科学通报 2025 年 第 70 卷 第 10 期: 1421 ~ 1427
解读“2024年度诺贝尔自然科学奖”专题 观 点
《中国科学》杂志社 SCIENCE CHINA PRESS

人工智能与蛋白质科学的融合: 2024年诺贝尔化学奖背后的蛋白质结构预测与设计革命

刘沈徽[1,2], 刘志杰[1,2*]

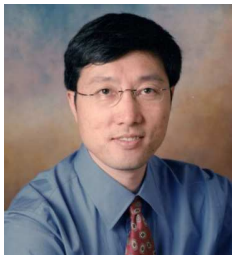
1. 上海科技大学iHuman研究所, 上海 201210
2. 上海科技大学生命科学与技术学院, 上海 200031
* 联系人, E-mail: liuzhj@shanghaitech.edu.cn

近年来, 人工智能技术的突破性进展使其以前所未有的速度在改变着人类社会的方方面面, 在科学研究上更是引发了颠覆性的进展. 2024年诺贝尔物理学奖和化学奖都不约而同地颁发给了人工智能相关的研究. 其中, 诺贝尔化学奖授予了谷歌DeepMind研究人员Demis Hassabis和John M. Jumper, 以表彰他们在蛋白质结构预测领域的杰出贡献. 另一半诺贝尔化学奖项则由华盛顿大学的David Baker教授获得, 表彰他引领了蛋白质设计领域的革命性进展. Baker团队开发的Rosetta软件包不但可以预测蛋白质结构, 也能够按照功能从零开始设计全新的, 甚至自然界不存在的蛋白质. 这意味着人类可以创造出具有各种功能的全新的生物大分子, 为制药、疫苗、传感器等多个领域提供了无限的可能.

蛋白质作为生命的基本组成单元, 是生命活动的直接执行者, 其结构与功能由20种氨基酸残组成的一级序列所决定. 组成蛋白质分子的肽链在细胞内由核糖体合成的同时便折叠成三维结构, 进而成为具有特定功能的生物大分子. 由于主链存在多个单键, 肽链可以围绕这些单键自由旋转, 因此在排除氨基酸残基侧链之间空间位阻的情况下, 蛋白质分子的折叠方式几乎是无穷大的. 自20世纪70年代以来, 科学家一直致力于通过蛋白质的一级序列预测其三维结构, 但未能取得较大突破. 2021年, 基于人工智能深度学习算法的蛋白质三维结构预测方法AlphaFold2(AF2)[1]的出现震惊了世界, 其展现的蛋白结构预测精度横扫了其他所有的预测手段.

1 蛋白质科学研究的历史进程

蛋白质科学研究一直是诺贝尔化学奖青睐的对象, 特别是在结构生物学领域, 历年来多项重要发现均受到了诺贝尔奖的垂青. 自20世纪50年代以来, X射线晶体学的迅速发展为科学家揭示蛋白质三维结构提供了强有力的工具. John Kendrew和Max Perutz成功解析了血红蛋白和肌红蛋白的晶体结构, 成为首批揭示蛋白质分子三维结构的科学家. 为表彰其卓越贡献, 他们于1962年共同获得诺贝尔化学奖[2~4]. 他们的成就不仅推动了结构生物学的发展, 也为后续的生物医药研究奠定了坚实基础.



刘志杰 上海科技大学讲席教授, 大道书院院长、iHuman研究所执行所长. 1994年获中国科学院生物物理研究所博士学位, 研究方向为GPCR结构与功能研究和基于结构的药物设计. 现任中国生物物理学会副理事长.

随着研究的深入, 核磁共振波谱学(NMR)的进步使得研究人员能够在溶液中观察蛋白质的结构和动态特性. 这项技术的优势在于能够在接近生理条件下分析蛋白质分子, 从而提供更为真实的结构信息[5]. 核磁共振波谱学在结构生物学中的应用也获得了2002年诺贝尔化学奖. 此外, 2017年的诺贝尔奖则颁发给了研究冷冻电子显微镜技术的先驱们[6], 冷冻电子显微镜技术在设备及软件方面的突破性进展使蛋白质的结构研究呈现了井喷式增长, 这项技术极大简化了生物大分子的成像过程, 成为当前结构生物学领域最受关注的技术之一.

2 蛋白质结构预测的历史发展

1972年, 美国生物化学家Christian Anfinsen因发现蛋白质的氨基酸序列决定了其肽链的折叠方式及三维结构而荣获诺贝尔化学奖[7], 自此科学家都有这样一个设想: 是否仅仅通过氨基酸的序列组成就可以预测任何蛋白质的三维结构.

