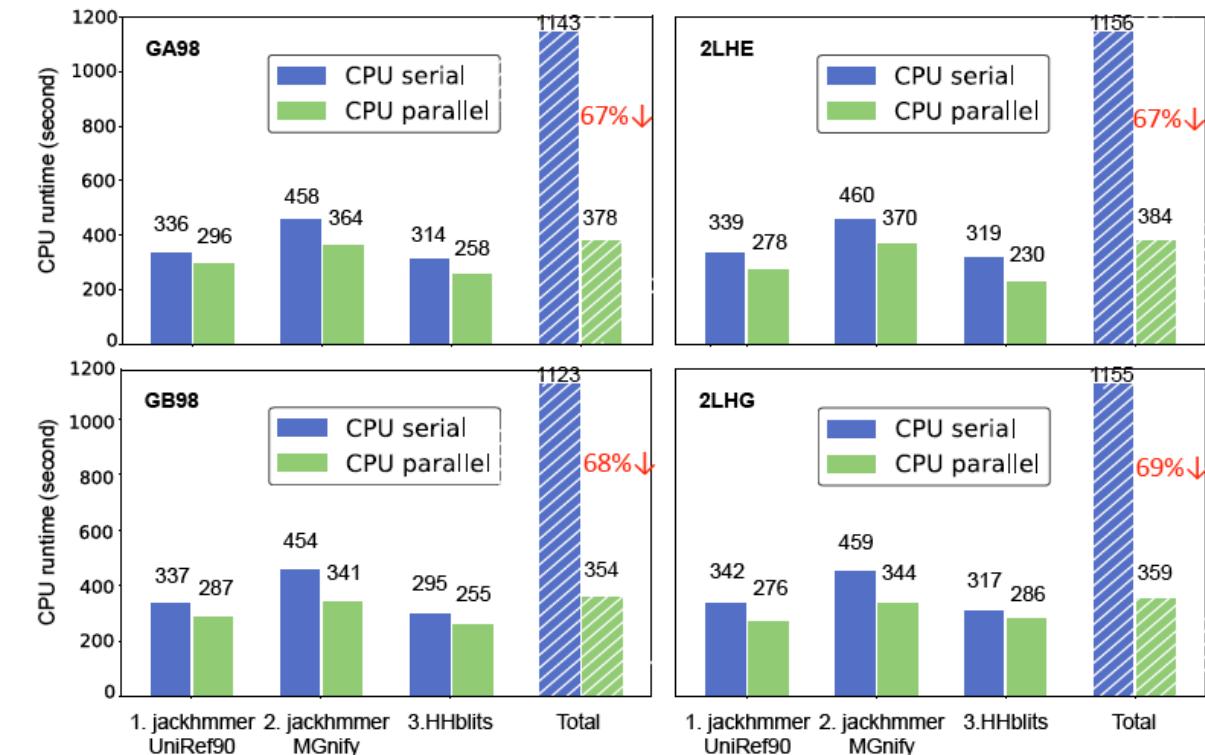
ParallelFold流程：CPU与GPU计算分离



拆分CPU与GPU部分

分拆CPU与GPU部分的计算

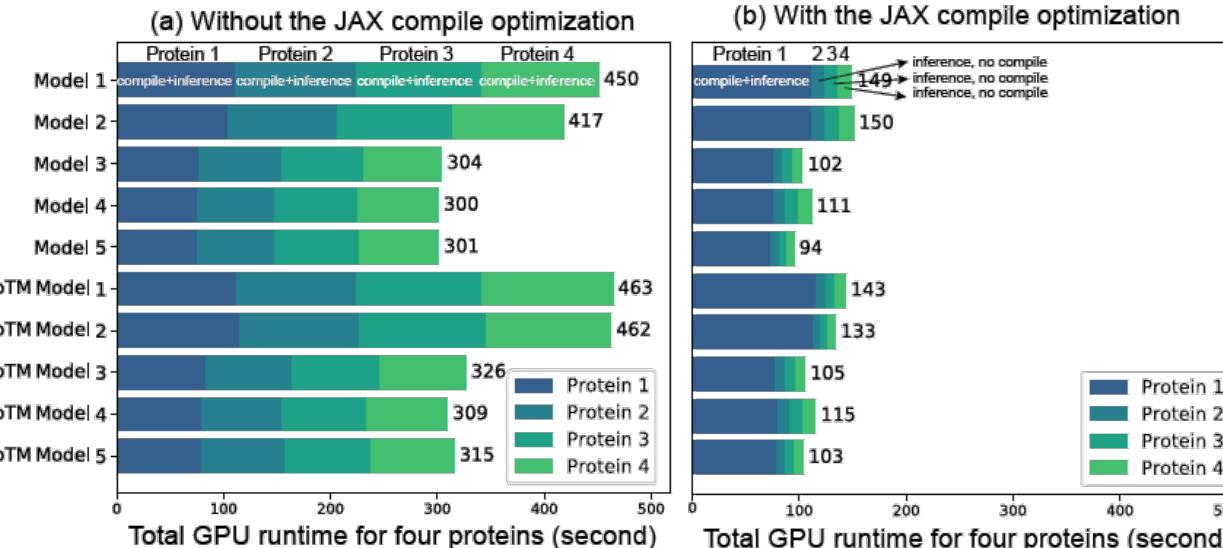
- CPU负责做MSA和模板搜索
- GPU负责神经网络部分



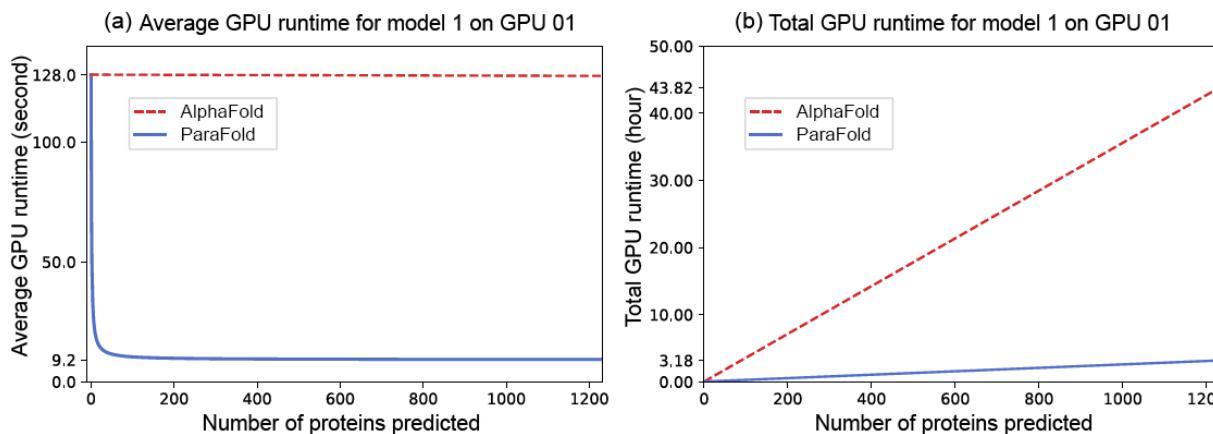
Python多线程并行优化MSA速度
测试蛋白长度56aa

ParallelFold流程：CPU与GPU计算分离

Ten models



GPU编译优化：多次编译到一次编译



Device	Number of Proteins	Total runtime (hour)	Average runtime (second)
GPU 01	1,232	3.9	11.5
GPU 02	1,232	5.4	15.7
GPU 03	1,232	3.9	11.5
GPU 04	1,232	5.1	14.8
GPU 05	1,232	3.9	11.5
GPU 06	1,232	5.0	14.7
GPU 07	1,232	3.8	11.2
GPU 08	1,232	5.0	14.6
GPU 09	1,232	5.3	15.5
GPU 10	1,232	5.0	14.5
GPU 11	1,232	3.9	11.5
GPU 12	1,232	5.0	14.7
GPU 13	1,232	5.2	15.2
GPU 14	1,232	5.1	14.8
GPU 15	1,232	5.1	14.8
GPU 16	1,224	5.0	14.8
Total on DGX-2	19,704	5.4	13.8

De Novo蛋白数据集：19,704个小蛋白用时5.4小时

仅含GPU部分，蛋白长度50aa，使用16*V100 (DGX-2 in SJTU HPC)

