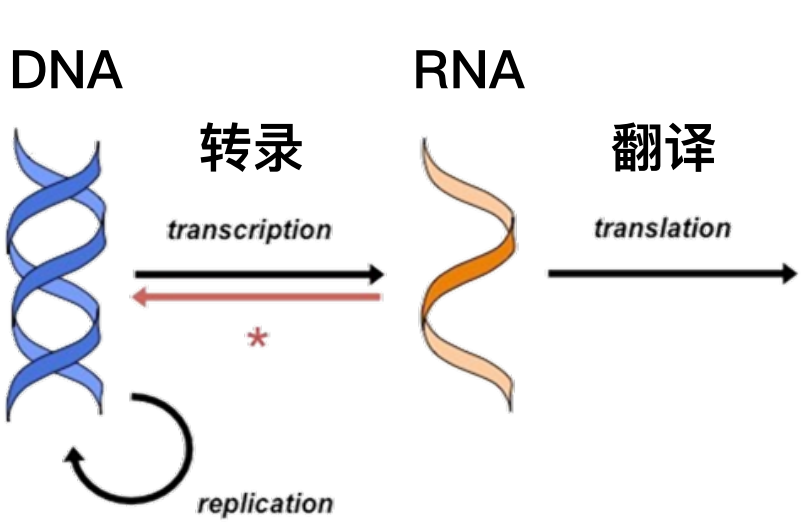
Alphafold2: 如何应用AI预测蛋白质三维结构

钟博子韬 上海交通大学

2021/07/22

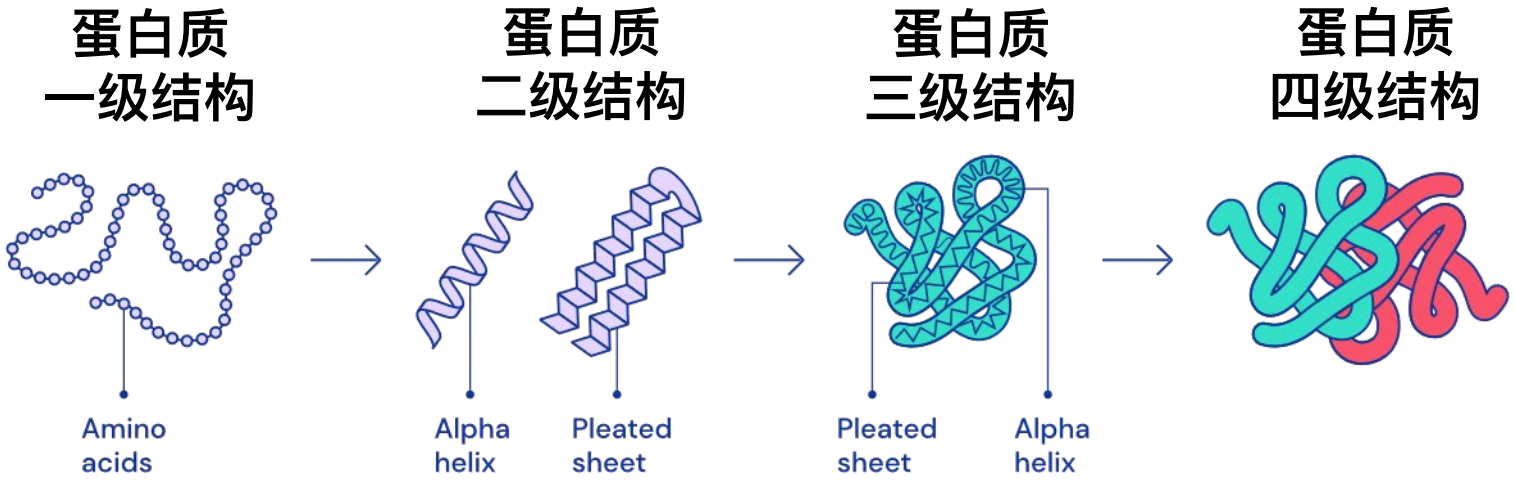
目录 Content

蛋白质的折叠过程



中心法则

DNA → RNA → 蛋白质



蛋白质折叠过程

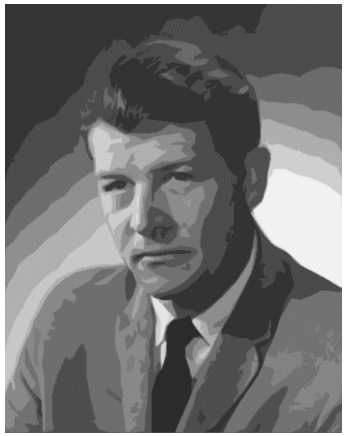
（蛋白中的氨基酸=残基）

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

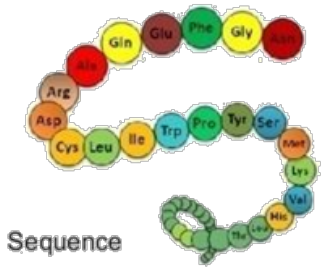
1972 Nobel Prize in Chemistry



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



生物学功能

药物靶点

抗体结合

工业化酶

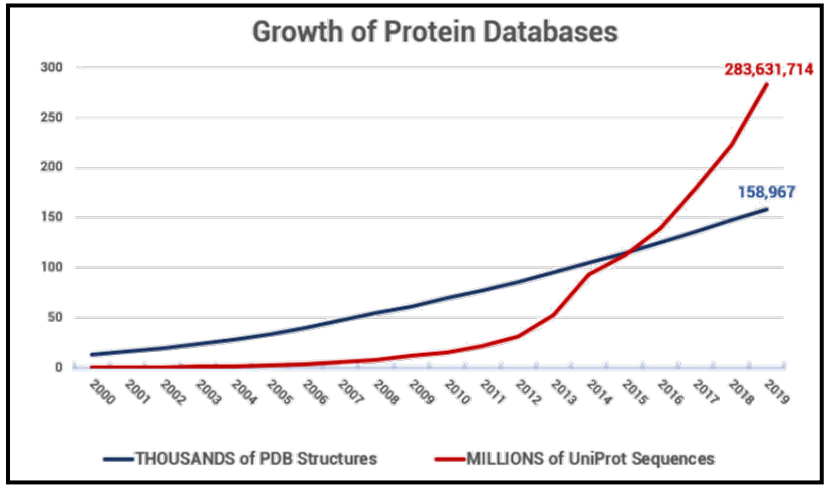
蛋白设计

实验方法

低通量、高准确度

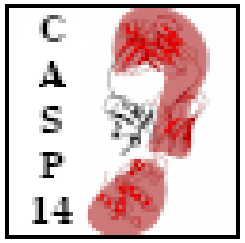
计算方法

高通量、低准确度



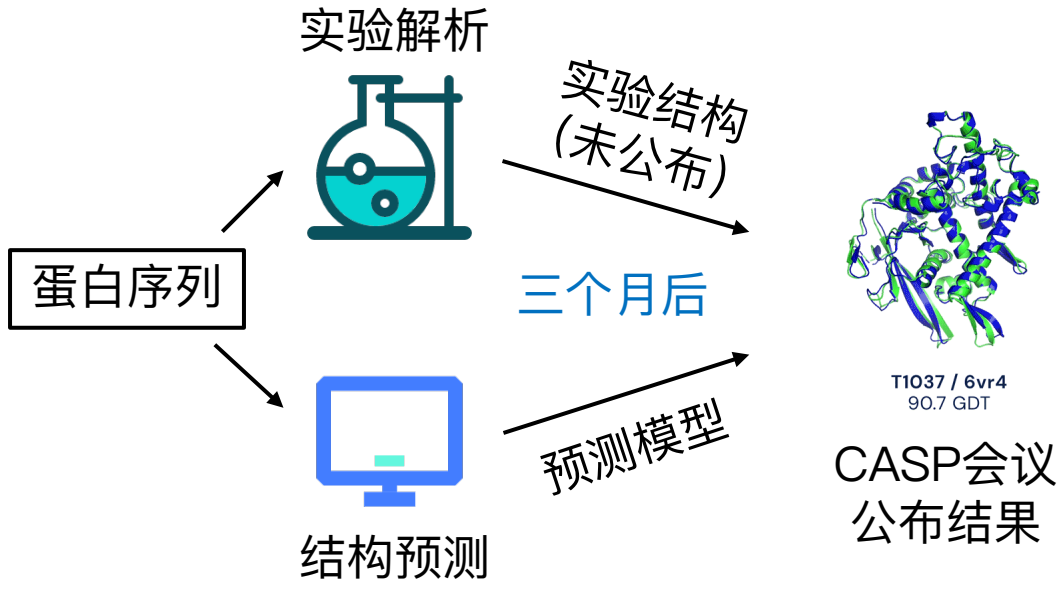
海量的序列信息，少量的结构信息

需要高精度高通量发现蛋白质结构的方法



CASP

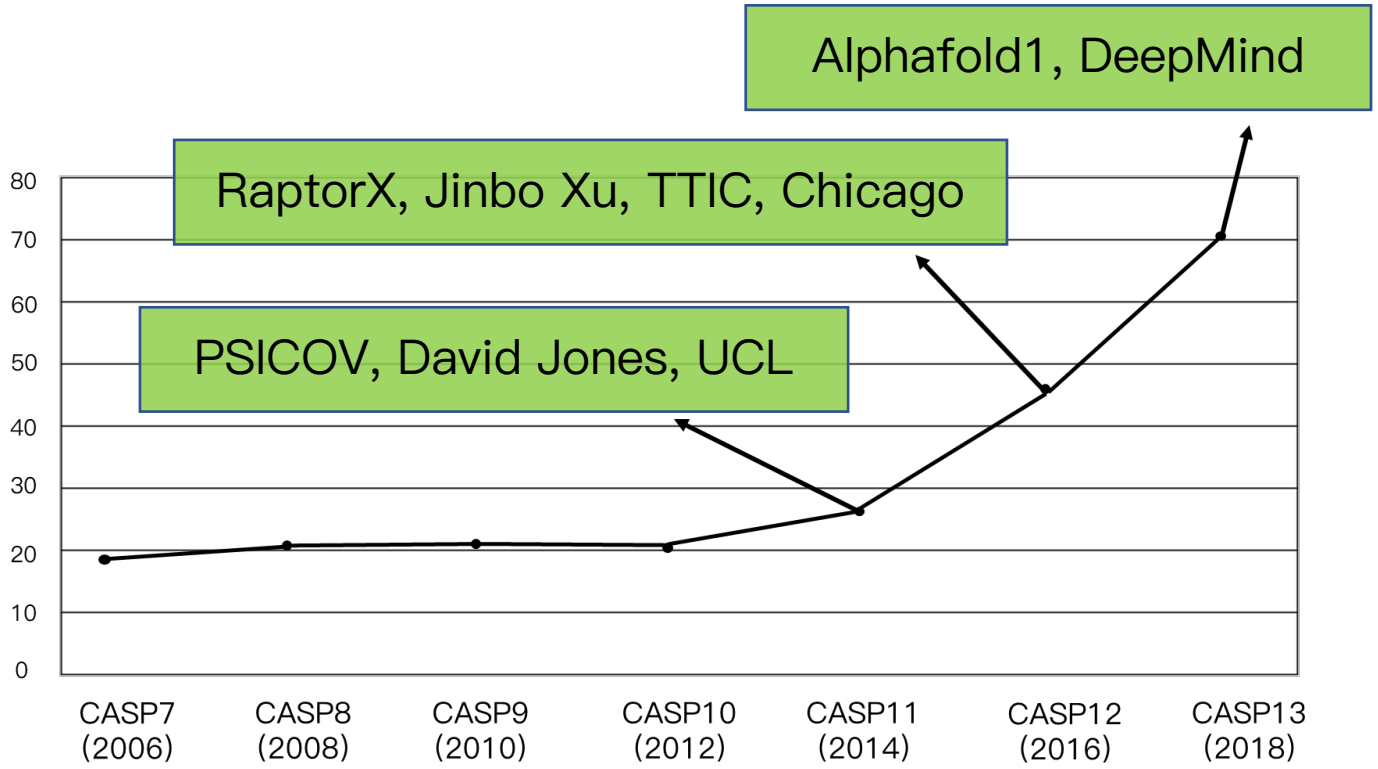
Critical Assessment of protein Structure Prediction, 1994