准确预测蛋白质结构需要对于蛋白质的折叠过程有深

引用格式: 刘沈徽, 刘志杰. 人工智能与蛋白质科学的融合: 2024年诺贝尔化学奖背后的蛋白质结构预测与设计革命. 科学通报, 2025, 70: 1421–1427
Liu S, Liu Z-J. The integration of artificial intelligence and protein science: the protein structure prediction and protein design revolution behind the 2024 Nobel Prize in Chemistry (in Chinese). Chin Sci Bull, 2025, 70: 1421–1427, doi: 10.1360/TB-2024-1178

入的理解以及对于蛋白质序列与天然构象之间的关系具有很深入的认知. 目前, 蛋白质结构预测的方法总体可以分为两大类, 一类是基于模板的蛋白质结构预测, 又称为template-based modeling, 另外一类则无需全局模板结构即可生成结构模型, 简称template-free modeling方法. 最早发展出的蛋白质结构预测方法是同源建模法[8], 同源建模基于序列相近的蛋白质其结构也相似的假设, 因此可以通过其同源蛋白的三维结构来推断目标蛋白的结构, 根据选取的靶蛋白与同源模板的比对结果, 可以使用结构建模工具例如MODELLER[9]构建其结构模型, 至今这种方法仍然被广泛使用, 但值得注意的是当序列相似性太低, 比如低于30%时, 该方法的准确性会迅速下降. 随后threading方法被提出, 该方法假设: 与蛋白质的氨基酸序列相比, 蛋白质的三维结构中的某些局部结构特征更加保守, 这些局部结构特征可以用来识别和对齐序列同源性较低但结构相似的蛋白质[10,11]. Bowie等人[11]将蛋白质中每个氨基酸的位置分为不同的环境类, 根据其所处的表面暴露程度和二级结构类型进行分类. 例如, 某个氨基酸可能属于“暴露的β折叠区”或者“埋藏的螺旋区”, 这些信息比氨基酸的序列信息更加保守, 这一方法突破了传统序列比对的局限性, 特别是在序列相似性较低的情况下, 能够有效识别结构上相似但序列上不同的蛋白质.

以上提及的方法属于基于模板的建模方法, 针对无模板建模的方法, 目前的认知是蛋白质的天然结构应该处于低能量态, 局部区域的残基与残基之间完美契合且无相互冲突. 因此该类方法主要基于第一性原理, 尝试使用能量函数来寻找蛋白质结构的最低能量构象或直接模拟折叠过程来实现预测[12,13]. 早期的这类方法主要用于优化X射线衍射方法解析的蛋白晶体结构, 通过细化键长、键角、二面角、范德华力等因素从而提高结构的物理特性[14]. 此外, Shaw等人[15]尝试通过分子动力学直接模拟蛋白质折叠过程, 然而由于能量函数不够准确以及高昂的计算成本, 该方法并没有被广泛使用. 此外, 还有碎片拼接方法, 例如Baker团队[16,17]于1997年开发了Rosetta建模软件, Rosetta通过模拟退火蒙特卡罗方法, 插入高评分的碎片来优化蛋白质结构. 然而由于碎片拼接法的计算量巨大, 基于梯度下降的快速折叠法被提出, 利用梯度下降(gradient descent)技术, 预测过程可以快速收敛到蛋白质结构的最低能量状态[18]. 由于蛋白质的构象系综及能量分布非常复杂, 为了防止梯度下降法陷入局部能量最小值, 近年来通过深度学习预测的几何空间约束(如氨基酸之间的距离)被用来平滑能量分布, 使得梯度下降方法能够更加准确预测蛋白质结构[18,19]. 例如, AlphaFold和trRosetta这两种方法就利用了深度学习技术来预测蛋白质结构, 并取得了突破性进展. AlphaFold在CASP13竞赛中首次展示了其优异的性能, 甚至在快速的梯度下降模拟下也能达到或接近传统碎片拼接方法的准确性.

自2010年以来, 蛋白质结构预测领域已经普遍融入了深度学习技术, 从而改变了其研究范式. 这种新方法通过深度神经网络直接从大规模数据集中学习, 不再依赖预设的分布假设. 这使得从蛋白质序列中预测其二级结构、残基间距离乃至整个三维结构变得更为精确. 早期的算法建模应用包括CMAPpro[20], 该方法通过深度学习改进接触预测; DL-Pro利用几何特征评估蛋白质模型质量[21]; DNSS使用序列信息通过深度学习预测蛋白质的二级结构[22]. 其中, RaptorX-Contact通过深度残差网络学习和细化残基间接触图, 相比传统方法展现出更高的性能[23]. 蛋白质语言模型是此范式的另一应用, 如Transformer等深度网络[24]被用来揭示蛋白质序列的隐含规律. 这些模型在大规模的蛋白质序列数据集上训练后, 能够有效预测蛋白质结构和功能. 深度学习技术还被用于从残基间距离图中识别常见模式, 并利用这些模式构建序列-结构对齐[25]. AlphaFold2等工具的最新进展应用了端到端的深度学习策略[1], 直接从序列预测蛋白质结构, 通过从结构到序列的直接反向传播, 使得学习过程更加直接和高效.