以50aa的miniprotein为例（1000+个任务）

- 初始流程：CPU+GPU 1352.62h
- 第一轮优化：GPU 43.82h
- 第二轮优化：GPU 3.18h



ParaFold 提升高通量结构预测效率

在看实例之前需要知道的：如何评估预测结果

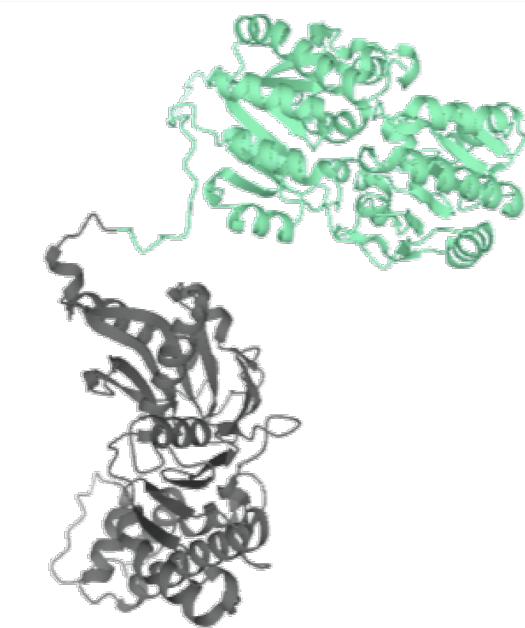
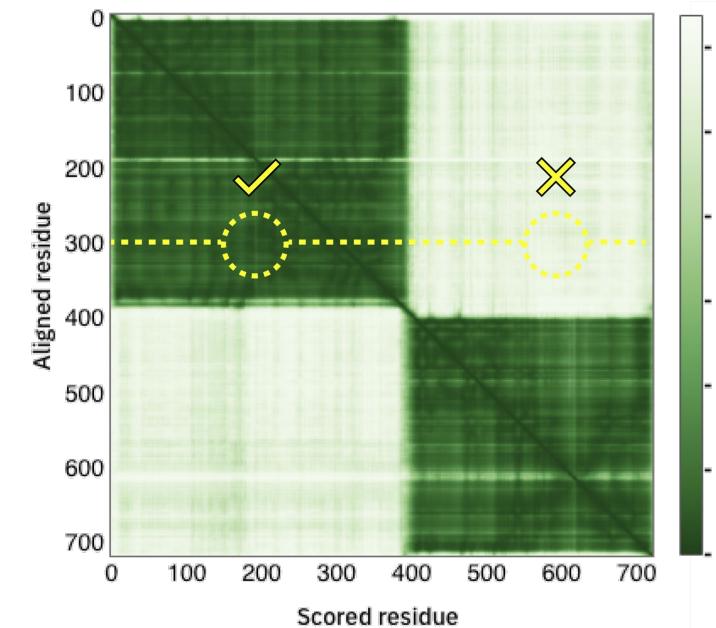
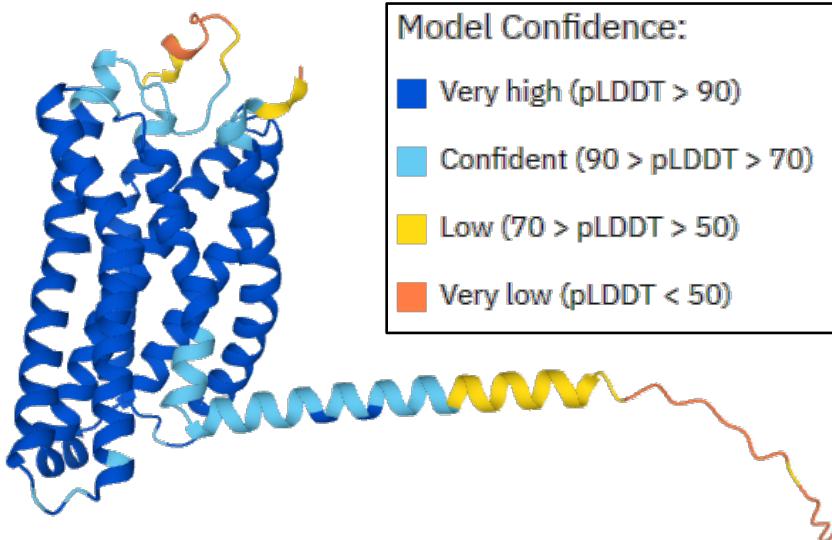
最常用的用于评估AlphaFold建模精度的指标：

pLDDT：反应每个氨基酸的预测精确度：

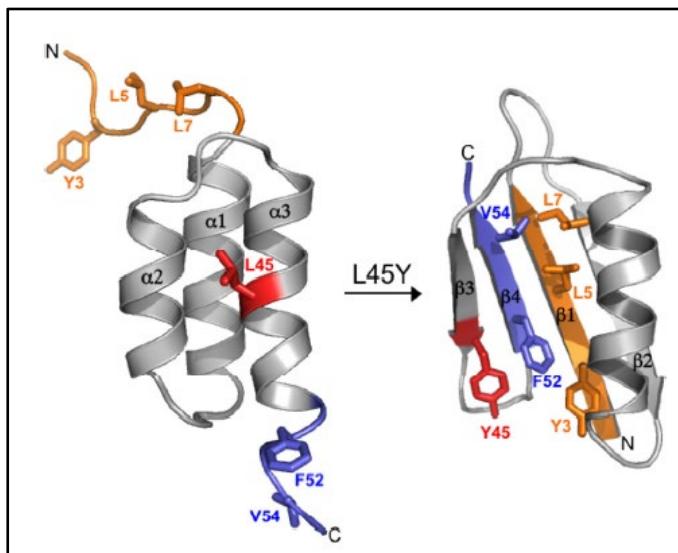
- 低于50表示非常不准确或为无规蛋白
- 低于70表示可信度低
- 高于90可信度极高

PAE：反应每对氨基酸距离与真实值的预测差值：

- 越低越好，单位为Angstrom
- 在反应多Domain或Complex结构精度是效果很好
- 只有pTM模型和Multimer模型才有PAE打分

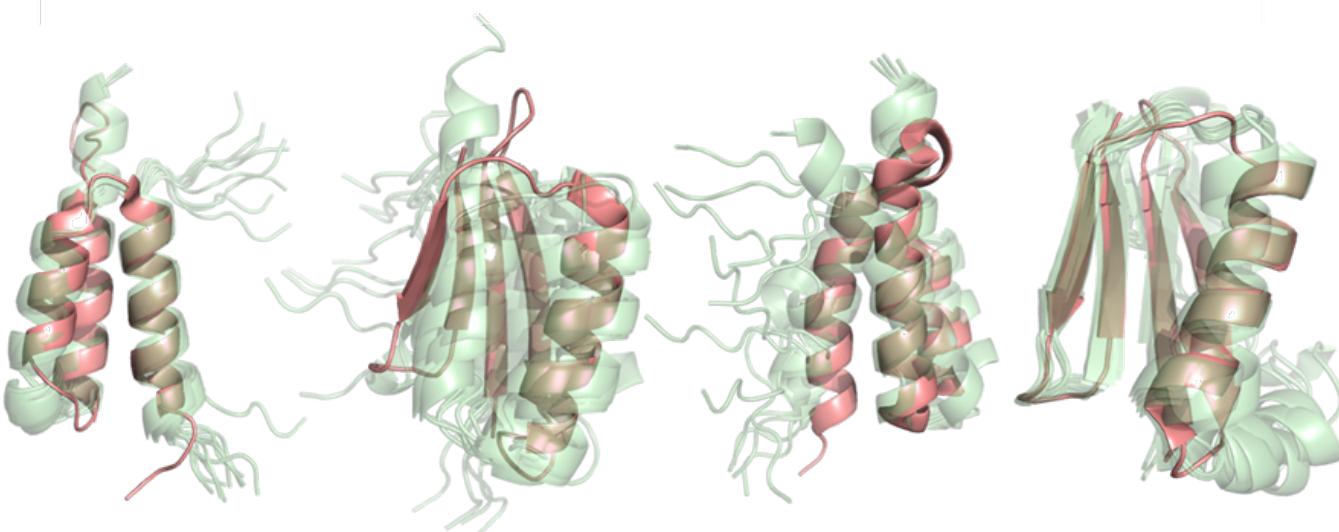
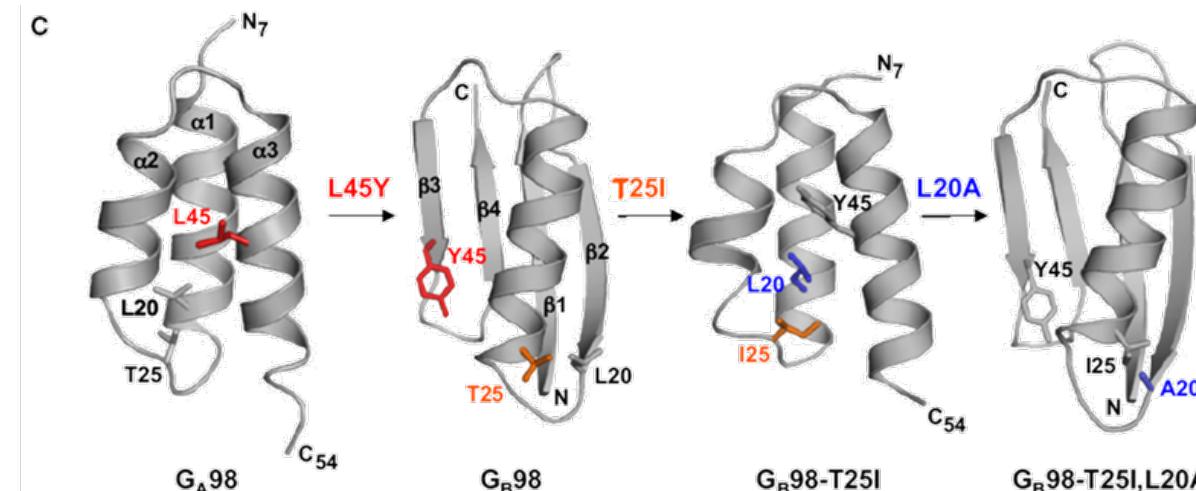


