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/2015916927

郭昊天生物学话题下的优秀答主 Sat Jul, 24 22:45

凡康的答案里面已经对AlphaFold2的工作做了很详尽的介绍了，各种优点我就不多赘述了。我来补充一些评论吧，回答一个“如何看待”的问题。

对于不想看长文讨论的同学而言，本文可以缩短成以下几个部分

1. 结构生物学家不会失业，相反，仍然会有很多工作

2. AI+biotech的团队有的会受到冲击，有的会开始起飞

3. AlphaFold2通过对人类蛋白质组的分析，展现了很多局限和不足

3. 蛋白质结构预测的圣杯仍然没有被摘下，**大家一起努力工作吧!**

1. AlphaFold2对业界的影响

1.1. 结构生物学家会失业吗?

先说我的结论：绝大多数结构生物学家都不会失业。

我们可以粗暴地将结构生物学家一分为二:

一种是做各种各样复杂的多组分蛋白质复合体，以及蛋白质动态结构的，哐哐发CNS，典型的例子就是施一公老师（虽然因为各种原因，施老师现在比较招黑，但是u1s1，人家结构生物学的工作还是很漂亮的）。这类工作目前还不能够被AlphaFold2所替代，原因我们先按下不表，我们后面来讲AlphaFold2目前的一些局限和不足。

另外一种就是去解析简单蛋白质的结构生物学工作，或者是以灌水的形式发表文章，或者是以和其他或基础研究或应用工作合作，作为文章的一小部分。这一大类工作，从历史经验上来看，也不会被AlphaFold2取代。

在过去AlphaFold2还没有出现的时候，其实各种计算结构生物学工具都已经能预测各种各样的蛋白质结构了，其中很多结果其实还是相当精确的。低复杂度的问题可以用比如说同源建模的方法，问题复杂度更高的可以使用trRosetta（也就是AlphaFold决战紫禁之巅的RoseTTAFold的前作）。但是蛋白质结构测定的工作，仍然只有极少的部分从结构生物学家的手中移交到了计算结构生物学家的手中。

一个表面的原因是，这些算法再好，仍然是有缺陷、有不准的时候。我们不做实验，怎么知道预测出来的结构对不对呢？但实际上在其他的科研领域，我们很容易就找到反例。比如说在合成生物学，有各种各样的计算工具，预测精度差得令人发指，大家仍然用得乐此不疲。

所以我觉得，实际原因，是科学家的共识问题。归根结底，在于科学共同体也是个社会小圈子。所有的科学活动都建立在同行评议的基础上。那么如果大家都觉得，你需要做实验解结构，那么就算有再好的工具，你发文章的时候也要解结构。这个习惯，会在一次又一次的科学发表和同行评议中被不断强化。

只有一代人都退出历史舞台之后，新的共识才能逐渐形成。

> Science progresses one funeral at a time.

当然，改变还是会存在的，解析简单蛋白质就能发CNS的时代，是一去不复返了（比如说抢发Covid-19各种结构的盛景）。

1.2. AI+Biotech

在我印象中，计算生物学使用AI相关技术由来已久。但是从DeepMind的AlphaGo暴打Lee Sedol开始，才有越来越多的人，开始讲一个AI和生物学的story。很惭愧，我也不能免俗，从2016年开始在合成生物学里面大量地使用机器学习的技术。包括今天我们创业做startup，AI也是我们的核心能力之一。只能说，真香！

到了可能大概是2018年前后吧，就冒出了一大堆AI+的生物技术创业公司。其中大家经常听到的相当一部分是AI制药的。

我觉得这些AI+生物的团队，不管是企业，还是科研院所的实验室，也可以粗暴地分为三类：

第一种，核心能力是AI，全部的工作内容是写AI。这部分团队会主动地更卷了。

比如像DeepMind，Baker lab，这种团队后续可能还会接着卷。像去年CASP14结束之后，Baker组根据放出来的一点点信息，就很快借鉴作业，搞了RosseTTAFold。虽然说预测准确度整体还是不如AlphaFold2吧，但是也有其他的优点，可谓各有所长。那么AlphaFold3是不是就不远了？

第二种，核心能力是AI，绝大多数的工作内容是用AI。这部分团队的要被迫更卷了。

原来这些团队里面也会有很大一部分人做算法开发。然后大家再用自己的算法和模型，做各种各样的蛋白质结构预测，在计算机上筛选小分子或者大分子药物，做docking等等。

除了极少数的能拿到非公开数据，以前大家都是靠公开数据库吃饭，那么各家比拼的就是算法和算力了呗。现在AlphaFold2出来之后，会让相当多团队的算法水平，起码在蛋白质结构预测算法上，被拉到同一个起跑线上。但是其他方向的算法护城河也未必就守得住。比如说，以今天

AlphaFold2开源的内容来看，小分子或者大分子结合态的结构预测，拿AlphaFold2稍作修改完全可以搞得定。甚至我们可以设想，如果DeepMind一直研发并继续开源下去，那用不了几年，大家比的就是抄DeepMind的速度有多快了。

最后大家就是拼算力了。但是……谁会有Google算力更强呢?

