·科技评述·

Science 2021 年十大科学突破解读

［编者按］ *Science* 杂志每年会在年底评选出当年十大科学突破。2021 年 12 月 16 日，*Science* 杂志公布了其评选出的 2021 年十大科学突破。为了让广大读者更深入地了解十大科学突破的科学价值，本刊特邀各领域著名科学家分别对其进行解读，以激发科研人员的创新思维，促进学术交流。

DOI:10.16262/j.cnki.1000-8217.2022.02.059

1 人工智能预测蛋白质结构

2021 年 7 月，世界知名人工智能团队深度思维宣布，已经利用 AI 智能软件程序——阿尔法折叠预测了人类表达的几乎所有蛋白质的结构，以及其他 20 种生物几乎完整的蛋白质组。AI 预测蛋白质结构将实现广泛应用，提供对基础生物学的见解并揭示潜在的药物靶点。同年 8 月，中国研究人员使用阿尔法折叠 2 绘制了近 200 种与 DNA 结合的蛋白质结构图。同年 11 月，德国和美国的研究人员利用阿尔法折叠 2 和冷冻电镜绘制了核孔复合物的结构图。现在，科学家正使用阿尔法折叠 2 来模拟奥密克戎变体刺突蛋白突变的影响。通过在蛋白质中插入更大的氨基酸，突变改变了它的形状，这也许足以阻止抗体与其结合并中和病毒。

专家点评：

张强锋 博士，清华大学生命科学学院副教授、博士生导师，清华大学—北京大学生命科学联合中心研究员，国家杰出青年科学基金获得者。主要从事结构生物学、基因组学、人工智能和大数据交叉领域研究，以通讯作者身份在 *Cell* 等杂志上发表学术文章多篇。

徐 魁 博士，清华大学生命科学学院助理研究员，专注于采用人工智能方法解析蛋白质冷冻电镜三维结构、RNA 结构等相关生物学问题，以第一作者在 *Cell Research*、*Nature Machine Intelligence*、*Bioinformatics*、*AAAI* 等期刊和会议上发表多篇论文。

从蛋白质分子的序列出发准确预测其结构是分子生物学的“圣杯”问题之一，具有高度的科学和应用价值。2021 年 11 月 17 日，*Science* 杂志公布了年度十大科学突破榜单，其中“人工智能预测蛋白质结构”第二次入选；该进展同时也被列为 *Nature* 杂志 2021 年十大科技新闻之一、*Nature Methods* 杂志 2021 年度技术。

基因组是编码生命信息的蓝图，而蛋白质是将这张蓝图付诸实施的工人。蛋白质执行生命活动功能的基础是其折叠形成的独特结构。解析蛋白质结构对探索生命活动分子机制和药物研发都具有重大意义。目前，解析蛋白质结构主要依靠 X 射线晶体衍射（X-ray Crystallography）、核磁共振（NMR）以及冷冻电镜（Cryo-EM）等实验技术，这些技术一直面临着通量低、成本高、周期长等问题，严重制约了解析蛋白质结构的速度。截至当前，已知的十几亿蛋白序列中仅有不到二十万对应的结构被实验解析。

幸运的是，早在 20 世纪 60 年代，Christian Anfinsen 就通过实验证明蛋白质的序列决定了其结构（Christian Anfinsen 因此获得 1972 年诺贝尔奖）。求解蛋白质序列和其三维结构之间的映射关系，即蛋白质结构预测，成为诸多分子生物学家的梦想。然而，这个映射关系非常复杂，传统的基于物理和数学的方法难以解决。随着实验数据的积累，基于进化分析的新思路，特别是人工智能方法的发展，预测精度近几年得到突破性提升。从 2016 年 RaptorX 首次将深度残差网络应用于残基接触图预测，到 2018 年 AlphaFold1 设计深达 200 多层的残

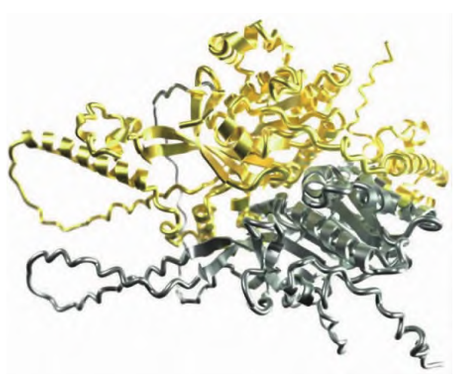

图 1 人工智能预测了两种蛋白质如何形成参与酵母 DNA 修复的复合体（图片来源：*Science* 官网）

差网络，再到 2020 年 AlphaFold2 通过端到端的神经网络设计，终于将蛋白质结构预测的精度提升到接近实验方法结构解析的精度，引起了整个生物学领域的巨大轰动。AlphaFold2 团队随后于 2021 年在 *Nature* 杂志连发两篇文章，公开了方法的实现细节和部分源代码，预测了整个人类蛋白质组的结构，并免费向公众开放（https://alphafold.ebi.ac.uk）。同年，华盛顿大学的 Baker 团队开发了 RoseTTAFold，同样利用深度学习技术准确、快速地预测蛋白质结构。