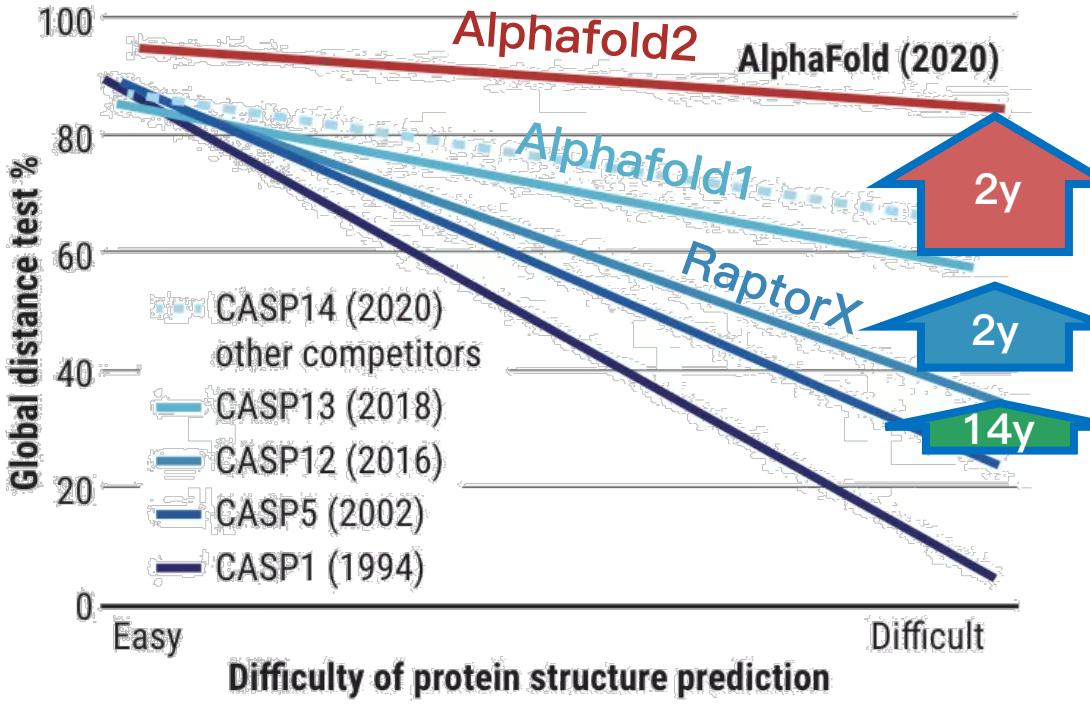


评估计算机预测蛋白质结构方法的准确度

https://www.dnastar.com/blog/structural-biology/why-structure-prediction-matters/

CASP中的关键提升



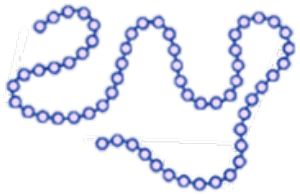


- 2014 (MSA): PSICOV, David Jones, UCL
- 2016 (Deep Learning/ResNet): RaptorX–Contact, Jinbo Xu, TTIC, Chicago
- 2020 (Transformer): Alphafold2, DeepMind

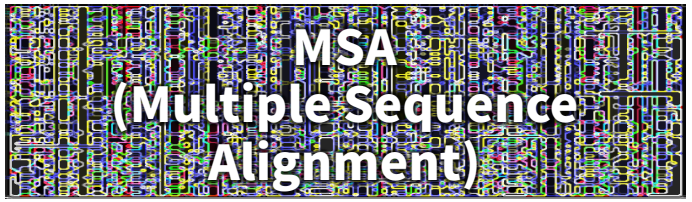
Picture from Jianyi Yang's Report and Science

预测蛋白质的Contact Map: 间接预测蛋白质结构

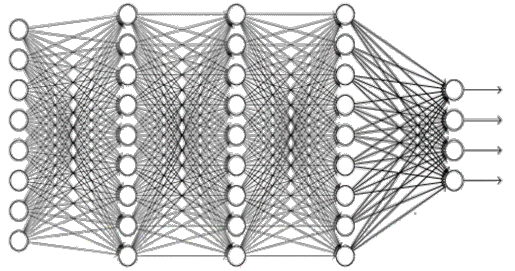
蛋白质序列



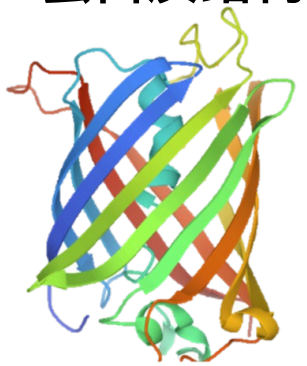
多序列比对



深度学习模型



蛋白质结构

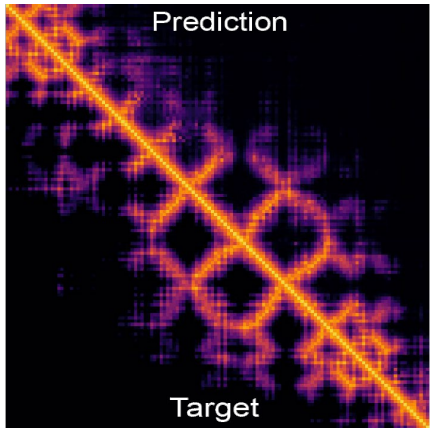


在序列数据库中找与原序列接近的序列
按对应的氨基酸位点做比对(Alignment)