预测突变造成结构变化的实例：GA98 & GB98



	1	10	20	30	40	50
G _A 77	TTYKLILNLKQAKEEAIKE	LVDAGIAEKYIKL	I	ANAKTVEGVWTLKDEI	KATVTE	
G _A 88	TTYKLILNLKQAKEEAIKE	LVDAGIAEKYIKL	I	ANAKTVEGVWTLKDEI	LFTVTE	
G _A 91	TTYKLILNLKQAKEEAIKE	LVDAGTAEKYIKL	I	ANAKTVEGVWTLKDEI	LFTVTE	
G _A 95	TTYKLILNLKQAKEEAIKE	LVDAGTAEKYIKL	I	ANAKTVEGVWTLKDEI	KTFVTE	
G _A 98	TTYKLILNLKQAKEEAIKE	LVDAGTAEKYFKL	I	ANAKTVEGVWTLKDEI	KTFVTE	
G _B 98	TTYKLILNLKQAKEEAIKE	LVDAGTAEKYFKL	I	ANAKTVEGVWTYKDEI	KTFVTE	
G _B 95	TTYKLILNLKQAKEEAIKE	A	LVDAGTAEKYFKL	ANAKTVEGVWTYKDEI	KTFVTE	
G _B 91	TTYKLILNLKQAKEEAIKE	A	LVDAGTAEKYFKL	ANAKTVEGVWTYKDEI	KTFVTE	
G _B 88b	TTYKLILNLKQAKEEAIKE	A	LVDAGTAEKYFKL	ANAKTVEGVWTYKDEI	KTFVTE	
G _B 77	TTYKLILNLKQAKEEAIKE	TD	A	LVDAGTAEKYFKL	ANAKTVEGVWTYKDET	KTFVTE

Below the sequence alignment, arrows indicate the positions of secondary structure elements: β_1 , β_2 , α_1 , β_3 , and β_4 .

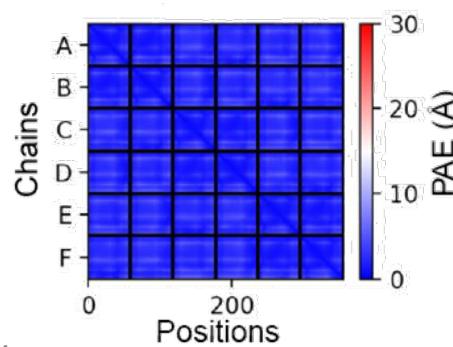
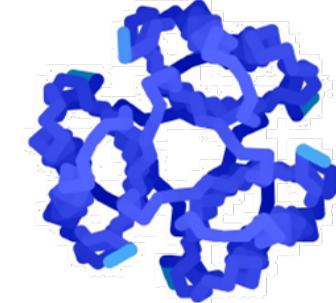
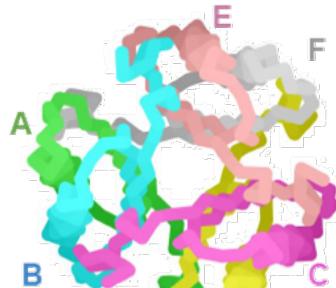


在一个典型的单氨基酸突变导致结构变化的例子中，我们测试结果显示AlphaFold仍然能预测到正确的结构，但也有时预测错误（使用MSA，未使用结构模板）

ColabFold: 预测蛋白质复合体

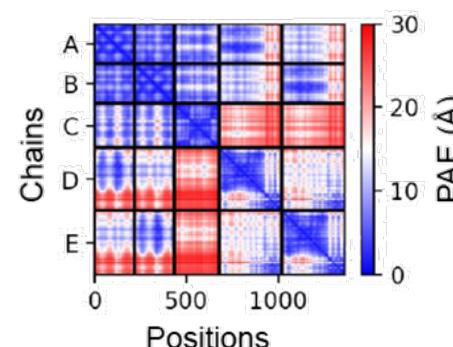
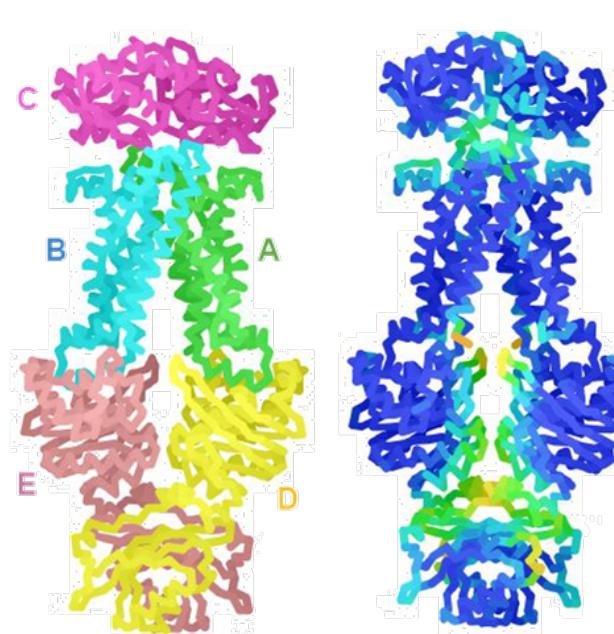
同源多聚体

A - Homo-oligomer (6)



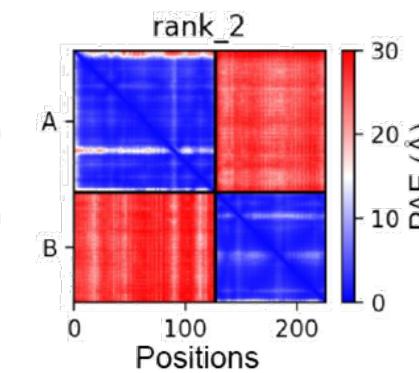
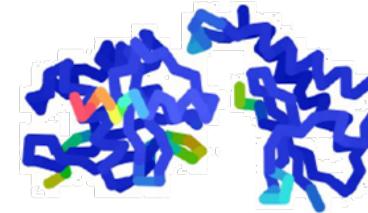
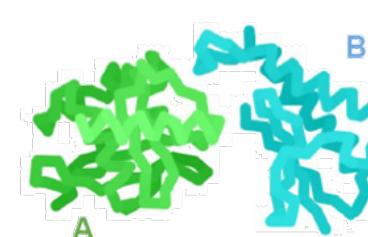
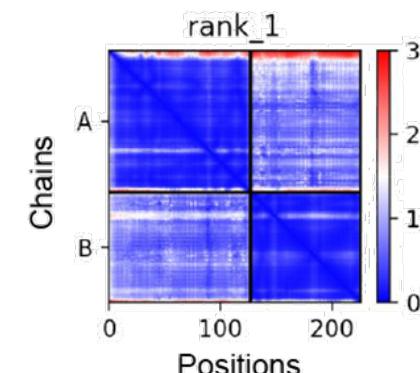
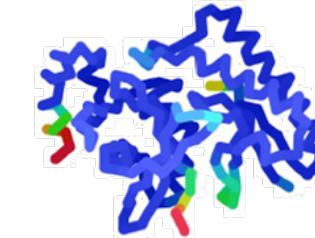
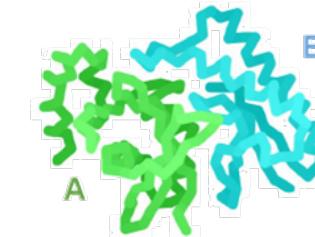
异源多聚体