3 AlphaFold2的技术突破

AlphaFold2是深度学习领域中的一项重要技术突破, 它成功解决了蛋白质结构预测这一长期存在的科学难题, 在2020年的CASP14竞赛中[26], 其预测结果与实验结果的接近程度达到了空前的高度, 以前所未有的准确度横扫了所有的参赛队伍. 标志着该技术在蛋白质结构预测领域实现了颠覆性的突破. AlphaFold2的成功主要得益于几个重要的技术突破: (1) 创新的神经网络架构. AlphaFold2使用了一种全新的神经网络架构, 结合了卷积神经网络(CNN)和注意力机制(attention mechanism), 有效捕捉到蛋白质序列的远程相互作用, 并将其转化为三维空间中的结构关系. 这种结合使预测模型能够处理蛋白质折叠过程中复杂的分子间作用力. (2) 多序列比对(MSA)和模板建模整合: AlphaFold2引入了多序列比对(MSA)和模板建模(template modeling)的信息, 通过深度学习从大量蛋白质序列及其同源信息中学习相互作用模式, 从而极大地提升了预测精度. MSA使预测模型能够理解保守的进化信息, 这对于准确预测蛋白质结构至关重要. (3) 引入结构重建模块: AlphaFold2在预测过程中引入了结构重建模块, 使模型在迭代更新中逐步逼近真实结构. 这种“结构重建与调整”的方法极大提高了预测质量. (4) 端到端的训练方式: 与传统方法相比, AlphaFold2采用了端到端的训练方式, 通过整合输入的蛋白质序列及多种辅助信息, 让模型直接输出蛋白质的三维坐标. 这种方法减少了对中间步骤的依赖, 使整个预测过程更为自动化和高效.

蛋白质结构预测被称为“生物学中的最大挑战之一”, AlphaFold2的成功使科学家能够快速、准确地了解蛋白质的三维结构, 从而为理解疾病机制、创新药物设计以及新材料合成等领域带来了巨大的潜力.

本课题组在AlphaFold2开源后迅速评估了其对G蛋白偶

联受体(GPCR)结构预测的准确性，GPCR是人体中最为庞大和具有结构多样性的膜蛋白家族，可被划分为以下几类家族：Class A、Class B1、Class B2、Class C、Class D1、Class F (图1(a))，以及最近被解析的苦味受体Class T家族[27,28]. 为检验AlphaFold2预测的结构与实验测定结构之间的差异，我们对截至2022年3月31日存入Protein Data Bank (PDB)[29]中的GPCR结构数据进行了整理和选取，保留了高分辨率(晶体结构优于3.2 Å,冷冻电子显微镜结构优于3.8 Å)的结构用于对AlphaFold2的评估. 此外，为了更好地评测AlphaFold2预测结构的精确度，我们将GPCR蛋白结构中的loop区删除，仅保留疏水跨膜螺旋区域. 我们通过计算TM-score[30]和RMSD[31]呈现AlphaFold2预测的结构与PDB中实验数据的差别，发现AlphaFold2对于Class A和Class F GPCR亚家族的预测准确性较高，TM-score和RMSD两项指标均优于对其他GPCR亚家族的预测结果(图1(b), (c)). 对于实验数据较多的Class A亚家族受体，大部分AlphaFold2预测结构的TM-score均超过0.90，然而对于非单一结构域的其他亚家族，AlphaFold2的预测精度普遍偏低，原因在于两个结构域之间的相对运动会导致整体对齐的得分下降(图1(d)). 但当把多结构域拆分成单结构域分别比对时，又会获得较好的表现(图1(e), (f))，说明AlphaFold2对于单结构域结构的预测有更高的准确性.

此外，为了确保评估的独立性，选取了我们课题组当时已解析但尚未发表的GPCR结构进行评估，其中包含了GPR12[32]、GPR139[33]、HCA2[34]以及Class T家族首个被解析的人源苦味受体TAS2R46[27]，这四个受体与预测结构的TM-score分别为0.90、0.88、0.94和0.93，RMSD分别为2.65、0.97、1.75和1.92，表明AlphaFold2对已发表和未发表结构的预测精度达到了几乎相同的水平.