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

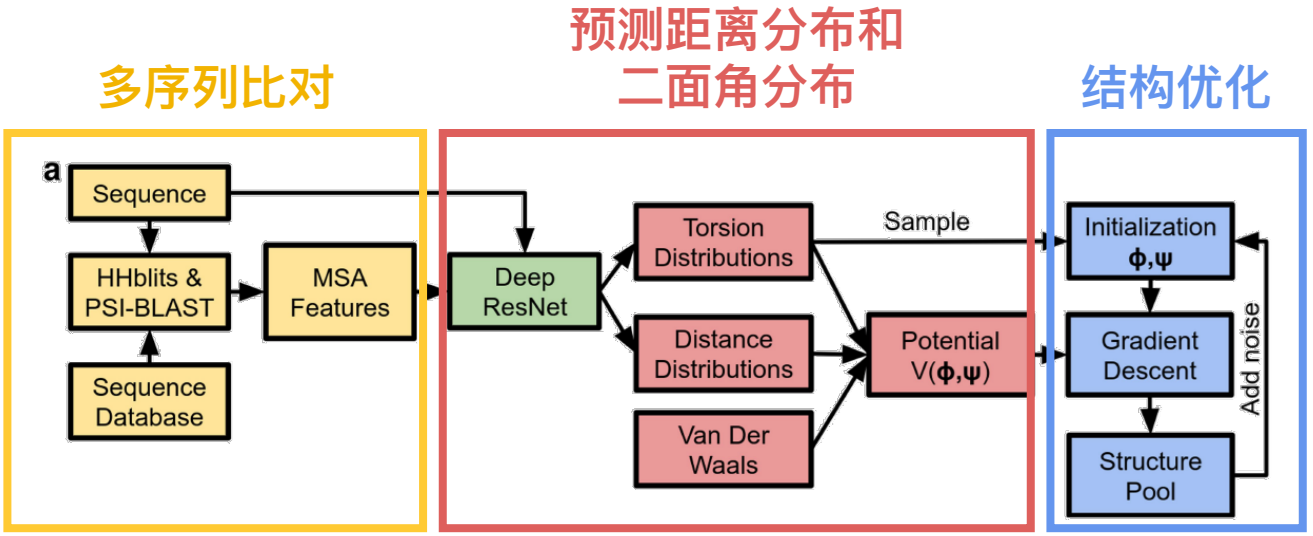
Alphafold 2018

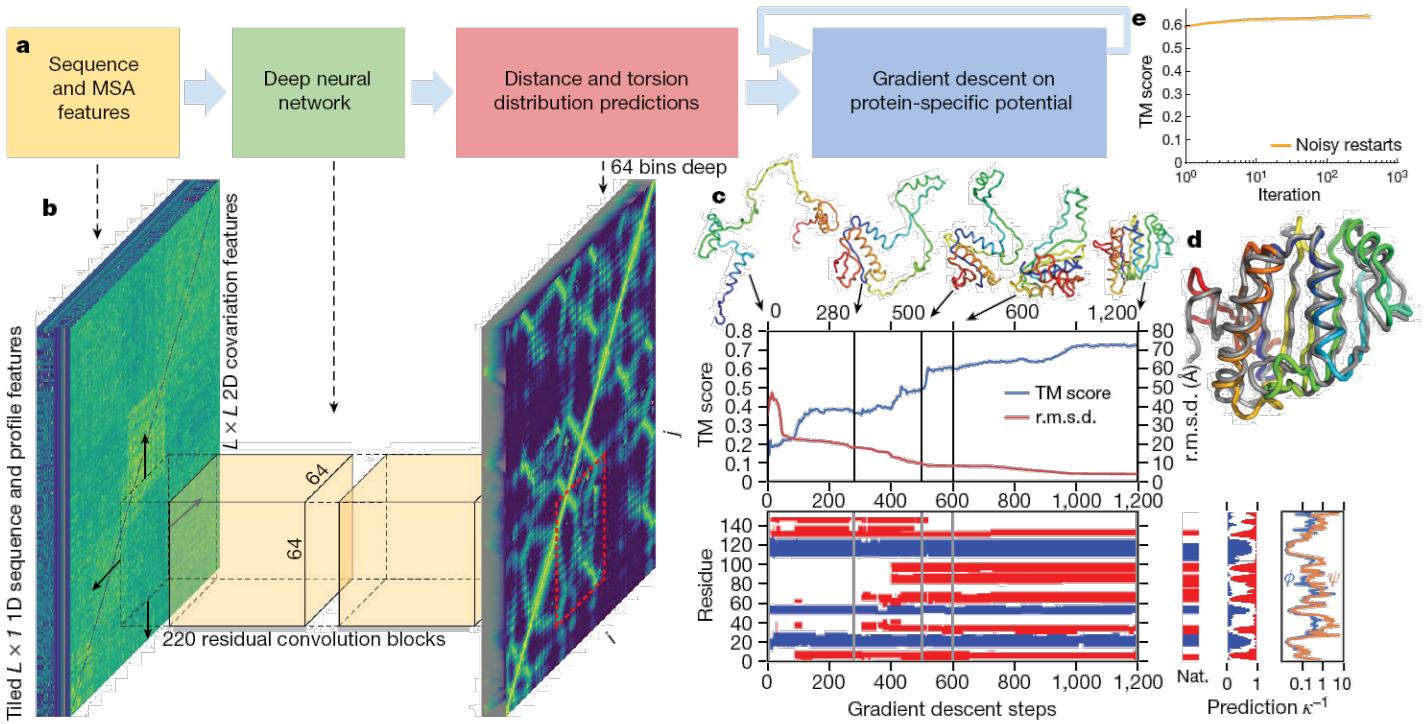
主要特点：

- 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



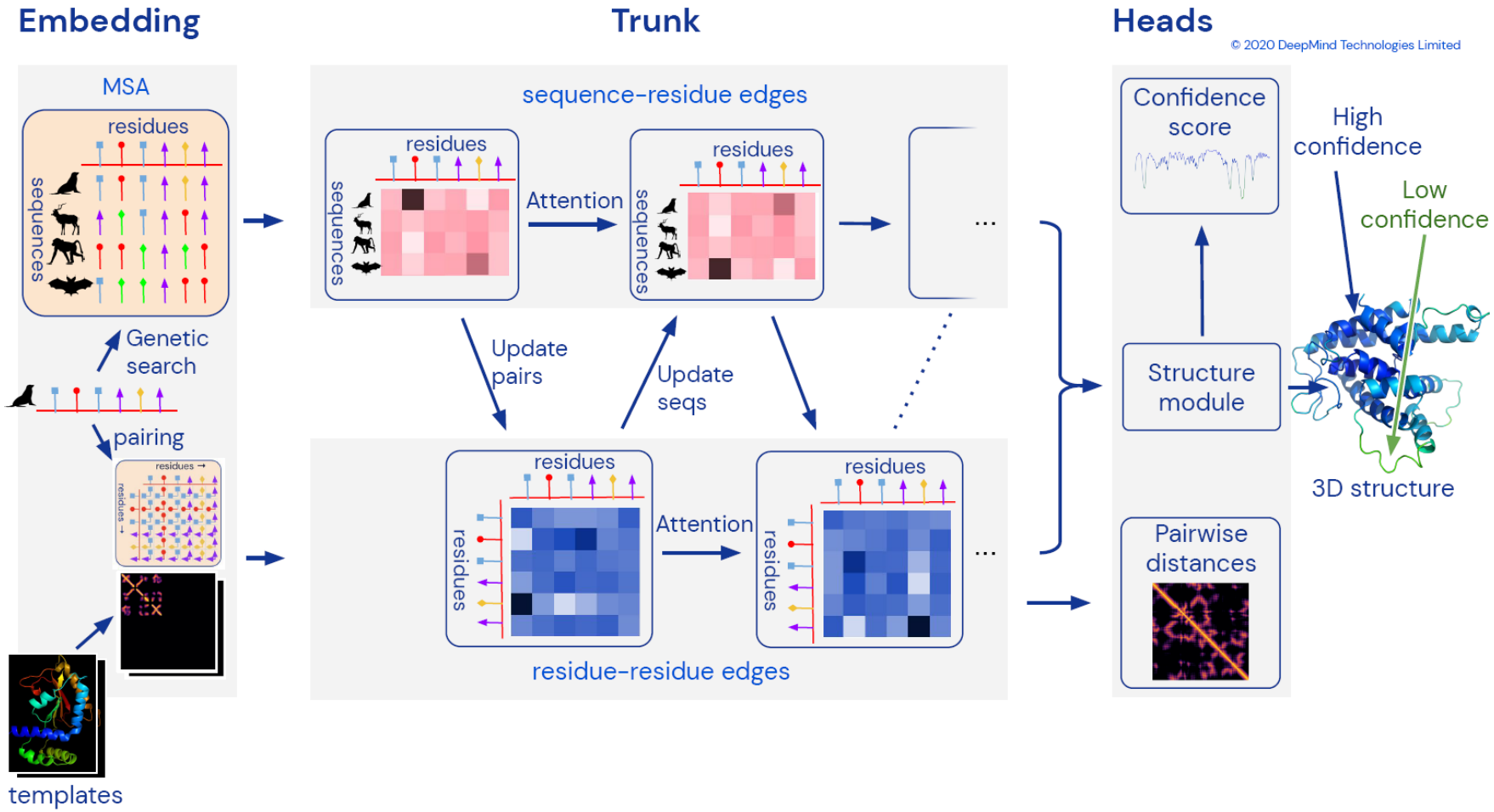


Senior A W, et al. Nature, 2020, 577(7792): 706–710

Alphafold2 的整体架构：在2020公布的信息

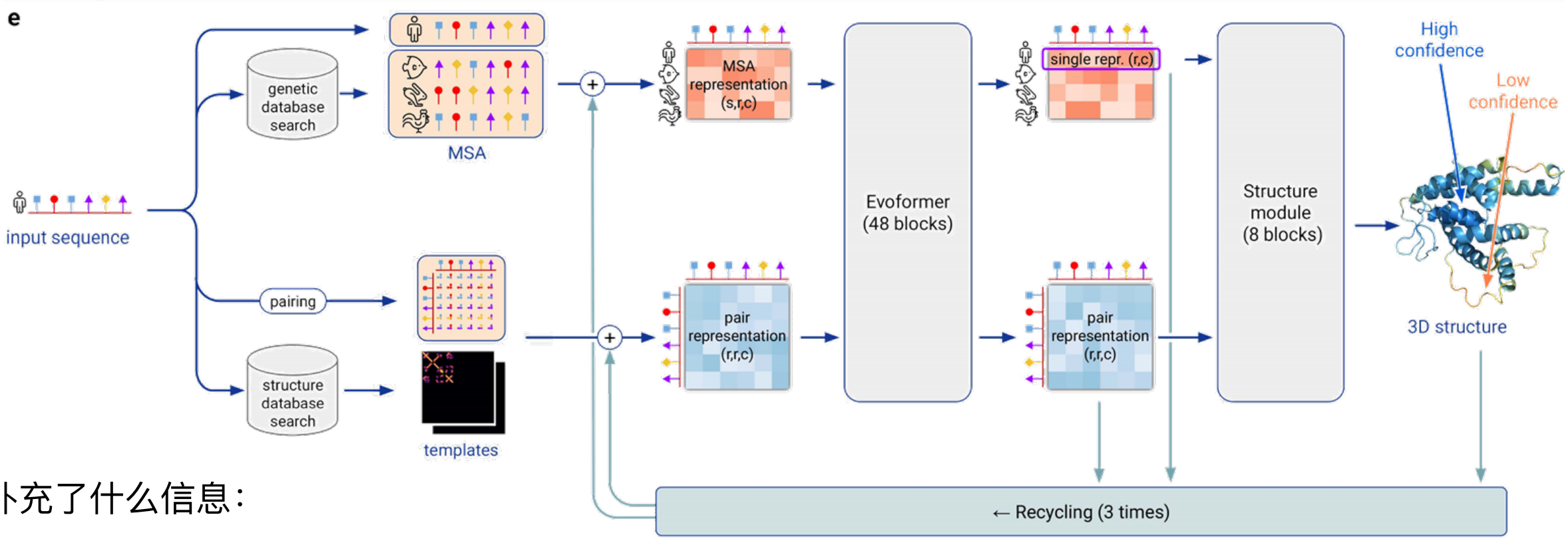
主要特点：

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



Picture taken from DeepMind slides in CASP14

Alphafold2 的整体架构：2021年发表的Nature论文



补充了什么信息：

- 架构上补充了Recycling
- Attention模块的细节：Evoformer
- 公布了Structure module的细节：IPA, residue gas

Jumper, J, et al. Nature, 2021

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
- 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