B - Homo/hetero-oligomer (2:1:2)



异源二聚体

C - Hetero-dimer (1:1)



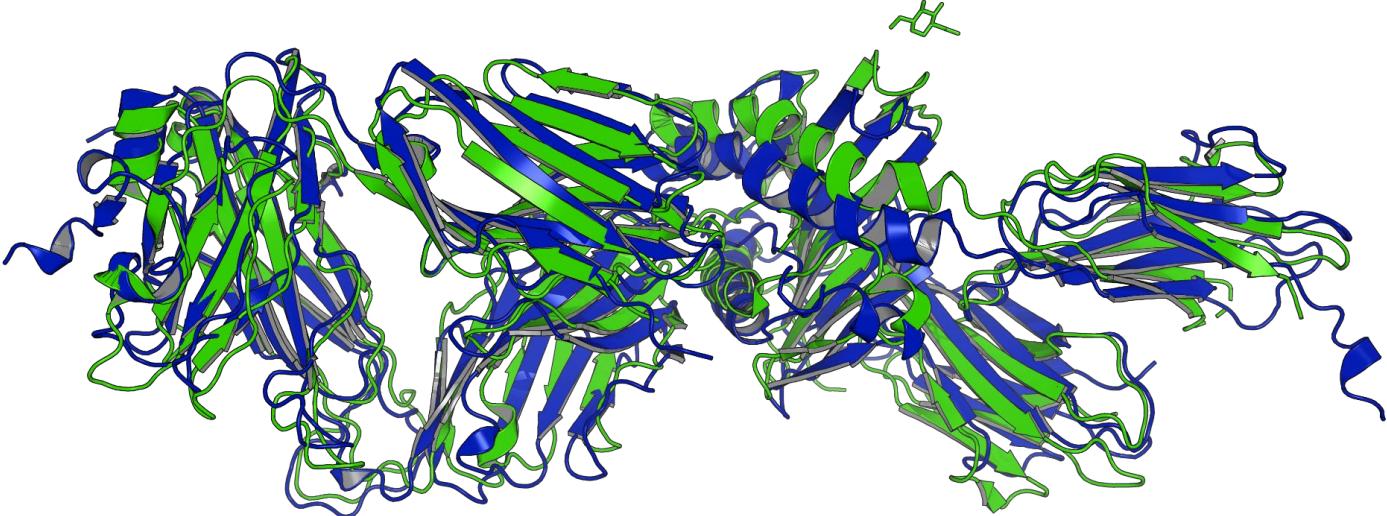
PAE能够显示多聚体的预测准确度

复合物测试体系

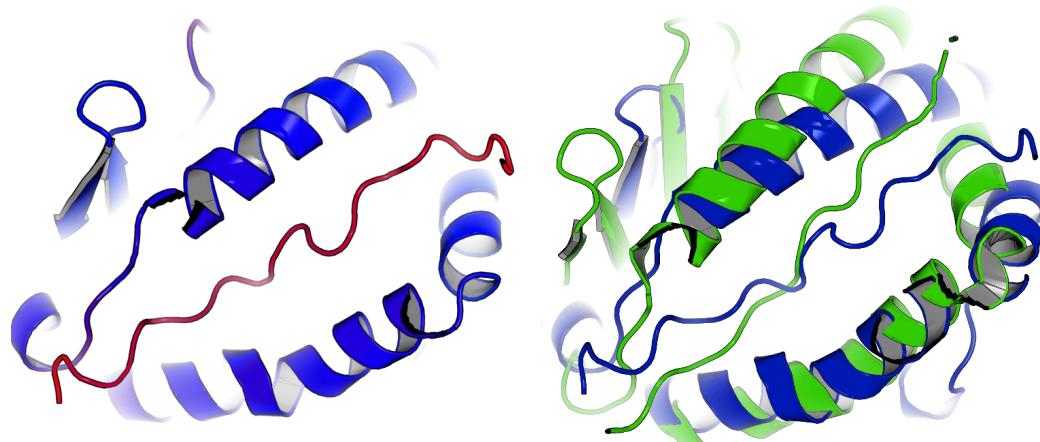


Ras-related protein & Ras-binding Raf

PDB: 1C1Y
RMSD: 0.449 Å
 $\text{ipTM} + \text{pTM} = 0.898$
Running Time: 3.3h



Crystal structure of a autoimmune TCR–MHC complex



Colored by pLDDT

Pred VS Exp

PDB: 4GRL
RMSD: 3.567 Å
 $\text{ipTM} + \text{pTM} = 0.552$
Running Time: 15.4h

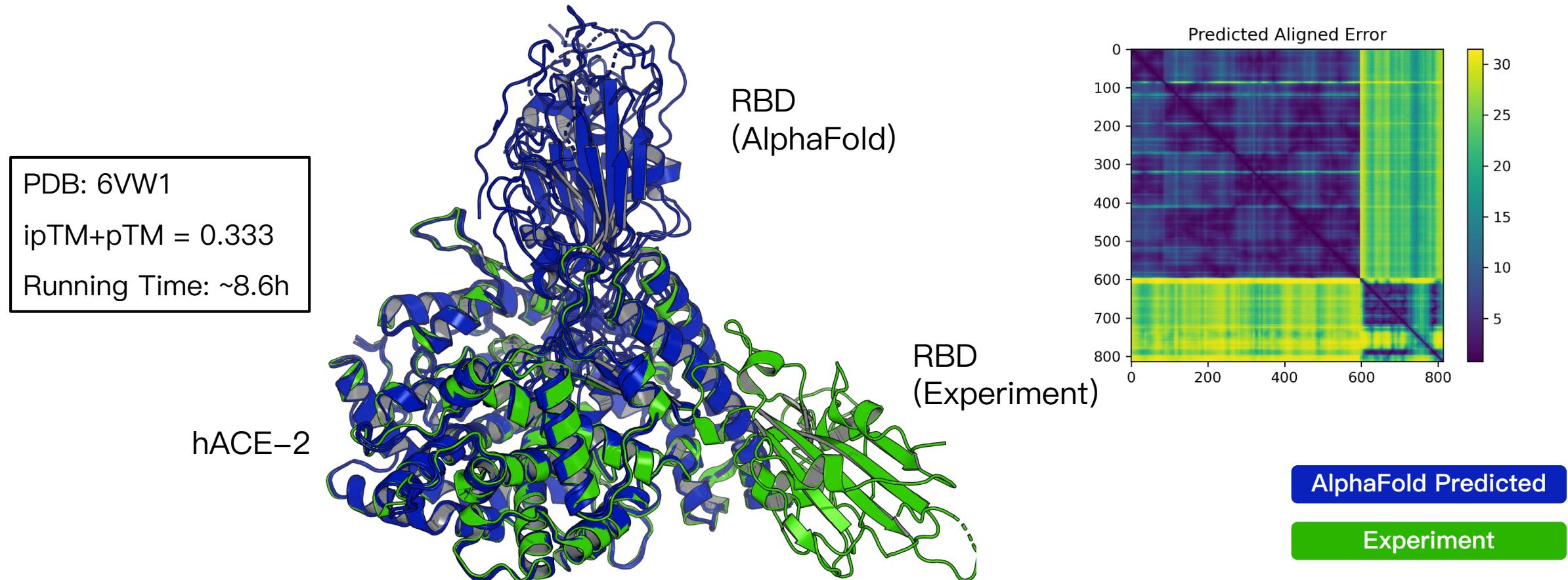
AlphaFold Predicted

Experiment

As a limitation, we observe anecdotally that AlphaFold–Multimer is generally not able to predict binding of antibodies and this remains an area for future work.

