AlphaFold 震撼发布 98.5% 的人类蛋白结构预测结果，有哪些重大突破？它们将带来哪些应用？ - 知乎

https://www.zhihu.com/question/474094187/answer/2016126950

钟博子疯子动力学模拟 ♂ Sat Jul, 24 22:39

本文正在持续更新中！

最开始说一下我认为Alphafold会对生命科学的影响吧。这篇Nature文章讲的不是技术核心，而是应用的起点，我这里当然更侧重Alphafold技术本身对未来科研的影响：

1. 结构生物学肯定是受到负面影响。“why not use alphafold”会成为大部分较为简单的结构类研究的灵魂拷问。当然，这肯定是不包括施老师等“大牛”的，Alphafold就是个一般解结构的，搞不定超大复合体、搞不定IDP、甚至暂时搞不定小分子药物。
2. 对于生命科学整体来说一定是好事。以前很多想法总是受制于没有结构（不想解、没钱解、不能解），只能从序列瞎猜关键残基。而以后没有了结构的束缚，大部分研究不再束手束脚，想法到实践之中的试错成本暴降，会有更多的发现（和论文）出现。
3. 当然，最开心的应该是会AI、懂蛋白的这群人。Alphafold为他们提供的是一个新世界的大门，门里都是全新研究领域，想抢多少抢多少。从最容易的蛋白质设计、功能预测，到全蛋白药靶筛选、药物设计（更有趣的就不说了），这些领域DeepMind只来得及开个头，剩下的全都等着你来抢。
4. 谁最难受：Alphafold的同行，这没啥好说的...

结论：21世纪果然是生物的世纪（又没说是学生物的人的世纪，毕竟你在1910年学的物理放1930年也被淘汰了啊）行业的蓬勃发展必然伴随着技术的高速迭代，技术的高速迭代就意味着从业者的技能会迅速贬值。所以，要想站在时代的浪潮头，就得尽力跟上技术的步伐呀

后面就是我细吹Alphafold的内容了，旁友们可以自助食用

（俺就是个小本科生，大家有闲心多挑挑毛病提提建议嗷！）

标题：Alphafold2: 如何应用AI预测蛋白质三维结构

摘要

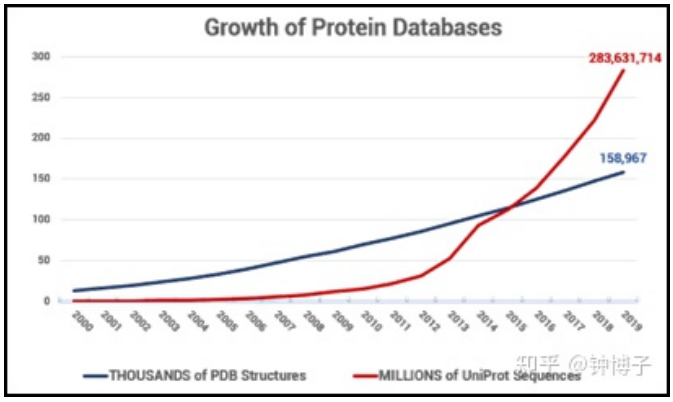
从整体来看，Alphafold可以说是在最大程度上的解决了蛋白质结构预测的问题。模型主要包含两大部件：专为蛋白质架构设计的Attention模块——Evoformer；和实现End-to-end、使模型输出能够直接是蛋白质结构坐标的结构模块(structure module)。这两大部件的精细设计，加上在各方面做到极致的训练细节，让Alphafold有着极高的结构预测精度，成为了生命科学领域中有革命

性意义的突破。Alphafold开源且公开模型参数，让人人都能用上这项技术，从而使它能得到充分的应用。这次Alphafold解析了大部分的人类单链蛋白结构，则是Alphafold展现自己威力的第一战，将会成为AI方法应用于生命科学研究的里程碑事件。

蛋白质结构预测

蛋白质是生命体发挥功能的最主要元件，几乎所有生命体功能都跟蛋白质有关，而蛋白的功能跟蛋白质结构有关。

但是十分尴尬的事情是，尽管人类通过测序技术的突破已经测得了大量的基因序列和蛋白质序列（比如耳熟能详的人类基因组计划），但蛋白质结构还是知道的很少。这里放个图对比一下差距：