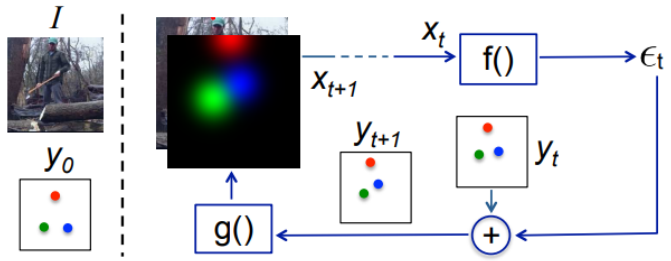
丰富的MSA数据库加上优秀的MSA搜索比对方法，得到高质量的MSA结果用于训练

Jumper, J, et al. Nature, 2021

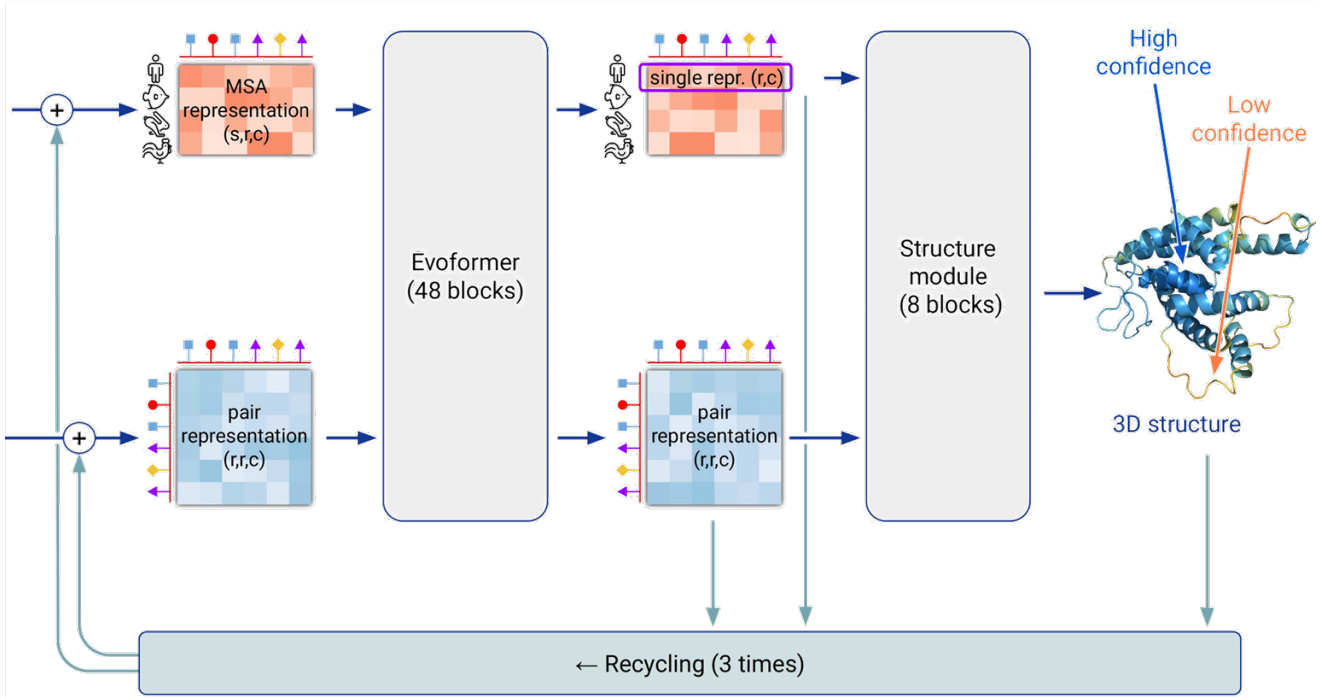
整体架构的精彩之二：

使用Recycling进行多轮迭代训练和测试

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



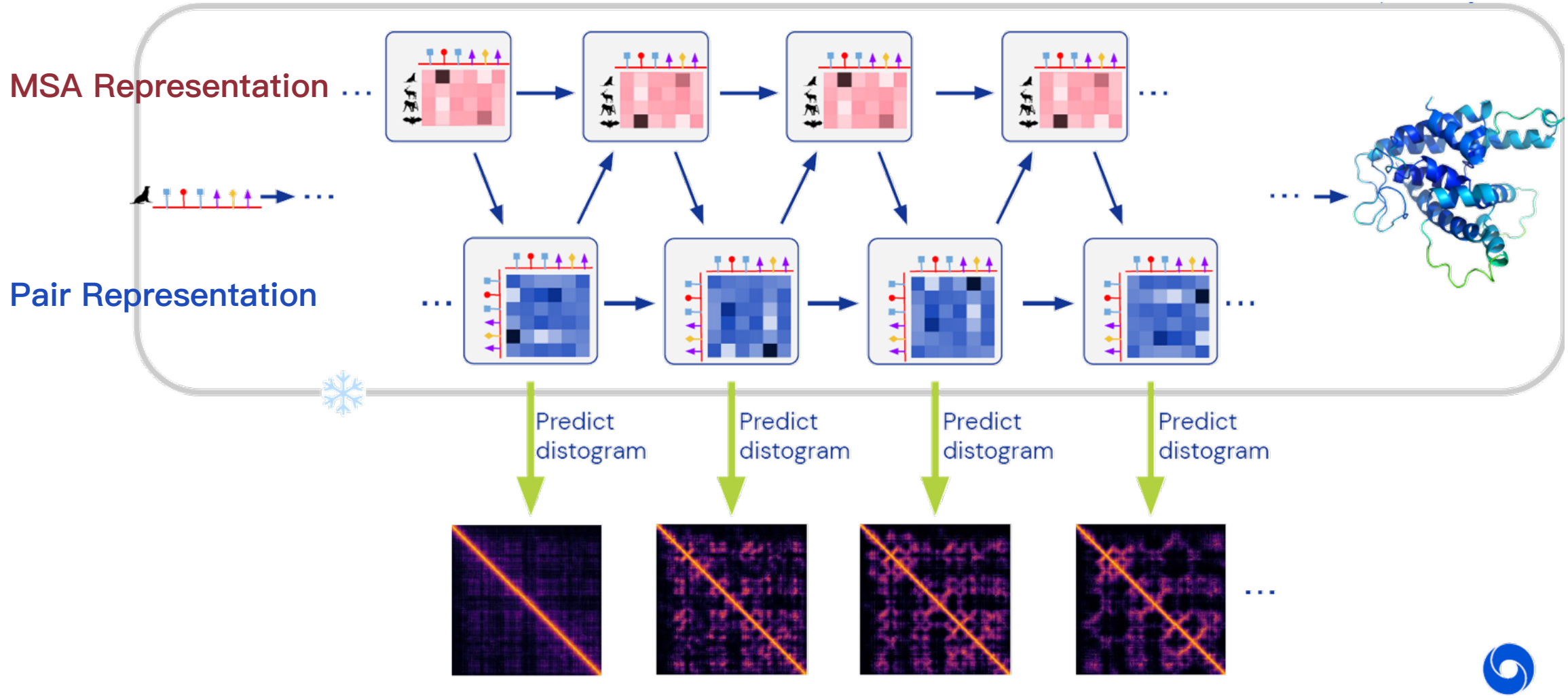
Recycling最初用于计算机视觉中的姿态估计(post estimation)问题，将训练的结果返回输入继续迭代训练