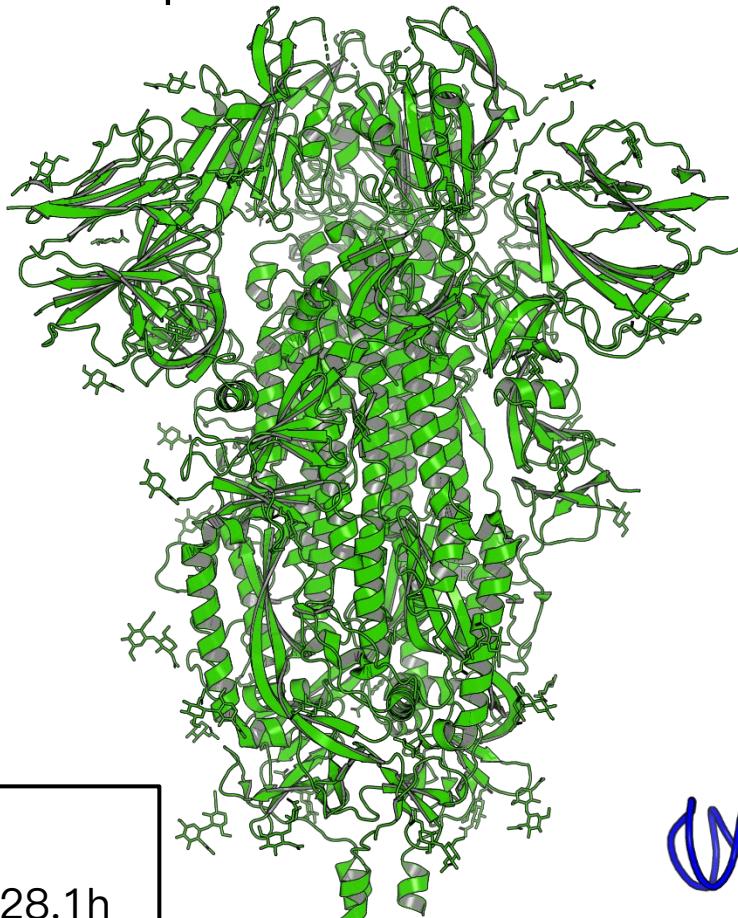
测试体系： RBD | ACE2

- SARS-CoV-2 Spike RBD (receptor-binding domain) & human ACE-2 Receptor



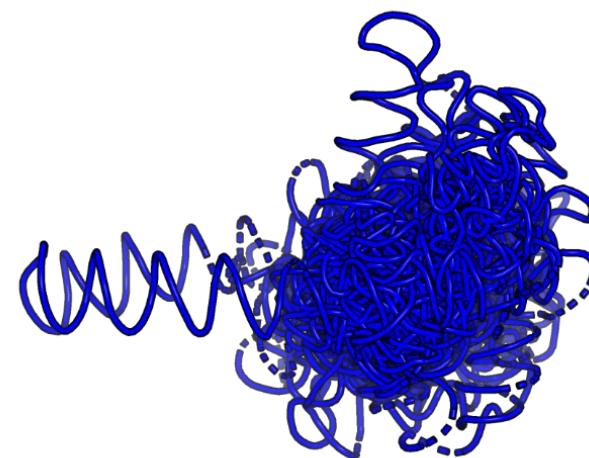
测试体系：SARS-CoV-2 Spike

- SARS-CoV-2 Spike



PDB: 6VXX

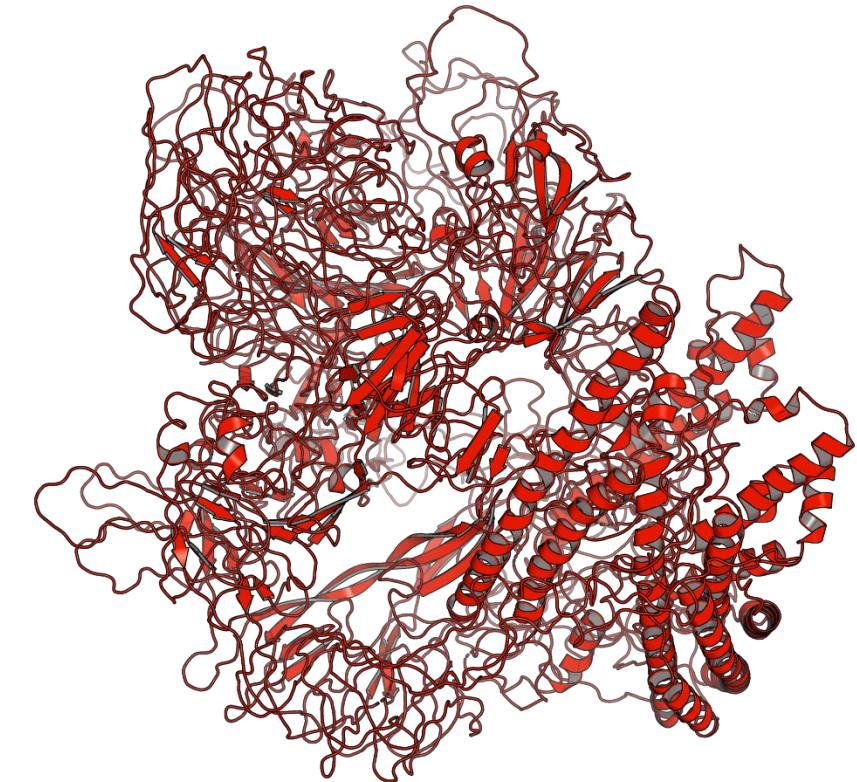
Running Time: 28.1h
(only 1 model)



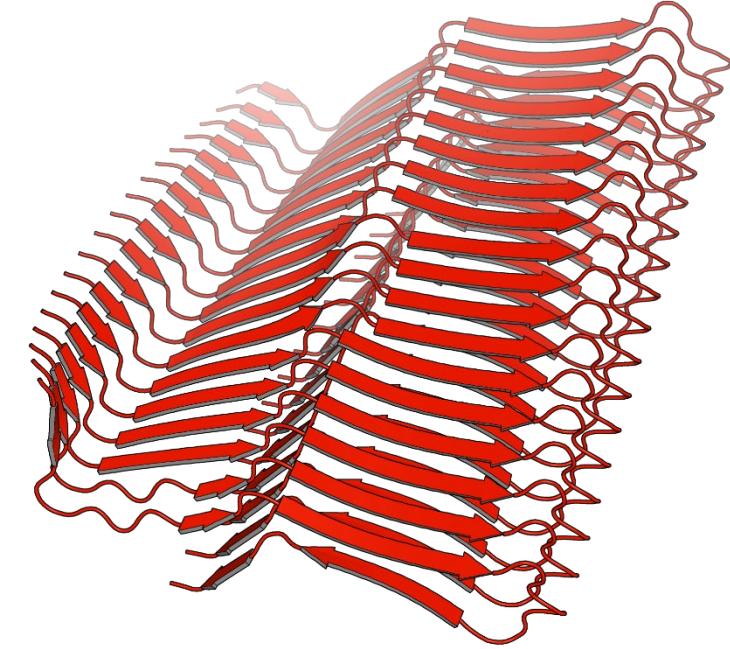
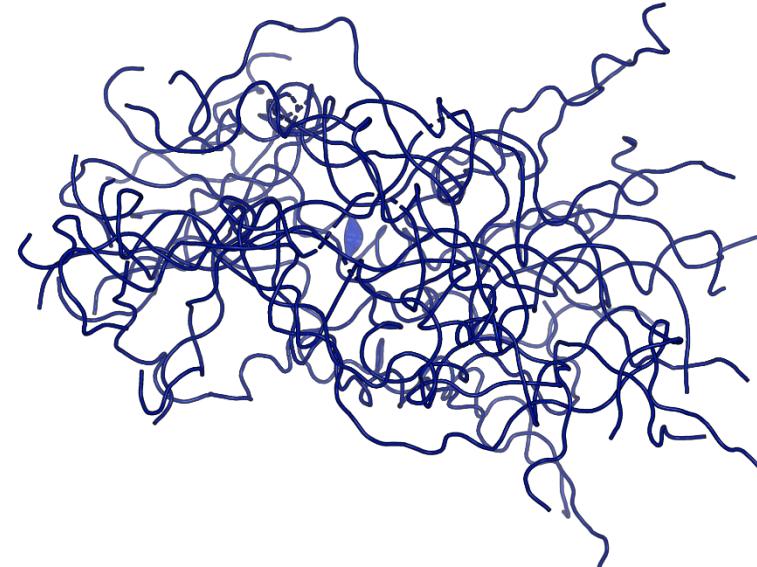
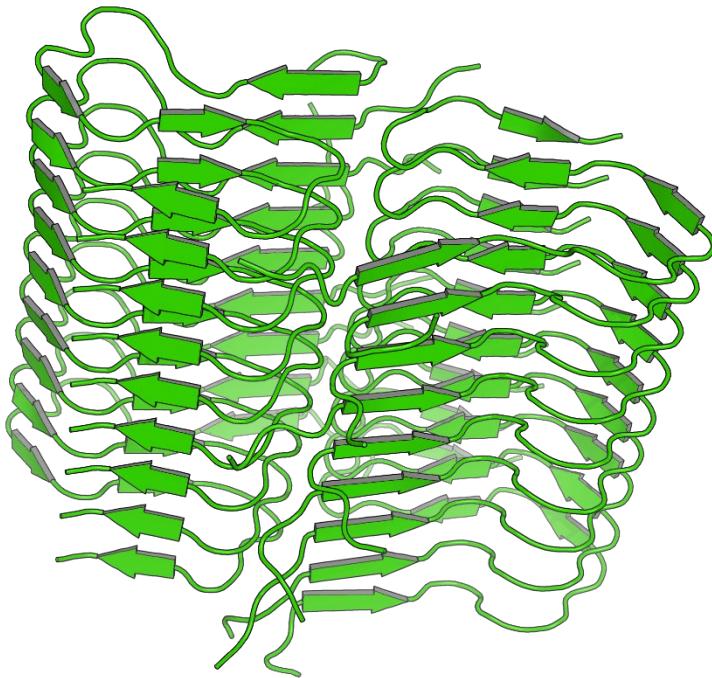
AlphaFold Predicted