第三种，核心能力是生物，一部分的工作内容是用AI。这部分团队喜大普奔，能用的铲子更多了。

这种团队里面，做计算的是一部分人，还有另一部分人是做实验的。这些做计算的人，每天又要做算法开发，又要做应用，偶尔还会被抓去实验室做搬砖的壮丁，忙得要死。现在好了，有了AlphaFold2，调包专心做应用就可以了，极大解放劳动力。以前可能要几个月才能搞定的事情，现在一两天就做完了。

1.3. AlphaFold2发布这么多蛋白质预测结果，是在抢人饭碗吗?

这个事情吧，我觉得还真的不是AlphaFold2非要发一大堆预测结果，出来吊打结构生物学家，或是各种AI+的同行。

本质上，这就是个常规操作。AlphaFold2利用大量的PDB数据进行训练之后，大家发现AlphaFold2在PDB数据上面表现很好，那自然而然地，大家就会关心：这个模型的泛化能力怎么样?

最好的测试方法，就是那一大群模型训练的时候从来没见过的数据，来测试AlphaFold2的表现。

那么…既然这些蛋白质都预测了，既然都要开源了，为啥不再灌一篇nature呢? 【狗头】

2. AlphaFold2的局限和缺点

AlphaFold2虽然强无敌，但是不得不说，还是通过这次人类蛋白质组的测试展现了很多局限性和缺点的。

我个人认为，这篇paper的学术意义甚至大于AlphaFold2的介绍（毕竟代码开源比介绍更直给）。通过AlphaFold2在非常广阔的蛋白组中的表现，暴露它的局限和问题，可以帮助我们指明未来的方向。

由于AlphaFold2还是比较吃算力的，这样的工作，如果不开源，大家要花非常多的时间和资源才能重复出来。通过分析这些局限，我们可以更清楚地看到未来大家可以研究什么。

2.1 吃硬件

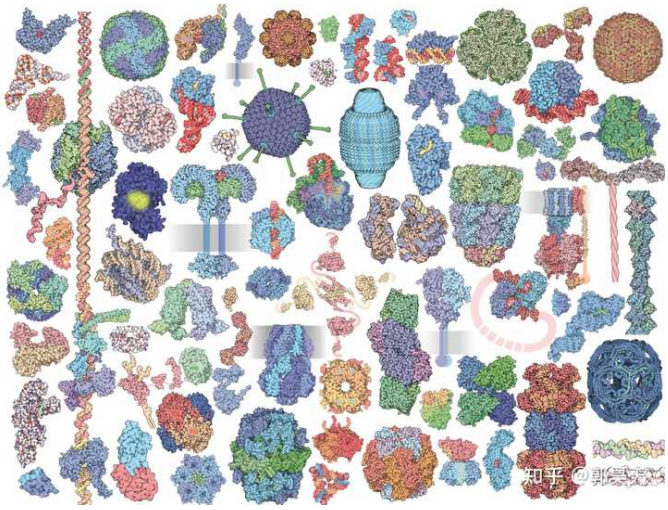
相当多的计算生物学和生物信息学软件都是一个普通的笔记本就能跑得起来的。

而AlphaFold2的部署，需要“3TB储存空间、85GB内存、和Navidia GPU”。

相比而言，RoseTTAFold可能会更友好一些，一个“普通”的高配电脑就能带得动。这可能会导致RoseTTAFold能更快地被封装成，并广为传播。

2.2 过于大的蛋白质或者蛋白质复合体，可能会跑不动

虽然说，从原理上来讲，AlphaFold2对单链的蛋白质或是多组分蛋白质，应该都能跑得起来。但是实操上，多组分的蛋白质复合体，原子总数要比单链蛋白质一般大得多得多。现在AlphaFold2的测试上限是不到3k个氨基酸，那么相当多的蛋白质复合体，是超过这个数字的。