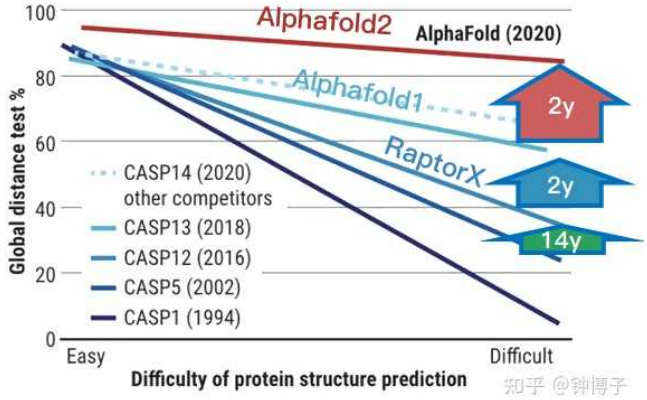


ps: 这里蛋白质数量领先只是纵轴的差异导致的，其实差距已经达到千倍了

为什么差距这么大呢？因为实验解析太困难了呀。蛋白结构解析顺利的话要几个星期，困难的话几年都解不出来。

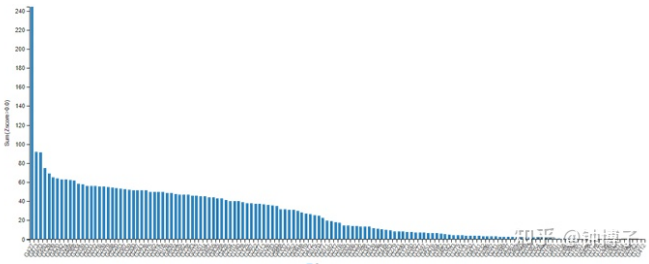
所以科学家们就想用计算的方法直接用序列预测蛋白质的结构，计算机跑得快，说不定没几年就赶上测序的速度了。

所以一群科学家创办了CASP比赛，让参赛团队们设计建模方案来做预测建模。实验室的解析的蛋白结构先不公布，只公布一个序列。等到比赛的时候再把蛋白结构拿出来，跟计算机建模的比一比之间的差距如何，能否达到一致。然而让大家十分无语的是，这个比赛在开办之后的很长时间都提升缓慢，远远达不到实验的精度。不过这一点在2016年深度学习方法入局之后发生了变化。



图源Science报道，我做了一些修改

可以看到，从2016年的CASP12开始，预测的效率开始飙升。到2020年时，Alphafold2实现了碾压了其他所有对手，在结构建模上获得了极高的进度，过半预测结果的RMSD95低于2Å，已经十分接近实验精度。看看Alphafold在去年比赛中的表现，大概是可以用“断层出道”来形容了（下图最左边那个是Alphafold2）。

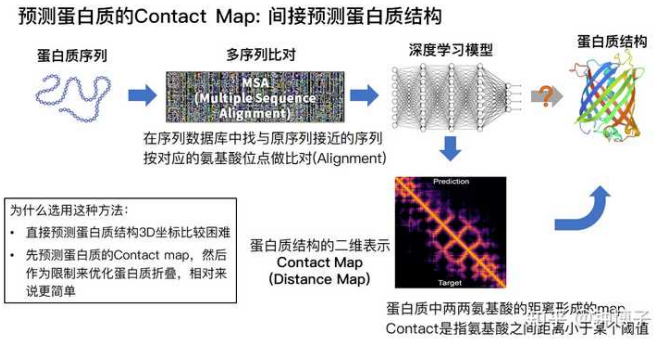


图源CASP14官网。看到这结果只能说：“既生Baker，何生Jumper”了

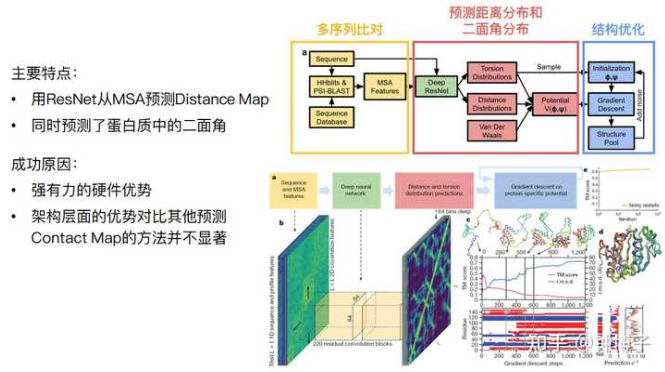
看到这里，相信你可能已经对蛋白质结构预测有了那么一点了解了（不就是用给电脑序列，电脑给我结构嘛，就跟siri听懂俺说话，再按照我的意思给我定闹钟差不太多）。现在所有人好奇的问题就是，DeepMind他们究竟是怎么造出Alphafold的？如同：三体人究竟是怎么造出水滴的？不过这次的结局不再是丁仪的“傻孩子们，快跑啊”，而是三体人亲自送来水滴帮你技术爆炸。