ColabFold Predicted

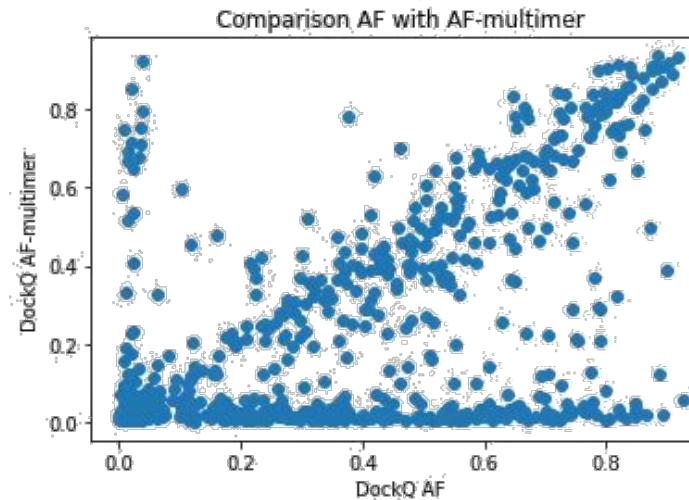
Experiment



测试体系：Amyloid Protein



ipTM+pTM = 0.138
Running Time: 44.0h



AlphaFold–Multimer似乎训练过度了？

- 大量结构预测结果坍缩成一个球
- 整体观感远不如使用原模型，用Gap拼接

现阶段建议Multimer和原版一起使用

AlphaFold的各种版本

现阶段大概存在四种不同的使用 AlphaFold 的方案



AlphaFold 2.0

最初的发行版
不支持复合物



AlphaFold 2.1

最新版
独立的复合物模型



ColabFold

MSA较快
不需要本地资源
使用单体模型预测复合物



ParaFold

高通量预测
运算效率提升

配置环境：AlphaFold Github

<https://github.com/deepmind/alphafold>

硬件软件需求

- Download **genetic databases** – Total: ~ 2.2 TB (download: 428 GB)
- Download **model parameters** (3.5 GB)

论文公布的一些训练和测试的条件

- 训练使用了128张TPUv3（较大的算力），初步训练用了约一周，进一步调试用了4天
- 测试蛋白所需的时间取决于蛋白长度：（测试体系均基于1张V100，2500残基是4张V100）
 - DeepMind: 256个残基需4.8分钟；384个残基需9.2分钟；2500个残基需18小时
 - 我们的测试(不含MSA、JAX编译时间): 56个残基需6秒/模型；841残基需要5分钟/模型
- 大蛋白的预测很容易超出显存，对于16G的V100来说，上限是约1300个残基。2500残基的蛋白用了4张V100（论文数据）

AlphaFold Database

- DeepMind 使用 AlphaFold 预测了很多常见生物的蛋白质组结构，包括人类、小鼠、大肠杆菌等
- 很多常用蛋白的结构都可以直接到AlphaFold DB下载，而且AlphaFold DB也已经接入蛋白质数据库UniProt

Species	Common Name	Predicted Structures
<i>Arabidopsis thaliana</i>	Arabidopsis	27,434
<i>Caenorhabditis elegans</i>	Nematode worm	19,694
<i>Candida albicans</i>	<i>C. albicans</i>	5,974
<i>Danio rerio</i>	Zebrafish	24,664
<i>Dictyostelium discoideum</i>	<i>Dictyostelium</i>	12,622
<i>Drosophila melanogaster</i>	Fruit fly	13,458
<i>Escherichia coli</i>	<i>E. coli</i>	4,363
<i>Glycine max</i>	Soybean	55,799
<i>Homo sapiens</i>	Human	23,391
<i>Leishmania infantum</i>	<i>L. infantum</i>	7,924
<i>Methanocaldococcus jannaschii</i>	<i>M. jannaschii</i>	1,773
<i>Mus musculus</i>	Mouse	21,615
<i>Mycobacterium tuberculosis</i>	<i>M. tuberculosis</i>	3,988
<i>Oryza sativa</i>	Asian rice	43,649
<i>Plasmodium falciparum</i>	<i>P. falciparum</i>	5,187
<i>Rattus norvegicus</i>	Rat	21,272
<i>Saccharomyces cerevisiae</i>	Budding yeast	6,040
<i>Schizosaccharomyces pombe</i>	Fission yeast	5,128
<i>Staphylococcus aureus</i>	<i>S. aureus</i>	2,888
<i>Trypanosoma cruzi</i>	<i>T. cruzi</i>	19,036
<i>Zea mays</i>	Maize	39,299

AlphaFold DB 中包含的蛋白质组

Docker 版本和 Conda 版本

Docker 版本 (AlphaFold 2.1)

- 高度封装的版本，加载即用，免除安装困难
- 可调参数比较少
 - 输入序列
 - 预设方案：monomer/ptm/multimer
 - 模板时间：设置结构模板搜索范围

- reduced_dbs: 使用了较少的BFD，没有ensembling，只需要8核CPU, 8G内存, 600G存储
- full_dbs: 比casp14快8倍，取消了ensembling，GDT只降低了0.1
- casp14: CASP14比赛中用的设定，8轮ensembling
- Ensembling: the trunk of the network is run multiple times with different random choices for the MSA cluster centres

Conda 版本 (ParaFold的方案)

- 相当于本地安装的版本，包含完整的AlphaFold代码和运行环境
- 需要安装AlphaFold/ParaFold环境，请参考 <https://github.com/Zuricho/ParallelFold>
- 支持自定义，如选取单体模型，pTM 模型，复合物模型的全部或部分、修改 Recycling 次数、选择是否 Amber 优化、设定 data 数据集位置等，均可通过修改参数实现

ColabFold 版本

ColabFold 本身的优点

- 使用极度方便，打开网页即可使用，不需要配置资源
- 支持很多功能：调整recycling；选择是否需要模板、AMBER优化；选择更快的序列比对算法MMseq2；建模蛋白质异源复合物(hetero oligomer)；建模蛋白质多聚体(homo oligomer)
- Jupyter Notebook形式，运行方便
- 优秀的可视化方法

ColabFold 本身的缺点

- Colab现在已经限制GPU资源，没有GPU不能运行AlphaFold
- ColabFold本身不支持高通量结构预测：MMseq2在远程服务器运行，Colab仍然使用串行流程