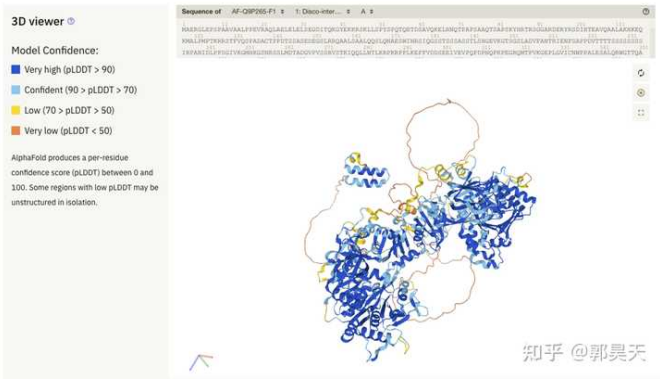
以PDB 101精选图为例，上面画的蛋白质复合体，一个赛一个大

2.3. 对无规结构毫无办法

AlphaFold2并没有整合多少先验知识。对于经典的各种二级结构，蛋白质的结构空间相对是比较有限的，而PDB提供的样本量足够多，因此AlphaFold2的预测结果很好。

但是对于无规结构，那一串氨基酸想怎么飘就怎么飘，稀奇古怪五花八门的结构都存在与天然的蛋白质构象当中。当可能性的多样性过高，这时候PDB提供的这点数据就远远不够了。相比而言，传统基于物理的方法可靠性反而要高得多。

结果就是在无规结构的预测上，AlphaFold2普遍放飞自我了。

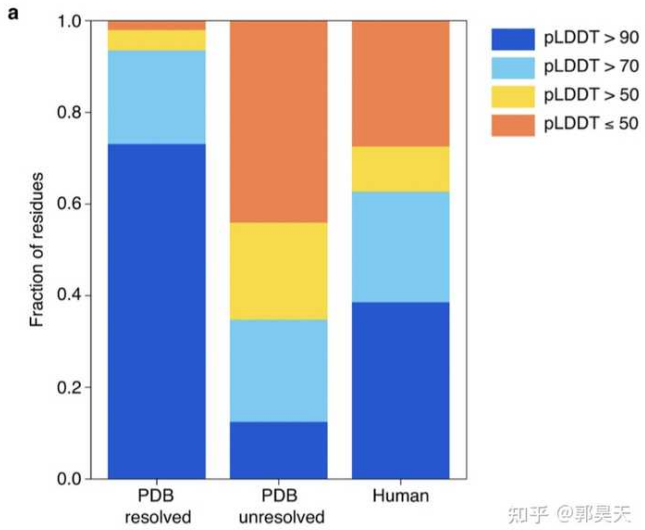


该有结构的无规序列，都变成了大大的圈圈

我们能得出什么结论？无序蛋白会变成后AlphaFold2一个非要重要的议题 —— All in 相分离！【狗头】

2.4. 可能对PDB数据存在一定程度上的过拟合

前面讲到，大家用PDB数据训练并测试了AlphaFold2之后，自然就会考虑测试AlphaFold2的泛化能力了。通过AlphaFold2自己对预测结果的置信度分析来看，结果只能说非常的不理想。



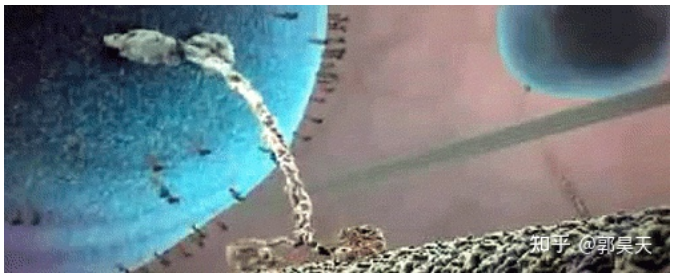
PDB上有数据的蛋白质，AlphaFold2普遍预测结果能高达90%的置信度；而对于PDB上没有数据的蛋白质，AlphaFold2的预测结果，有将近一半置信度不足50%

这个也比较好理解。PDB有数据的蛋白质，和PDB上没有实验结果的蛋白质，是两个高度有偏的数据集：统计上来讲，很可能越是容易被解析的蛋白质，PDB上的数据可能越多，数据质量也越好；而越难被解析的蛋白质（难以蛋白表达、纯化、或者分析结构），PDB上的数据就越少，数据质量也越差。

那么AlphaFold2可能学习到了很多feature，是容易被解析的蛋白质所共有，但是缺很难泛化的特征。也就说，这些是PDB数据的feature，而不是天然蛋白质的feature，更不是所有蛋白质的feature。上面讲到的AlphaFold2对无规结构的无力，也可能与此有关。