半年过去，Nature在7月16日公开了Alphafold的论文。技术细节的公开，代码开源，参数可下载，人们终于知道了为什么Alphafold为什么能这么强。尽管同日在Science公开的RoseTTAFold整个事情显得十分戏剧化，也不可能掩盖掉Alphafold作为主角的光辉。

在Alphafold2之前，大家都在预测Contact Map，Distance Map，Torision Angle等等，走的是间接预测一些结构限制(constrains)再回来优化蛋白结构的操作。Alphafold一代也就按照这个架构做了，靠算力也拿了个第一，但是这样的第一DeepMind怎么能满足呢？

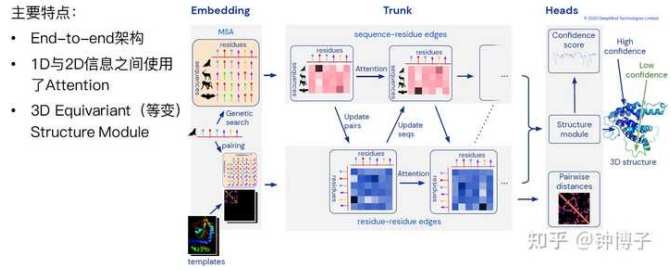


原本大家都是这么搞的（图源我的PPT）

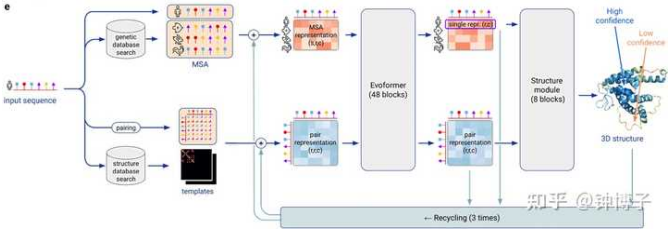


Alphafold一代不过如此（图源Alphafold一代论文）

Alphafold2就抛弃了之前的方案，就是要做端到端，要把物理信息加入模型，要上Attention。用上最好的架构，再在每个小环节都做到最好，就是现阶段近乎完美的解。所以他们就推掉原来的架构，（大概只保留了MSA作为输入这一点），换上了这样的架构：



Alphafold2在CASP14上公布的架构（兄弟们记住这个图，后面有人抄作业可以对比！）



Alphafold2论文中的架构（欢迎跟上图找茬）

我给Alphafold总结了5个精彩之处，正是靠着这些精彩之处，Alphafold才能最终拿下结构预测的王座（当然，这个精彩之处仅代表个人观点嗷）

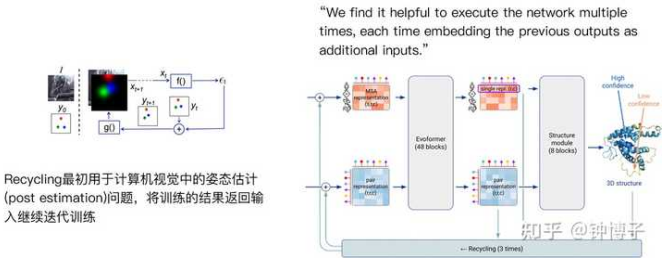
精彩之处一：模型输入——更强大的MSA & Templates

这一点非常简单，Alphafold搞来了尽可能多的数据库来做多序列比对(MSA)，牛要长得好，吃草就得吃得饱。模型的输入足够多才能满足Alphafold的检索特征的需求。