如果想使用ColabFold，那可以试试本地版本的ColabFold，能够使用更多的资源

<https://github.com/YoshitakaMo/localcolabfold>

各种运行时的bug：常见错误和解决方案汇总

MSA报错

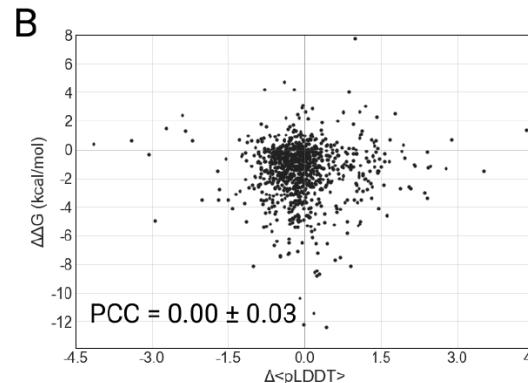
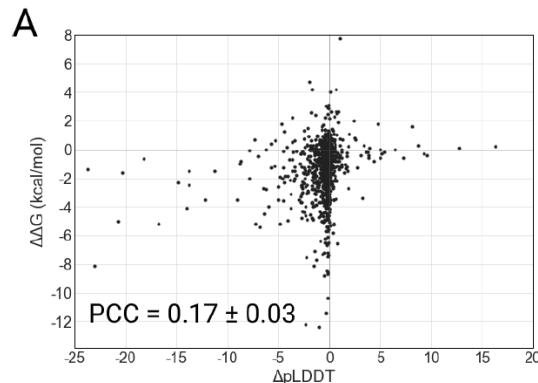
- 内存不足：一般会出现segmentation fault的字样，增加内存（在集群中就是增加CPU核数），换大内存结点等都可解决这个问题
- HHblits failed：这个错误更常见，而且长期被误解，网上也没有提供有效的方案。这里提供一个正确的解决方案：把用到的UniClust数据库从2018-08更新到2020-06版本（只要更新就行）

GPU报错

- Index Error：你是不是把复合物任务用单体模型算了？
- CUDA memory error：显存不足，需要多卡共享显存
- Tensor larger than 2GB：提取的feature过大，需要限制MSA的深度（AF2.1已经优化，很难出现这个问题）
- 运行极慢 (Slow Compile)：大概率没找到GPU，Python和CUDA环境配置需要重新检查

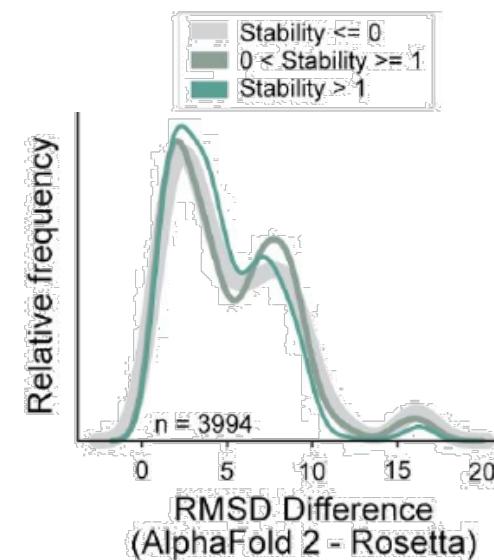
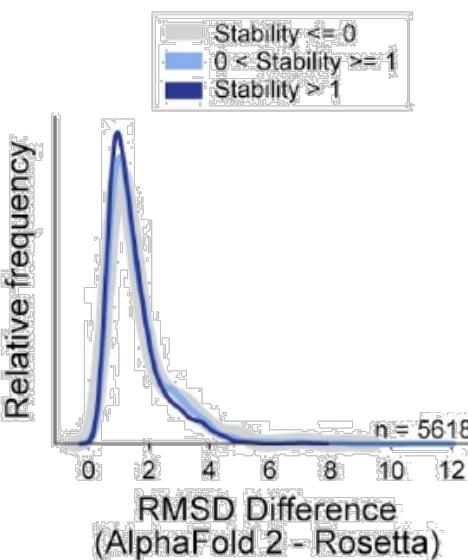
还有新的问题可以去AlphaFold或者ParaFold的GitHub提Issue

能预测蛋白质的稳定性吗？不能

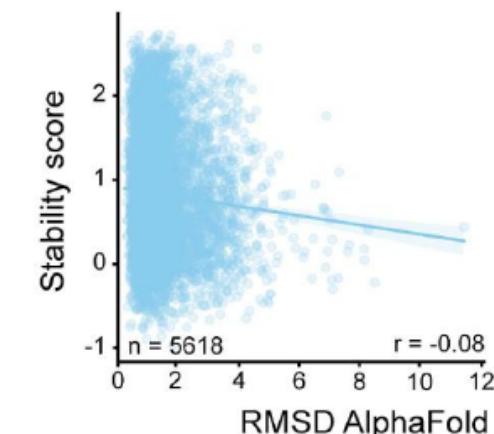


突变导致的稳定性变化：pLDDT与 $\Delta\Delta G$ 相关性差

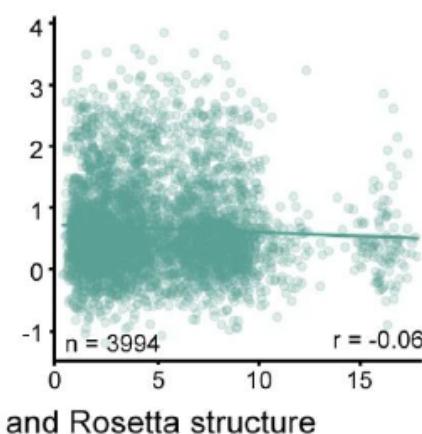
实验数据来源：ThermomutDB



Restricted designs
(Round 5)

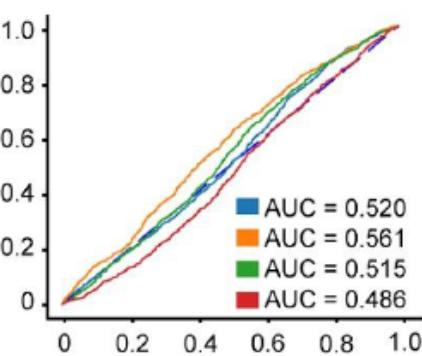
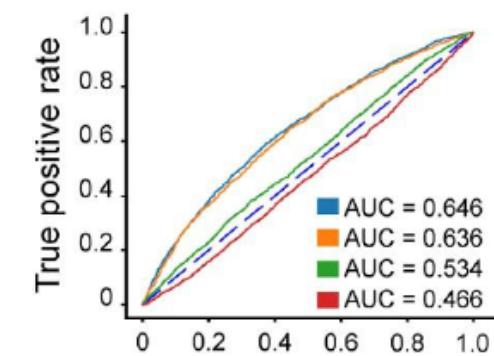


Diversity-oriented designs
(Round 6)