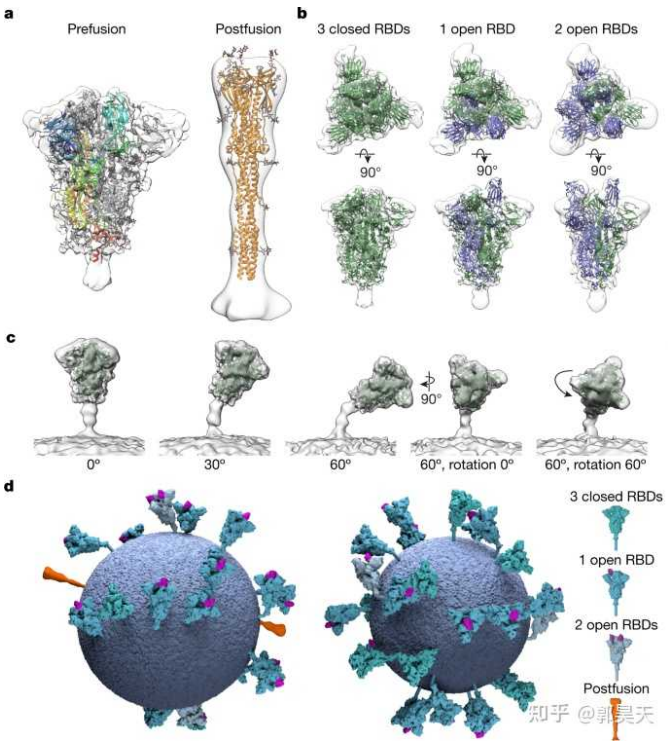
2.5. 大量的蛋白质拥有非常动态的结构，目前AlphaFold2只能预测静态解

这个说起来，也得说是一个PDB数据的feature，而不是蛋白质的feature。早年的大量数据是来自晶体学的，那么一个蛋白质满打满算也不会有多少构象被解析出来。但是实际上很多蛋白质的结构是高度动态的，而并非一个静态的稳定的结构。



比如说经典的kinesin，是前脚跟着后脚，反复向前走

随着冷冻电镜的开发和使用，有越来越多蛋白质的柔性动态数据被揭示出来。最近的例子就是新冠病毒Covid-19的S蛋白[1]，就有柔性结构，而且和其免疫抗原性息息相关。



Structures and distributions of SARS-CoV-2 spike proteins on intact virions

All in 冷冻电镜！【狗头】

3. 计算结构生物学的圣杯被攻克了吗?

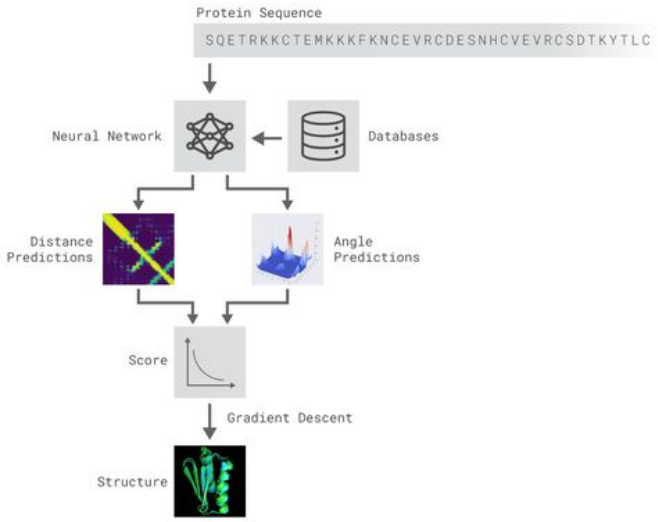
AlphaFold2刚在CASP14问世的时候，大家普遍都讲什么蛋白质结构的问题已经被解决了。

我觉得今天这篇人类蛋白质结构预测的文章很大程度上能让大家回复理智——蛋白质结构预测其实还远远没有被解决。仍然有非常多的问题亟待解决。

大家可以对AlphaFold2的预测结果进行分析，从而可以更好地优化AlphaFold2。而这么大量的数据如果要分析的话，肯定也少不了AI的应用。

仍然可能有会有新的算法突破。我在AlphaFold 1问世的时候就讨论过不同路径的可能性。我今天仍然认为，先验知识是有用的，以蛋白质序列空间的复杂性而言，完全的数据驱动是不能解决所有问题的。

[如何看待 AlphaFold 在蛋白质结构预测领域的成功？ www.zhihu.com]



而且，今天这个实验数据数据量仍然还是不够。**结构生物学家们，请继续努力！**

参考

1. ^Structures and distributions of SARS-CoV-2 spike proteins on intact virions https://www.nature.com/articles/s41586-020-2665-2