我累了不想码字了...直接挂PPT吧

精彩之处二：使用Recycling进行多轮迭代训练和测试

我第一个找的茬就是Alphafold在CASP14上没标的Recycling步骤。查了查论文发现这个步骤最初是用于做计算机视觉中姿态估计的，被他们搬来这个架构了

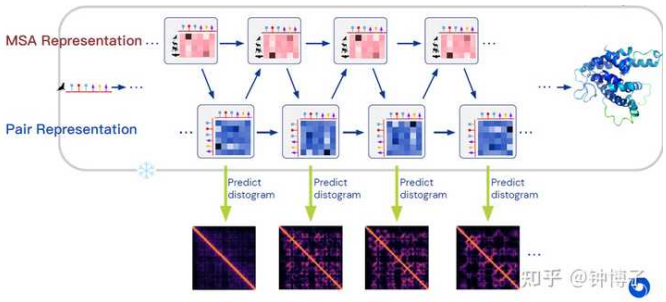


就是右图最下面那个

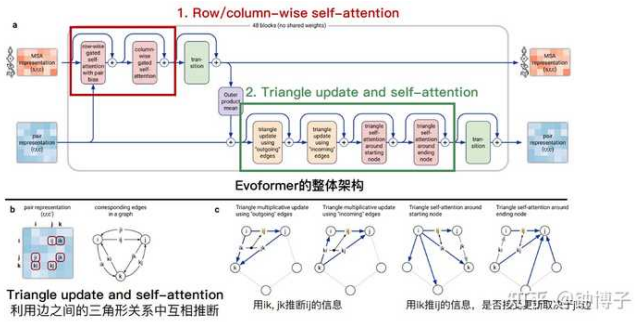
精彩之处三：Evoformer——用于结构预测的Attention架构

其实到这里才是Alphafold的主菜：Attention

这里Alphafold用了两个表示来表示它拥有的信息：一个是MSA Representation来表示MSA中的信息，另一个是Pair Representation来表示残基间的相关性的信息（有点像contact map，但又不只是contact）。在Evoformer中，这俩信息互相update，最后得到一个优化的足够好的embedding。



Evoformer中的两个representation互相update



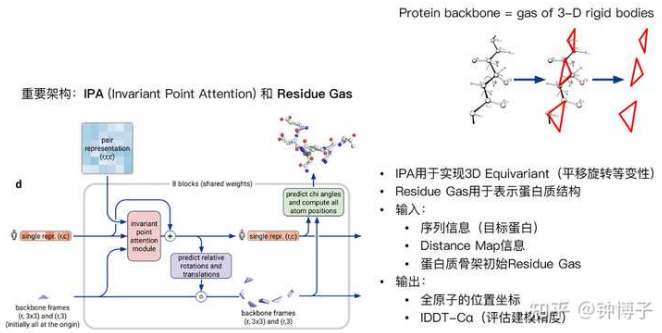
Evoformer真的是细节拉满

Evoformer整体来说细节拉满。在MSA一端做每行的self-attention，做每一列的self-attention。更新过来之后再在Pair这边先用三角形反复自我更新，再来三角形self-attention。反复的attention过程让两个representation都能够达到足够好的状态用于后续的结构优化。

这里聊一聊三角形update吧，感觉算是Evoformer中很有趣的一部分，也是Alphafold在精细上的一个体现。这里他们没有用各种粗糙的方案，而是从三角形的特性入手，来设计了一个方法互相更新。三角形中任意两条边的信息当然是能够影响第三条边的，举一个最简单的例子，两边之和大于第三边，就是两条边的信息对第三条边的限制。Evoformer中的triangle multiplicative update就是借鉴了这样的思路，用两条边去更新第三条边的信息。

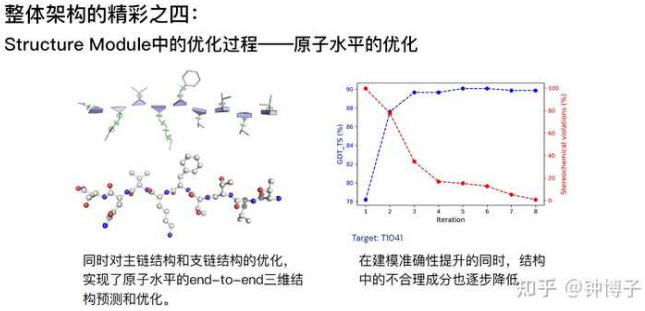
【Triangle self-attention懒得讲了，思路差不太多】

精彩之处四：Structure Module的关键——Equivariant



Structure module的目的就是让模型的输出可以是蛋白质的每个原子的坐标。要做到这点可不容易，他们认为需要实现3D-Equivariant，才能很好的实现这个模型。如何做到在模型中平移旋转等变，其中的奇妙就在这个IPA(invariant point attention)模块中

当然，他们还做了一个抽象化，把蛋白质主链中抽出了一个三角形（图右上），称为residue gas。在结构模块中，他们先用在IPA中更新，再用等变性更新这个residue gas，从而不断优化就用residue gas表示出了蛋白主链的结构。有了主链的位置，还有侧链的取向，预测全原子坐标自然轻而易举，补了一个小模型就得到了蛋白质的全原子坐标了：