突变导致的稳定性变化：RMSD与稳定性相关性差

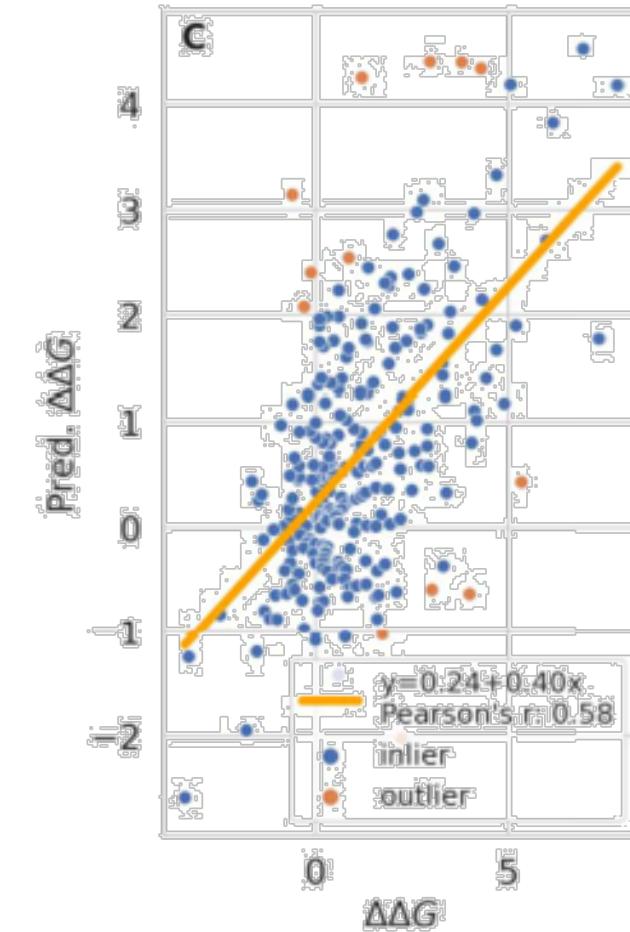
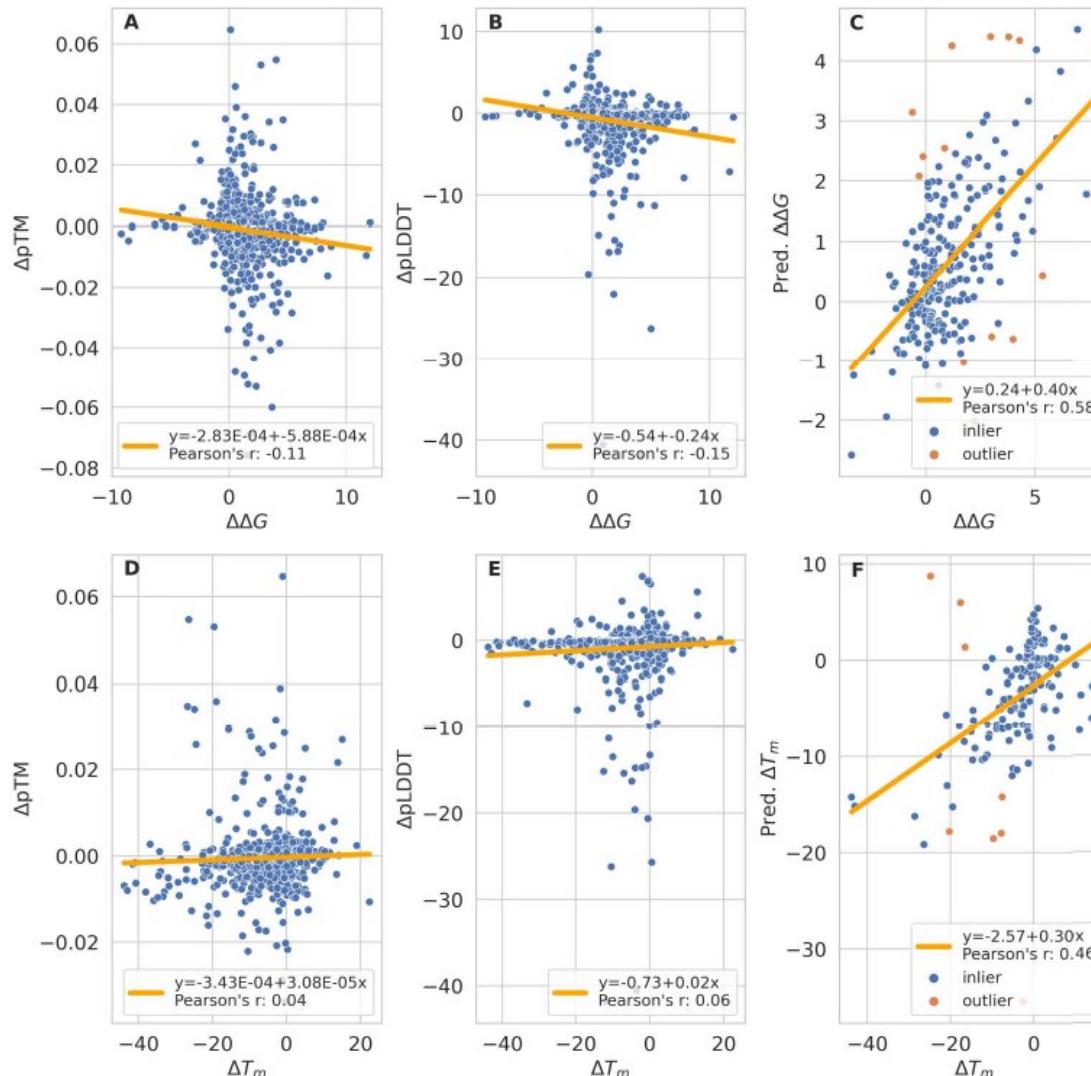
实验数据来源：高通量实验



Rosetta score
Rosetta score (AF2 model)
RMSD AF2 and Rosetta
pLDDT

RMSD和pLDDT的效果均不如
Rosetta score

能预测蛋白质的稳定性吗？似乎又行了

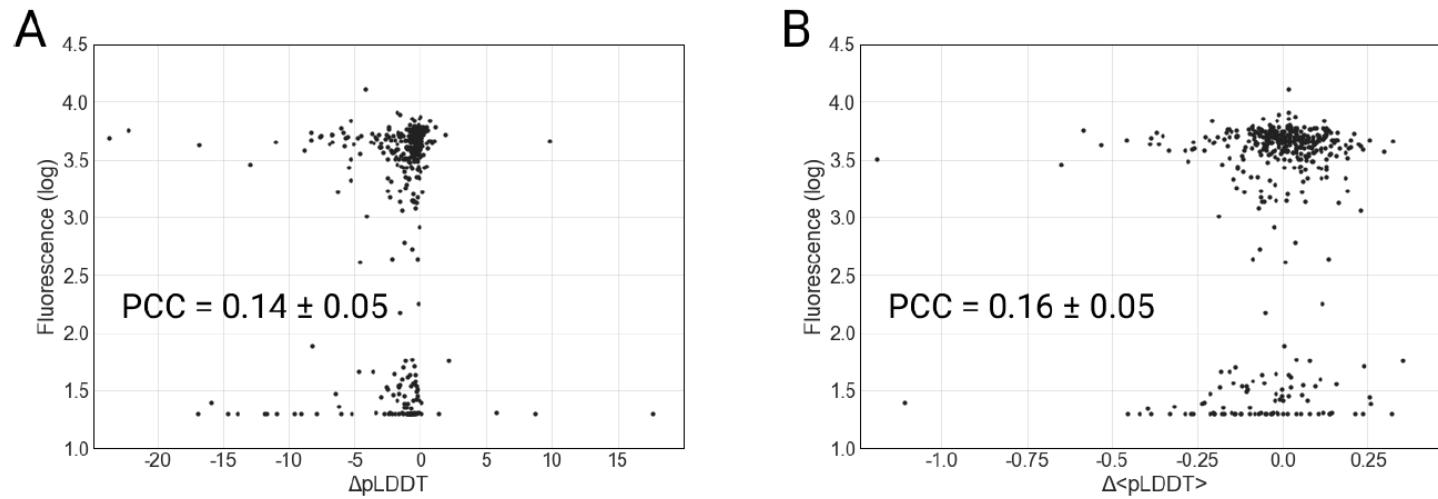


突变导致的稳定性变化：

- 搭配上神经网络能够预测 $\Delta\Delta G$ (simple MLP)

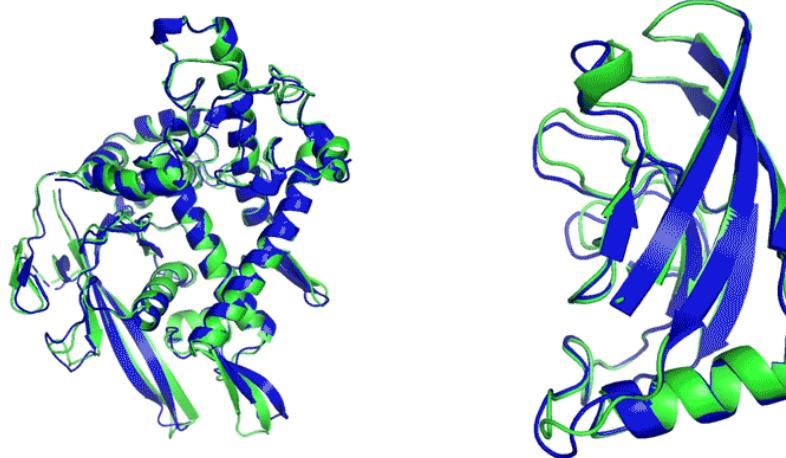
实验数据来源：FireProtDB

能预测突变导致的蛋白质的功能变化吗？这次真的不能



突变导致的GFP荧光强度变化：pLDdT与Fluorescence相关性差

实验数据来源：Sarkisyan et al. Nature (2016)



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

- Experimental result
- Computational prediction

谢谢聆听

Bozitao Zhong
2022/01/16