Carreira J, et al. CVPR 2016: 4733–4742

整体架构的精彩之三：

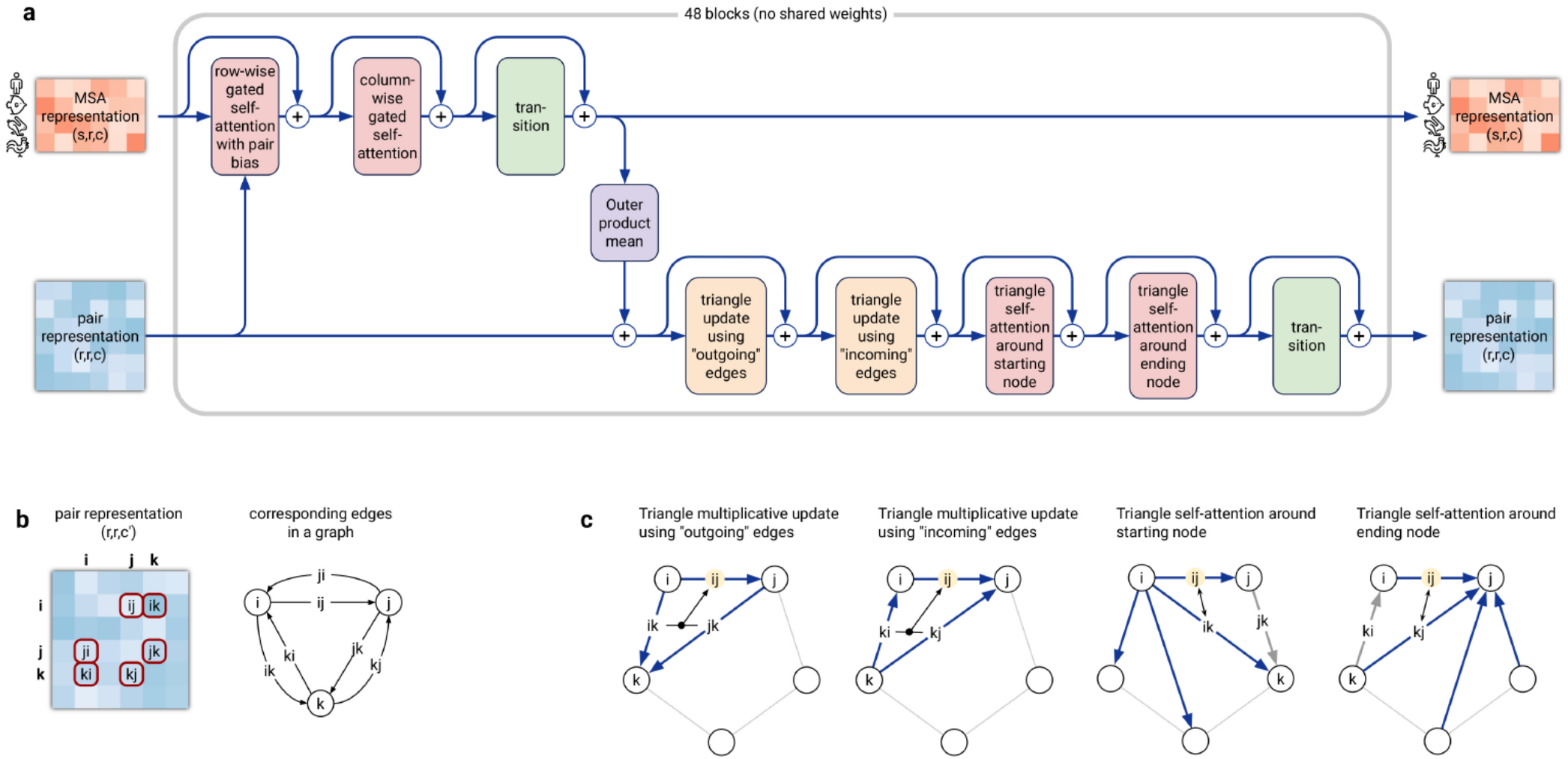
Evoformer: 用于结构预测的Attention架构



Picture taken from DeepMind slides in CASP14

整体架构的精彩之三：

Evoformer: 实现细节

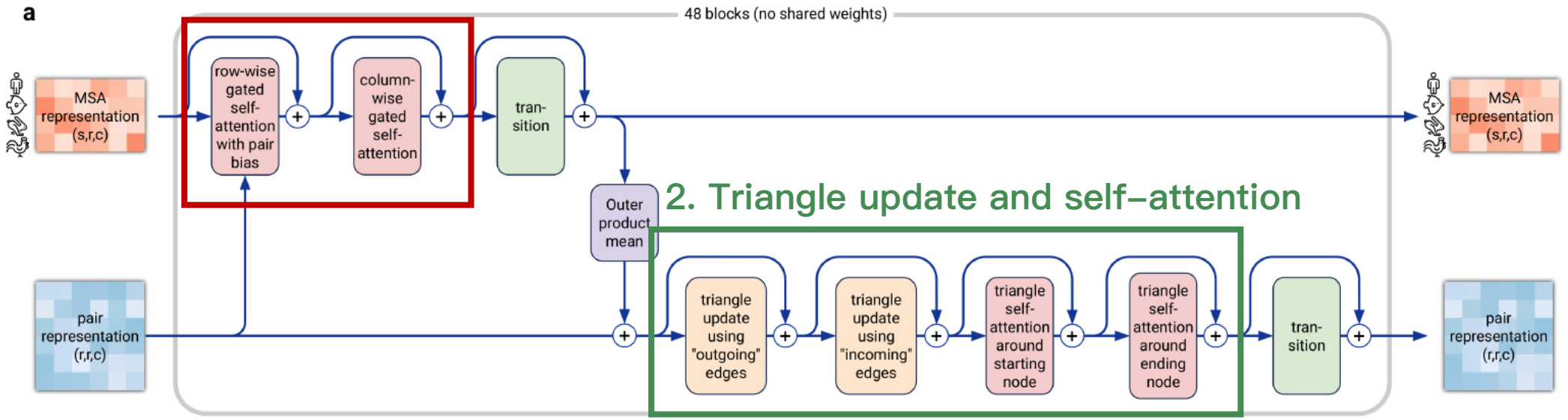


Jumper, J, et al. Nature, 2021

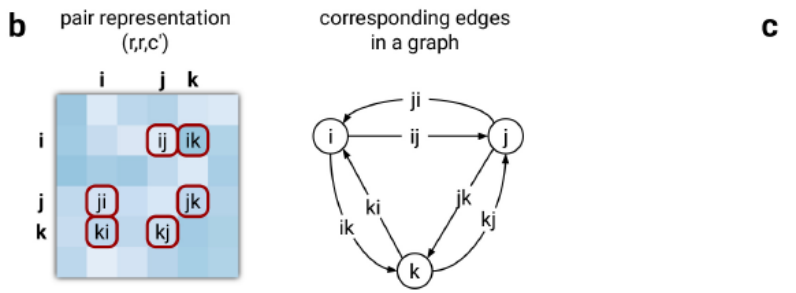
整体架构的精彩之三：

Evoformer: 实现细节

1. Row/column-wise self-attention

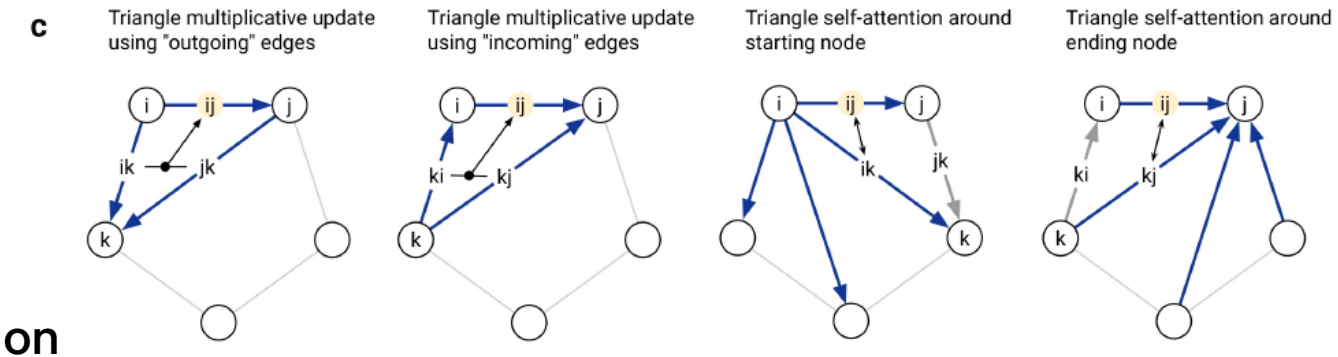


Evoformer的整体架构



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

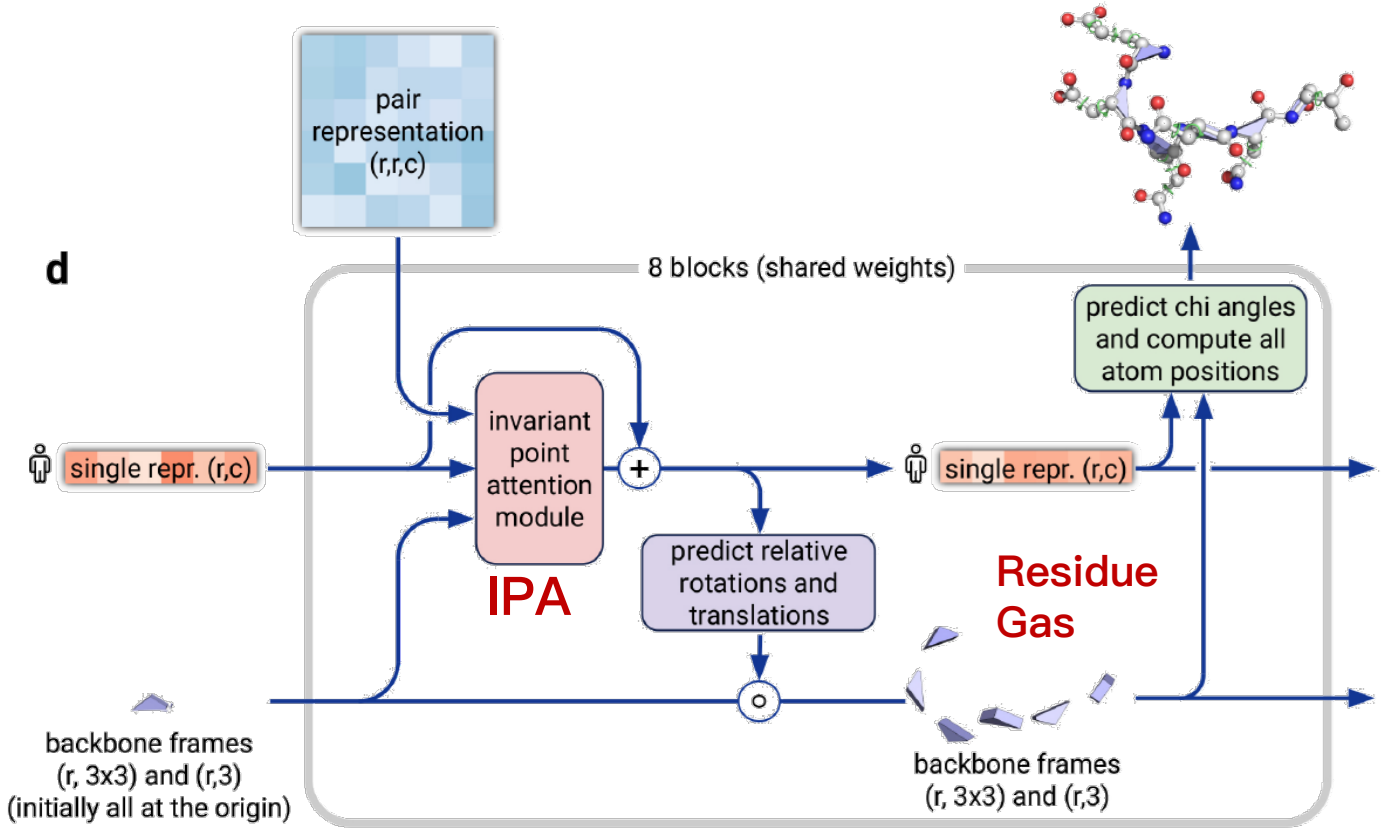
用ik推ij的信息，是否接受更新取决于jk边

Jumper, J, et al. Nature, 2021

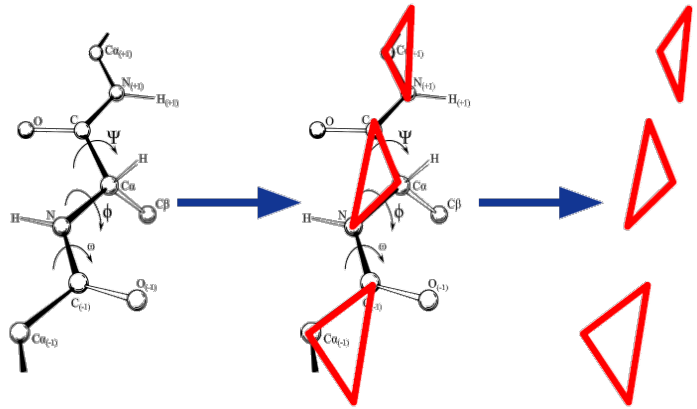
整体架构的精彩之四：

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

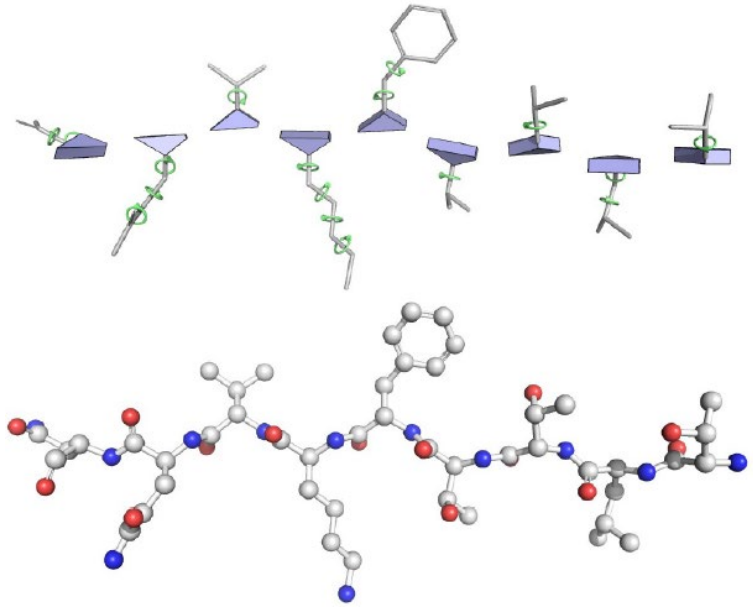


- IPA用于实现3D Equivariant（平移旋转等变性）
- Residue Gas用于表示蛋白质结构
- 输入：
  - 序列信息（目标蛋白）
  - Distance Map信息
  - 蛋白质骨架初始Residue Gas
- 输出：
  - 全原子的位置坐标
  - IDDT-Cα（评估建模精度）

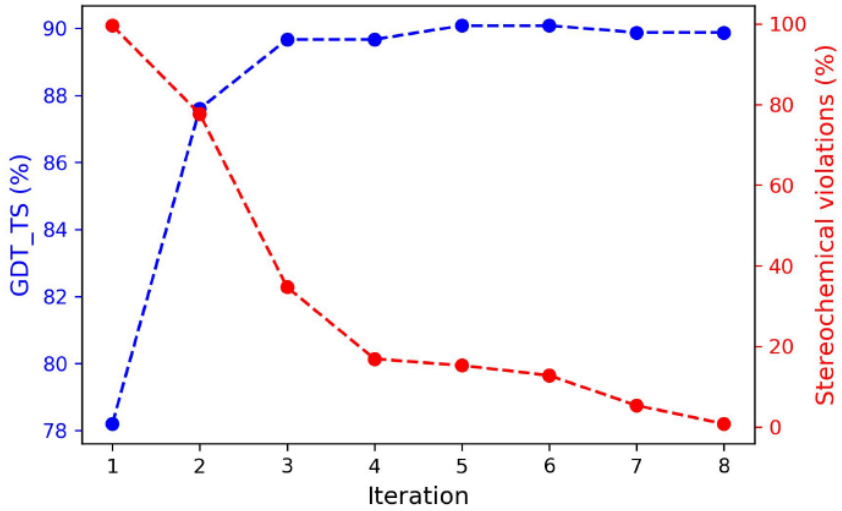
Jumper, J, et al. Nature, 2021

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: 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