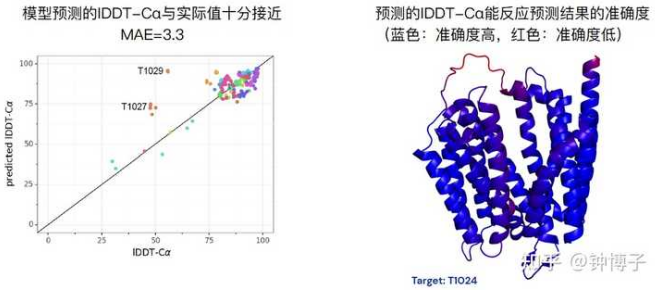


从而，在structure module中就实现了原子水平上的结构优化。对了，这里不仅是让结构更接近真实结构了（提升准确率），同时也纠正了结果中的错误(violations)（降低错误率），这样refinement也就不在需要那你，直接可以输出结构啦！

（其实事实上还有一步AMBER优化，但这一步不重要）

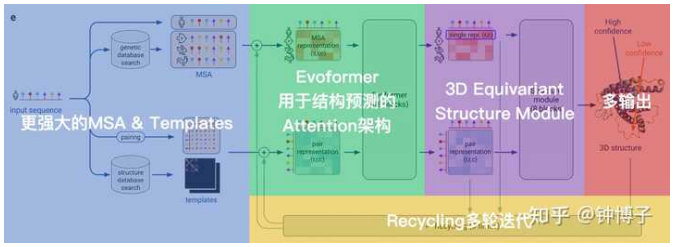
精彩之处五：多输出——如何知道预测的结构的精确度

最后一个精彩之处，我Alphafold不仅能预测，还能预测我预测的准不准。（禁止套娃！）也就是说模型还会自己给自己打分（lDDT-Cα），告诉用户：我这里预测的不错，你大可放心；这里预测的不太行，你要谨慎一点：



“是谁预测了我？而我又预测了谁？”“是我...预测了我！”“回答正确”

好了到这里就结束了，让我们回顾一下这个架构：



选个颜色真难

当然还有一些没提到的比如Self Distillation的训练策略，还是建议大家读读原文，这就不赘述了

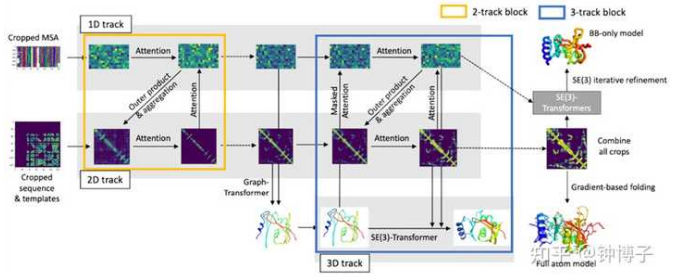
最后总结一下Alphafold的成就：

- 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

当然也存在不足（我只能吹毛求疵的找了呀）：

- 建模输入限制于单链
- 只能建模蛋白（20种常见氨基酸），不能**直接**识别修饰、核酸、小分子、金属离子

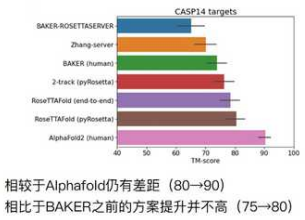
最后来瞅一眼RoseTTAFold

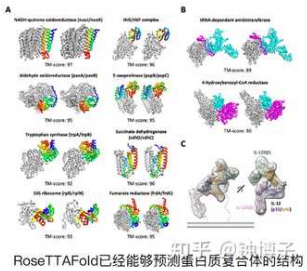


跟2020年Alphafold那张图像吗?

相似的MSA Represention（这里的1D track），相似的Pair Representation（这里的2D track）。相似的1D和2D互相update，加上self-attention（就Evoformer）。相似的把1D和2D信息交给structure module，还都具有旋转平移等变性。只能说，贝老师复现的好啊!

可惜呢，临阵磨枪的时间有点短，（可能人手也不太足？）很多细节没法自己优化了。Evoformer没有了三角形优化、SE(3)-Transformer没有了Residue Gas，没有Recycling，也没有精心设计的loss function，最终结果也难随人愿。





RoseTTAFold说我能建模复合体，但Alphafold也没说自己不能啊 [手动狗头]

结果如何呢？RoseTTAFold考了80，Alphafold当年可是90，还是差了些。更不爽的是，当年贝老师已经考虑75了，“致敬”End-to-end之后涨分5分，略有一点不够意思。

Alphafold效果如何

【请听下回分解！】