AlphaFold2 的突破是人工智能工程实践的巨大成功。从蛋白质结构预测角度来看，AlphaFold2 并没有引入新的思路。然而，AlphaFold2 成功实现了端到端的蛋白结构预测框架，直接通过 PDB 文件中原子的三维坐标来学习模型的参数。AlphaFold2 大量应用了深度学习的新技术和技巧，包括针对三维结构等变性质的点不变注意力模块、融合序列和结构信息的变换神经网络 Evoformer 模块、迭代优化和知识蒸馏策略等。AlphaFold2 开发团队包括多位深度学习工程师，他们使用了谷歌公司大量计算资源，在海量多序列比对信息上进行学习，对其复杂模型进行各种调试、优化，最后得到一套精确的预测模型。

AlphaFold2 的突破将带领结构生物学研究和药物开发进入新的时代。AlphaFold2 可以赋能结构生物学新的研究方向，比如加速 X 射线晶体衍射和冷冻电镜等实验技术解析蛋白质结构的速度、结合冷冻电镜单颗粒分析（SPA）和电子断层扫描（Cryo-ET）解析超大蛋白质复合物结构，甚至是其在细胞内的原位结构。AlphaFold2 将启发新的药物发现方法的出现。药物分子通过与靶标蛋白作用而发挥药效。根据 AlphaFold2 预测的目标蛋白结构，可以对该蛋白的成药性进行大致判断并进行药物分子设计。AlphaFold2 还将促进新功能蛋白质的设计，比如可以利用 AlphaFold2 针对新型冠状病毒设计更高效抗体，为解决新冠病毒大流行提供新型解决方案。

AlphaFold2 的突破甚至可能加速生命科学新研究范式的来临。如前所述，AlphaFold2 的突破是人工智能的巨大成功，是大数据的巨大成功，也是工业化科研的巨大成功。随着测序、影像、结构等新生物技术的开发和广泛应用，生命科学已经发展成为数据驱动的学科。生命系统的复杂性为人工智能提供了科学问题的沃土。这两者在工业化组织形式下，必将催生基于大数据和人工智能的新生物学研究时代。

当然，我们也要看到，虽然 AlphaFold2 取得了巨大的突破，但蛋白质结构预测的问题并没有得到完全解决。目前，AlphaFold2 对于序列较长的蛋白质、不存在同源序列或同源序列较少的孤儿蛋白预测效果明显变差；对于蛋白质无序区域，蛋白质结合位点，特别是大的蛋白质复合体结构预测上仍然有较大挑战；对于造成蛋白质结构变化的突变，AlphaFold2 也不能很好地预测。这些问题期待在未来，随着结构生物学与人工智能的进一步发展及更深度交叉融合下取得新的进展。但有鉴于此，在现阶段我们不宜过分夸大基于 AlphaFold2 的结构预测在结构生物学研究、药物开发中的作用。

华人科学家在蛋白质结构预测领域取得了诸多里程碑式的成果。如在早期 CASP 比赛中一直处于前列的 I-TASSER、RaptorX、trRosetta 等方法的领导者和核心成员都是海外华人。随着中国经济的高速发展和科研投入的不断加大，中国科学家在蛋白质结构解析和人工智能研究领域已经跻身世界前列，这为进一步发展蛋白质结构预测及相关研究奠定了基础。期待在不久的未来，中国科学家在蛋白质结构预测、冷冻电镜结构解析、新功能蛋白质设计、药物设计等领域开发出更多原创性的方法，获得更多具有突破性的成果。

2　解锁古老泥土 DNA 宝库

最近，科学家们从洞穴地面的土壤中解锁了一个更大的古代 DNA 宝库。研究人员使用这种“泥土 DNA”来重建世界各地穴居人的身份。在西班牙的 Estatuas 洞穴，核 DNA 揭示了 8 万至 11.3 万年前生活在那里的人类的遗传特征和性别，并表明尼安德特人的一个谱系在 10 万年前结束的冰川期之后取代了其他几个谱系。在美国佐治亚州 Satsurblia 洞穴有 2.5 万年历史的土壤中，科学家们发现了来自以前未知的尼安德特人系的女性人类基因组，以及野牛和现已灭绝的狼的遗传痕迹。通过将墨西哥奇基维特洞穴中 1.2 万年前的黑熊 DNA 与现代熊 DNA 进行比较，科学家们发现，在最后一个冰河时代之后，洞中黑熊的后代向北迁徙至阿拉斯加。