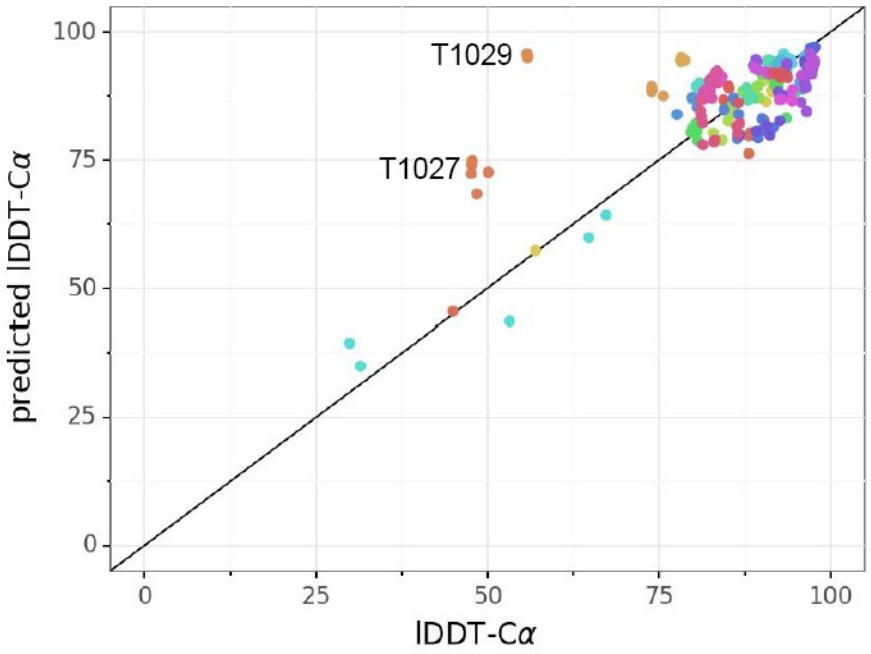
Jumper, J, et al. Nature, 2021

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

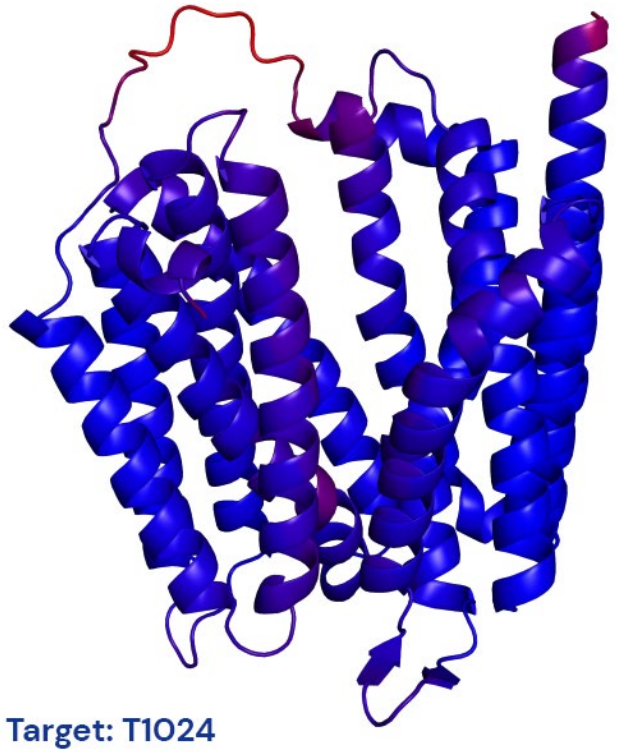
模型预测的lDDT–Cα与实际值十分接近

MAE=3.3



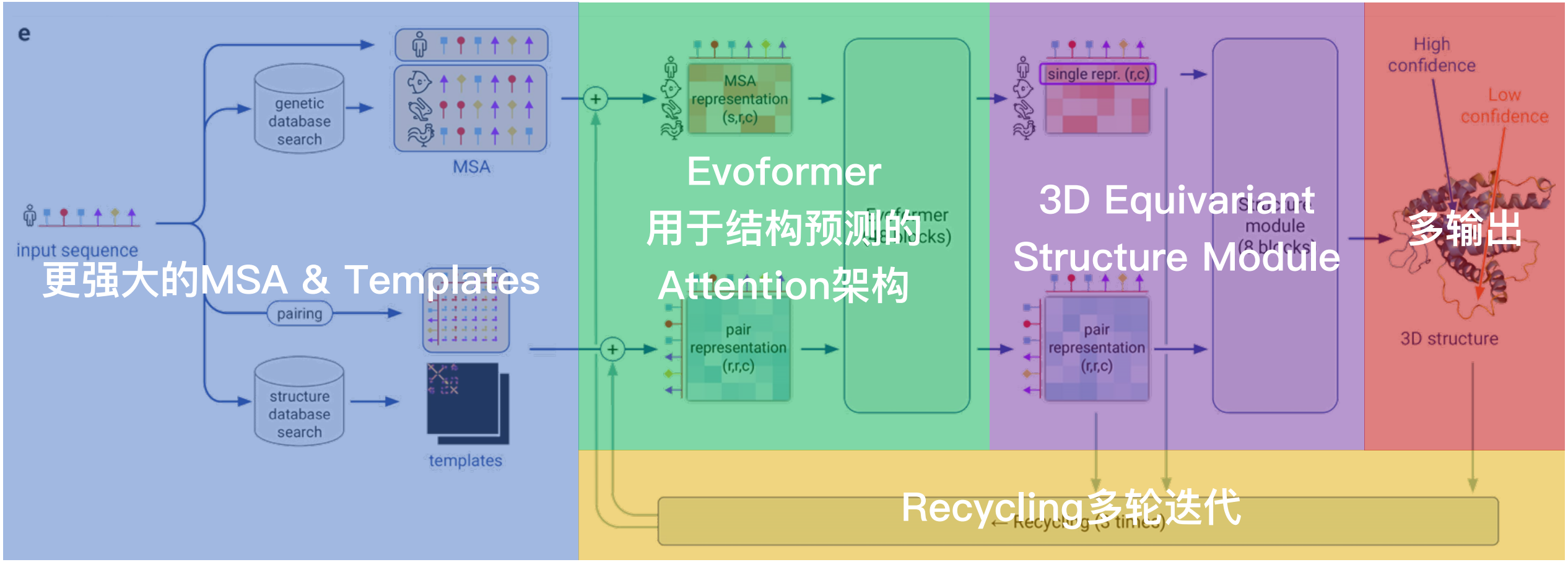
预测的lDDT–Cα能反应预测结果的准确度

（蓝色：准确度高，红色：准确度低）



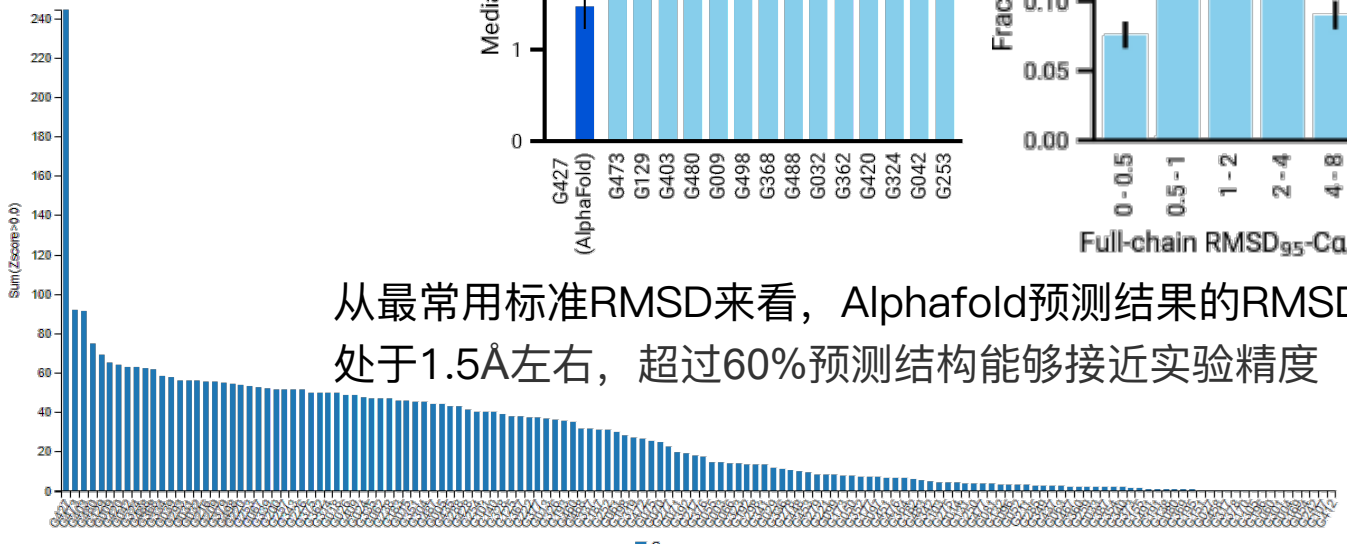
Jumper, J, et al. Nature, 2021

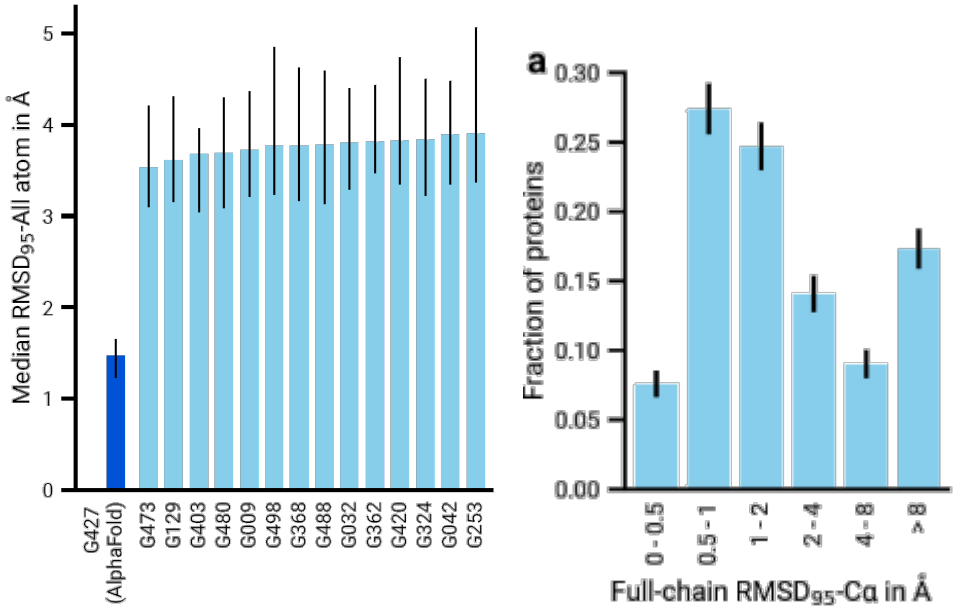
再回顾Alphafold2的整体架构



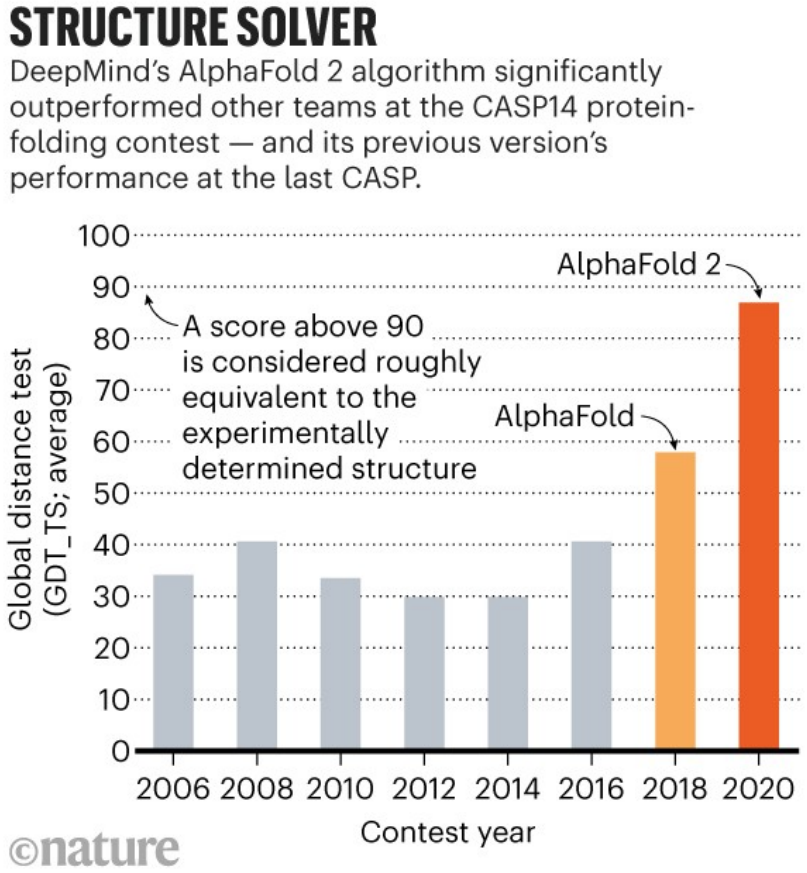
Jumper, J, et al. Nature, 2021

CASP14中Alphafold2的碾压性优势





从最常用标准RMSD来看，Alphafold预测结果的RMSD中位数处于1.5Å左右，超过60%预测结构能够接近实验精度（2Å以下）



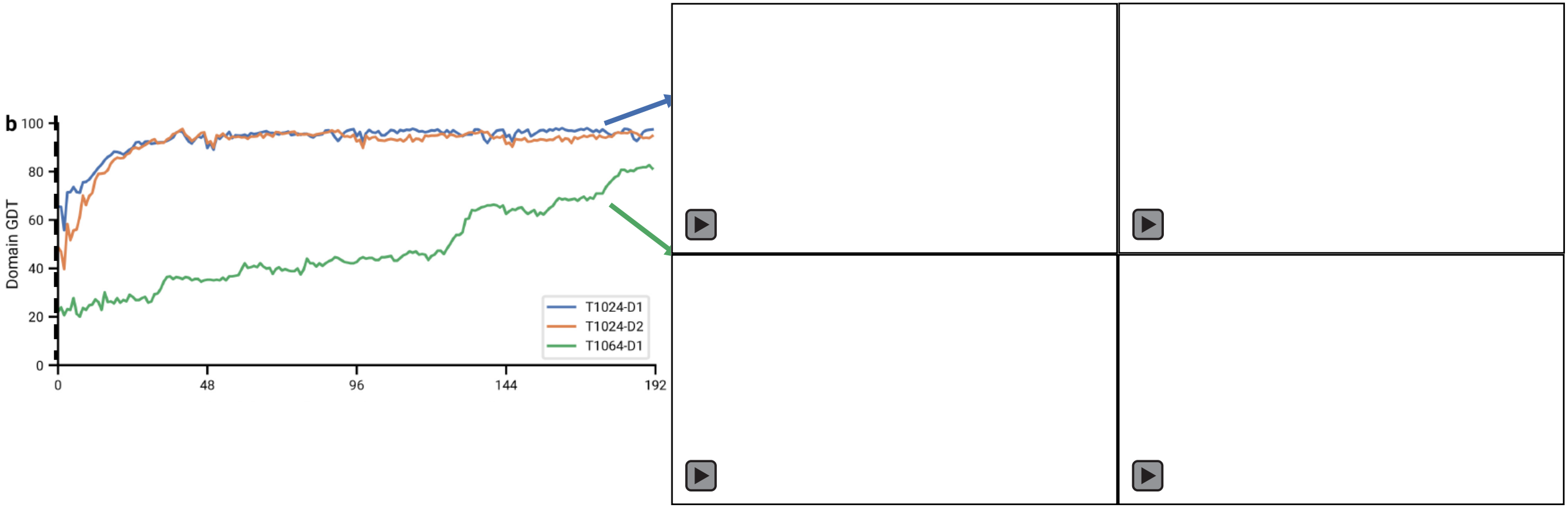
本届CASP14比赛比赛中，Alphafold的GDT–TS总分取得了碾压性的优势，远超第二名的BAKER团队

对比历届CASP的结果，Alphafold的提升是非常显著的

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Alphafold的预测结果

Picture from CASP14 and Nature
Jumper, J, et al. Nature, 2021

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

Jumper, J, et al. Nature, 2021

MSA覆盖最好高于60%，深度高于30

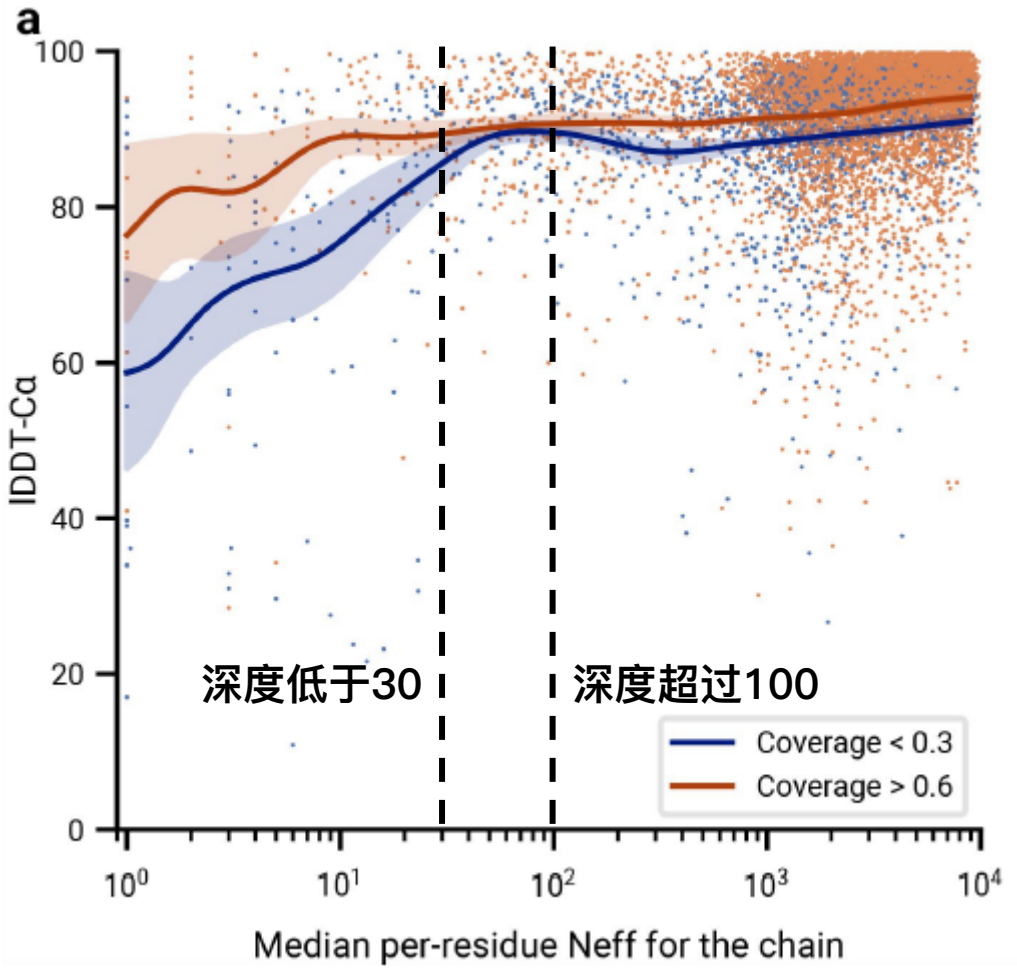
我们需要何种质量的MSA?

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MSA的覆盖率

- 覆盖率高于60%的结果明显好于低于30%的结果

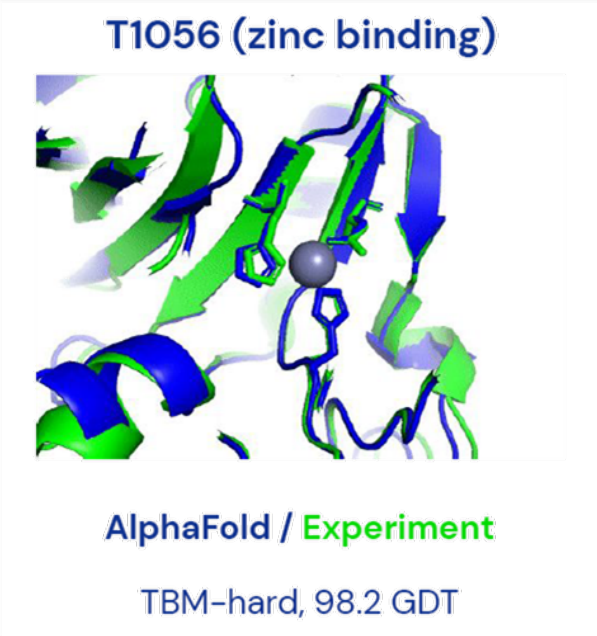


Neff: Normalized number of effective sequences

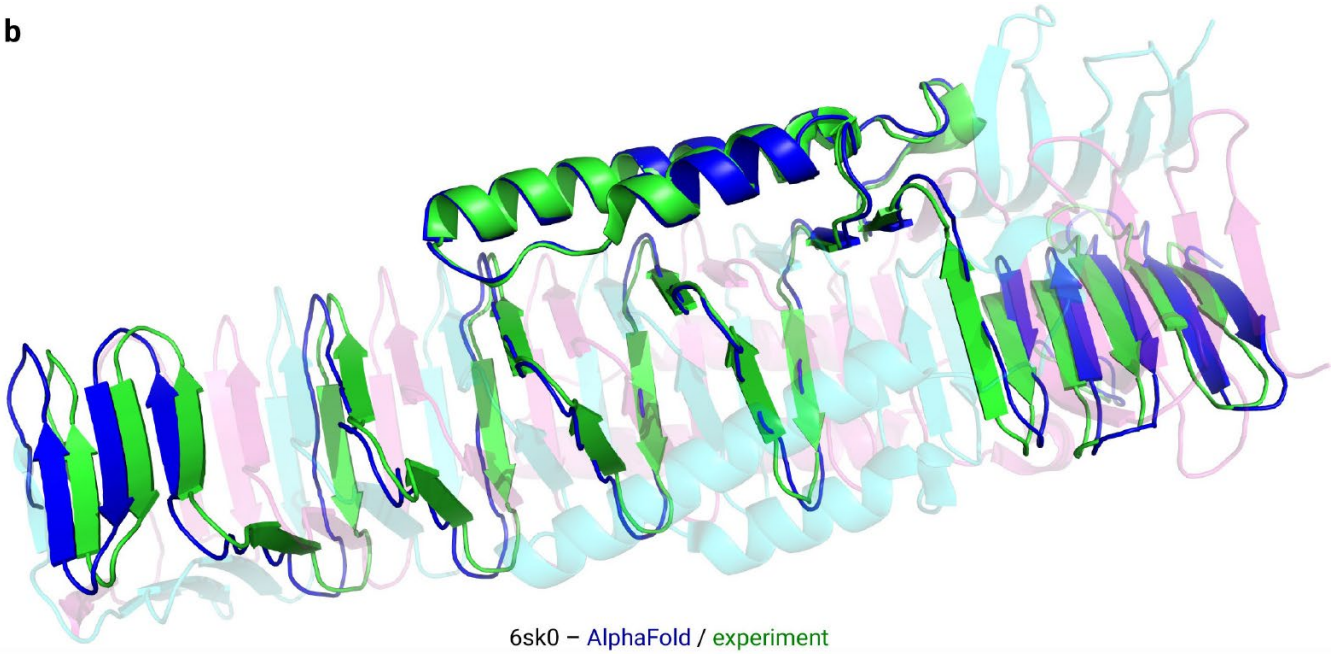
Jumper, J, et al. Nature, 2021

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

Jumper, J, et al. Nature, 2021

Alphafold 优点总结

- **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
- **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

Alphafold 的成就与不足

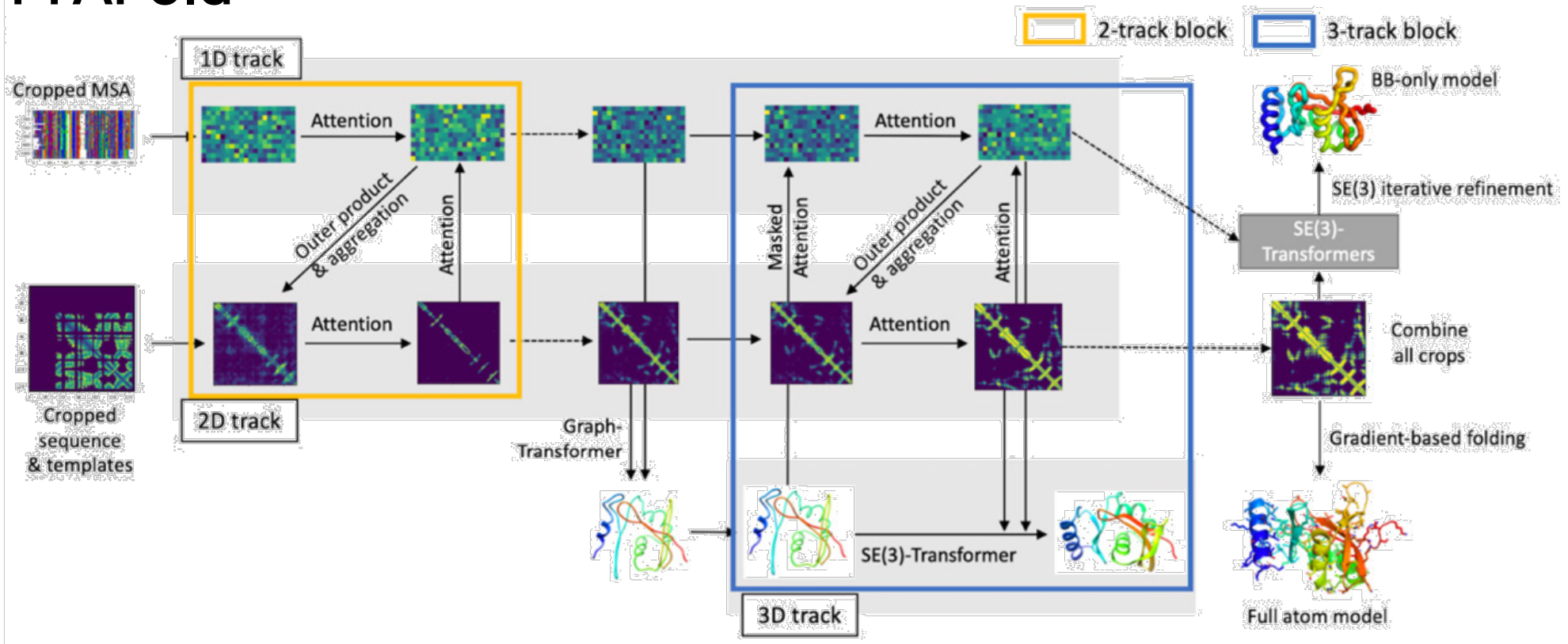
成就

- 完整建立了用于蛋白质结构预测的端到端(end-to-end)架构
- 将物理信息和几何信息融入模型，而不是使用搜索方式找到结构
- 模型能够预测自己的准确性，可以用于建模打分和排序
- 实现了计算机蛋白质建模极高的精确度

不足

- 建模输入限制于单链
- 只能建模蛋白（20种常见氨基酸），不能识别修饰、核酸、小分子、金属离子
- 本质上来说，得到的结果是晶体结构，但晶体结构并不一定能够代表真实结构

RoseTTAFold



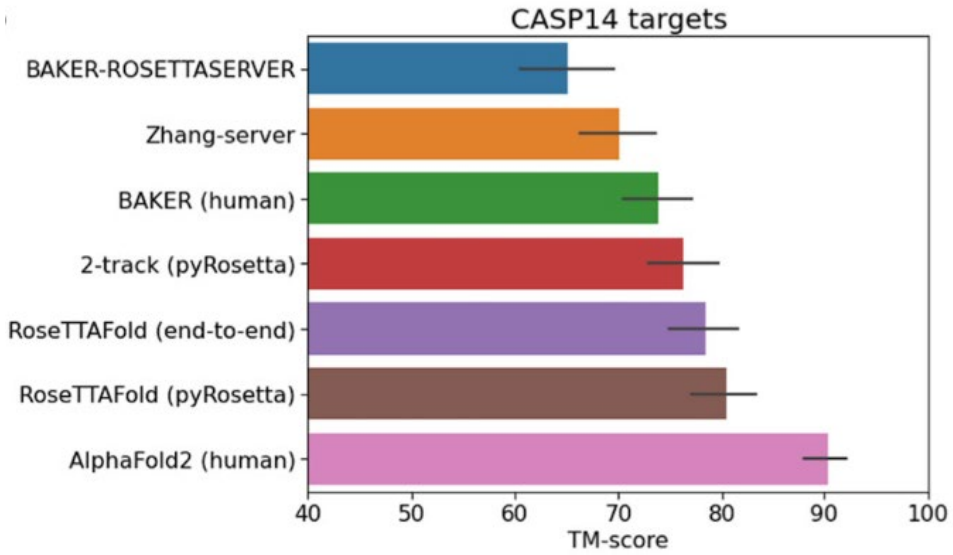
比较明显的区别：没有Recycling（2020年Alphafold并未提到）

细节上的区别：Attention模块差异巨大，SE(3)–Transformer与Structure Module差异巨大

- Alphafold中的Attention模块有很多的小trick，比如三角形优化等
- Structure Module也有很多专为蛋白质特化的设计，如Residue gas

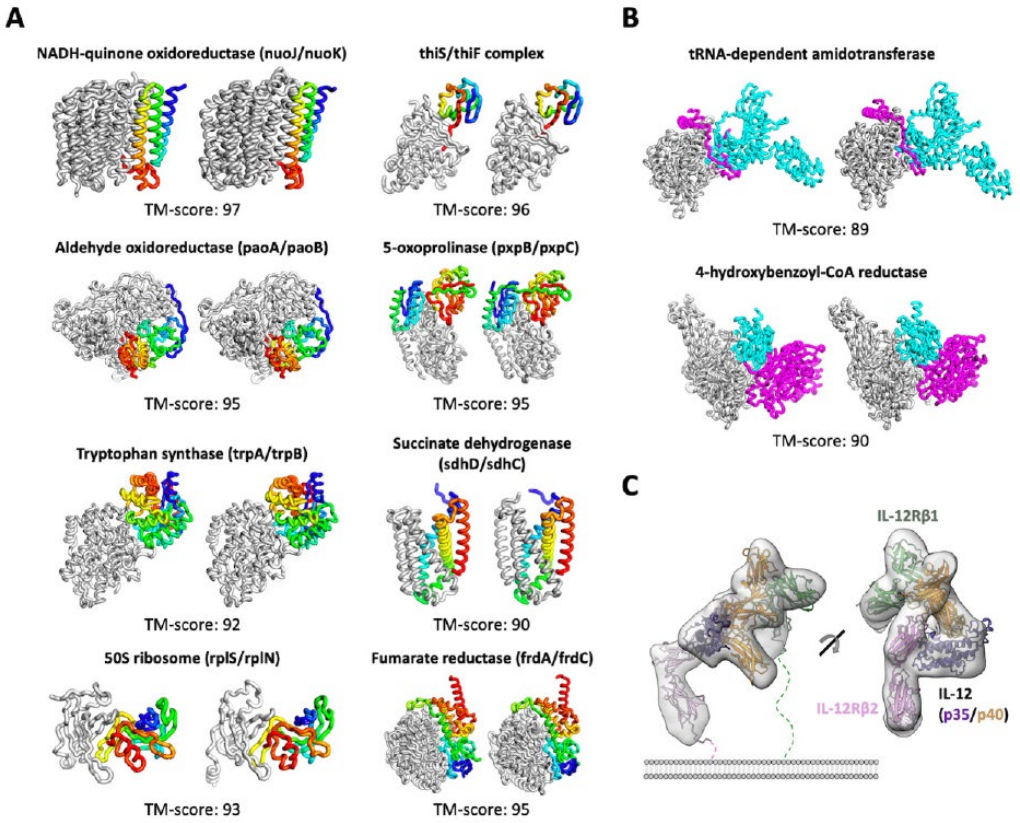
Baek M, et al. Science, 2021

RoseTTAFold结果



相较于Alphafold仍有差距（80→90）

相比于BAKER之前的方案提升并不高（75→80）



RoseTTAFold已经能够预测蛋白质复合体的结构

Baek M, et al. Science, 2021

Alphafold Github

https://github.com/deepmind/alphafold

硬件软件需求

- **Docker** (potential network issues in China)
- 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（非常巨大的算力），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：
  - 一张V100，256个残基需要4.8分钟；384个残基需要9.2分钟；2500个残基需要18小时
- 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

Alphafold 本地实现可能遇到的问题

https://github.com/deepmind/alphafold/issues

硬件要求并不高

- 运行Alphafold只需要一张A100或者1~2张V100即可
  - 显存大小是限制模型运行的因素，显存较小的显卡在运行大蛋白时可能报错
  - 多卡训练：using jaxlib 0.1.69 for CUDA unified memory
- MSA数据的存储空间：3T
  - 可以通过其他MSA方法如MMseq2来减少存储空间，不过这样会损失MSA数据量上的优势
- Model本身占用空间很小（3.5G）

Docker可能导致本地复现失败

- Docker需要root权限，或者需要管理员开启Docker权限
- Github issue中很多来自中国的科研人员提到了Docker或HHsuite的网络问题
- 现有的例子中cuda版本需要高于10.2

Alphafold 本地实现

SJTU π2.0超算已经部署：https://notes.sjtu.edu.cn/s/1clsCujfa

本地服务器已经部署

- Non-Docker实现可参考的资源：
  - https://github.com/kuixu/alphafold
  - https://github.com/kalininalab/alphafold_non_docker

Alphafold 云端实现

推荐Google Cloud实现，DeepMind的配置可以对应参考：

- This was tested on Google Cloud with a machine using the `nvidia-gpu-cloud-image` with 12 vCPUs, 85 GB of RAM, a 100 GB boot disk, the databases on an additional 3 TB disk, and an A100 GPU.

我只想试试Alphafold，还有更简单的方法吗?

Martin Steinegger(论文作者之一)和Sergey Ovchinnikov在Google Colab复现了Alphafold

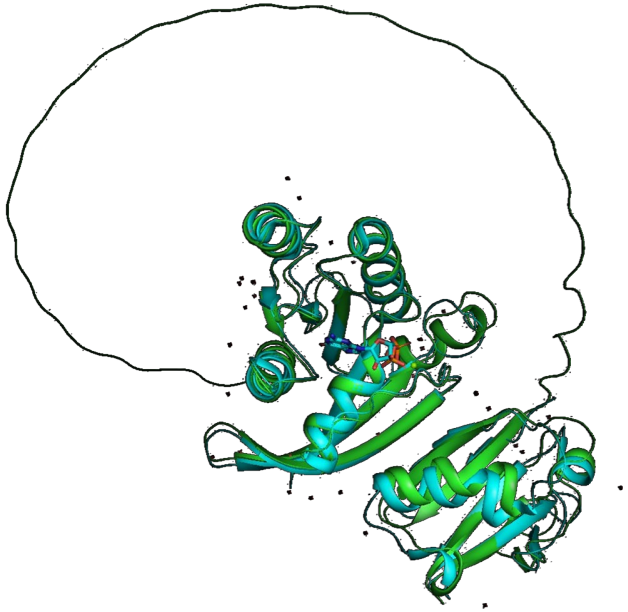
- Google Colab是基于Google云端硬盘的应用，使用共享的计算资源在云端运行深度学习模型
- **只需访问notebook，输入想建模的蛋白序列，再全部运行即可**
- 已经测试过一个长度为**79aa**的小蛋白，用时约**7分钟**完成建模
- 缺点：MSA方法使用的是MMseq2，相比于Alphafold会有一定差距
- Colab提供的硬件较差，更大的蛋白建议要用更好的显卡(V100, A100, 多卡训练等)

Google Colab 需要能访问外网

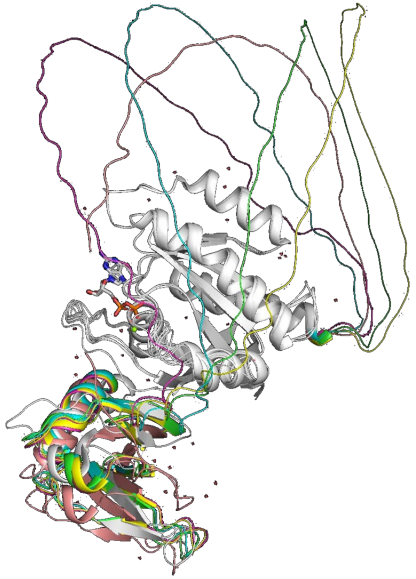
https://colab.research.google.com/drive/1LVPSOf4L502F21RWBmYJJYYLDlOU2NTL#scrollTo=qp9c7pTSdlbu

Alphafold 妙用：蛋白质复合体建模

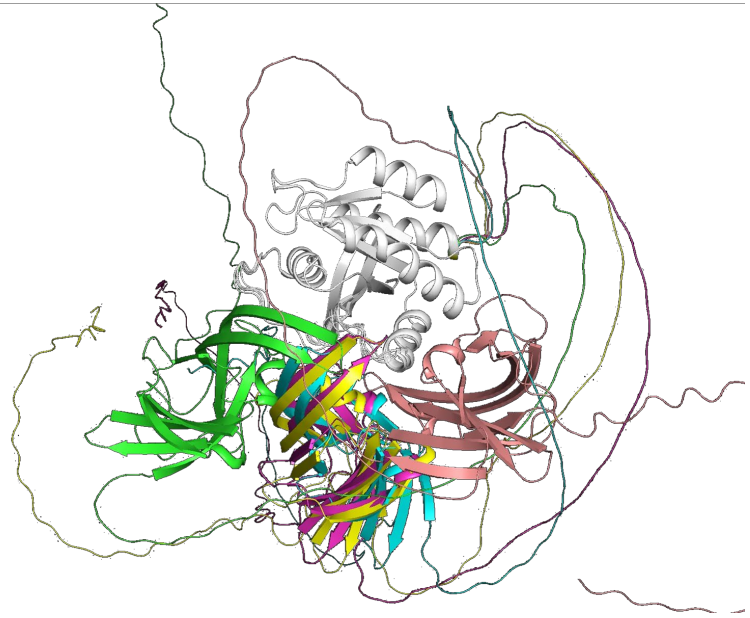
用linker连接Protein Complex建模结构



可结合的蛋白有相对稳定的复合物结构
不可结合的蛋白预测结构不一致

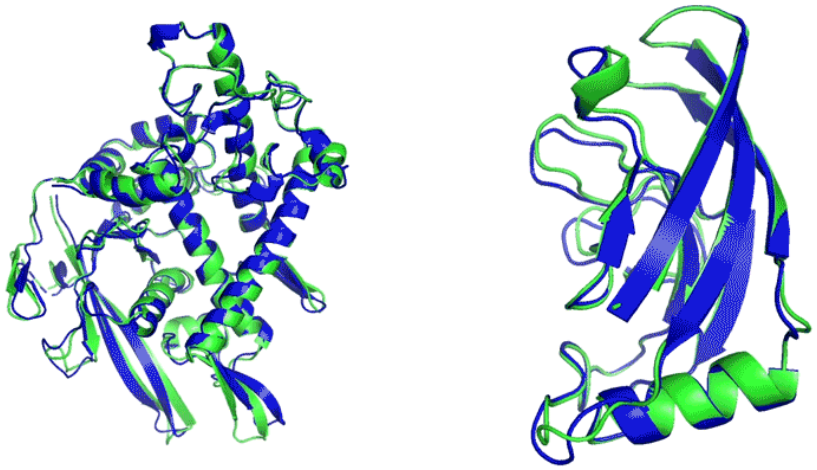


形成Complex



未形成Complex

https://twitter.com/thesteinegger/status/1415737015639764995?s=05



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

谢谢！

欢迎大家提出问题