由于GPCR存在apo、激活及拮抗等不同构象状态，我们好奇于AlphaFold2预测的结构与何种构象状态更相似. 我们从PDB数据库中选取了同时存在激活和拮抗两种状态的GPCR作为研究对象，评估结果显示AlphaFold2对于Class A GPCR亚家族预测的结构更偏向拮抗状态(图1(h), (i)). 另外，GPCR的配体发现一直是该领域的重大挑战之一，也是靶向GPCR药物发现的关键步骤. 配体对GPCR的调控作用主要是通过与在GPCR中的正构口袋(Orthosteric pocket)或别构口袋(Allosteric pocket)中的关键氨基酸残基的相互作用完成的. 因此，预测的蛋白质分子的主链和侧链结构精度都很重要. 据此，我们对AlphaFold2预测的ClassA家族GPCR的正构口袋残基进行了更为全面的评估，我们分别提取并计算了AlphaFold2预测结构与实验测定结构中组成正构口袋残基的Cα原子、Cα和Cβ原子、主链原子及组成正构口袋所有残基原子的RMSD，发现AlphaFold2对GPCR正构口袋残基Cα原子和主链骨架结构的预测均较为精准，然而在引入侧链原子后，也即全原子预测时的准确度明显降低(图1(j))，说明对于蛋白质分子残基侧链构象的预测，AlphaFold2还有提升的空间.

4 蛋白质设计的飞速发展

如果说AlphaFold2是蛋白质结构预测的革命者，那么David Baker就拥有蛋白质设计的上帝之手. 1998年前后，David Baker团队开始在蛋白质折叠领域取得初步成果，其开发出的Rosetta软件基于物理学能量最小化原理对蛋白质构象进行稳定性评估，作为一款革命性的蛋白质结构预测和设计工具，其团队不断拓展Rosetta的应用. 2003年，David Baker团队创造出第一个人工设计的蛋白top7[35]，随后使用X射线晶体学方法测定并验证了该蛋白的结构，但此蛋白并不具备功能. 2008年，Baker发布了蛋白质预测和设计相关的Foldit小游戏，邀请全世界科学家参与蛋白质设计，并于2011年在玩家的帮助下仅用10天就破解了一种艾滋病逆转录酶的结构[36]，成功解决了科研工作者花费15年时间都未能完成的难题. 2008年，David Baker团队利用Rosetta完成了第一个完全由计算机设计的蛋白质酶Kemp eliminase[37]，这标志着他们能够从头设计具有特定生物催化活性的蛋白质分子. 随后十年间，David Baker团队不断完善蛋白质设计方法，他们设计的蛋白质分子在医药、生物材料等领域展现出多种应用潜力. 2019年，他们设计出自组装的纳米颗粒疫苗，能够诱导强效的中和抗体反应[38]. 2020年，面对新冠病毒COVID-19大流行，David Baker团队迅速响应，设计了新型的小蛋白与SARS-CoV-2刺突蛋白紧密结合，并阻止其与宿主细胞受体的结合，这种微型蛋白为新冠病毒的防治提供了非常好的启示[39]. 2021年，他们开源了蛋白质预测工具RosettaFold[40]后，David Baker更专注于利用深度学习的方法进行蛋白质设计. 2022年开发出新一代的蛋白质设计引擎ProteinMPNN[41]，利用图神经网络对氨基酸之间的关系进行建模，可以更准确地预测蛋白质的三维结构和相互作用，并根据蛋白质骨架生成最优的序列. 随后，David Baker团队在2023年根据扩散模型推出了RFDiffusion[42]这一强大工具，用于生成与真实蛋白类似，但从未在自然界出现过的全新蛋白质分子，这一突破极大提升了蛋白质设计效率，并具有非常高的通用性和准确性.

值得注意的是，在蛋白质设计领域并非只有David Baker团队在推动其发展. DeepMind团队[1]在AlphaFold2方面的工作，极大地推动了蛋白质结构预测的准确性，其背后的深度学习方法也为蛋白质设计提供了新的视角. 2020年，Huang等人[43]开发出了一种构建蛋白质骨架的新模型Ig-VAE，该模型具有扭转和距离感知能力，可用于直接生成蛋白质骨架全部原子三维坐标. 2022年，清华大学马剑竹团队[44]开发了DiffAb模型，用于针对特定抗原结构生成对应抗体，该模型具有强大的普适性，能够进行序列-结构协同设计、给定骨架结构的序列设计和抗体优化. 此外，中国科学技术大学刘海燕团队在以往基于能量计算的蛋白质设计算法SCUBA的基础上引入扩散模型，并且引入了对抗损失，从而既能避免设计过程中对天然结构的偏好，又能避免在物理上的不合理性，极大提高了蛋白质设计的成功率[45]. 近期，DeepMind公司宣布

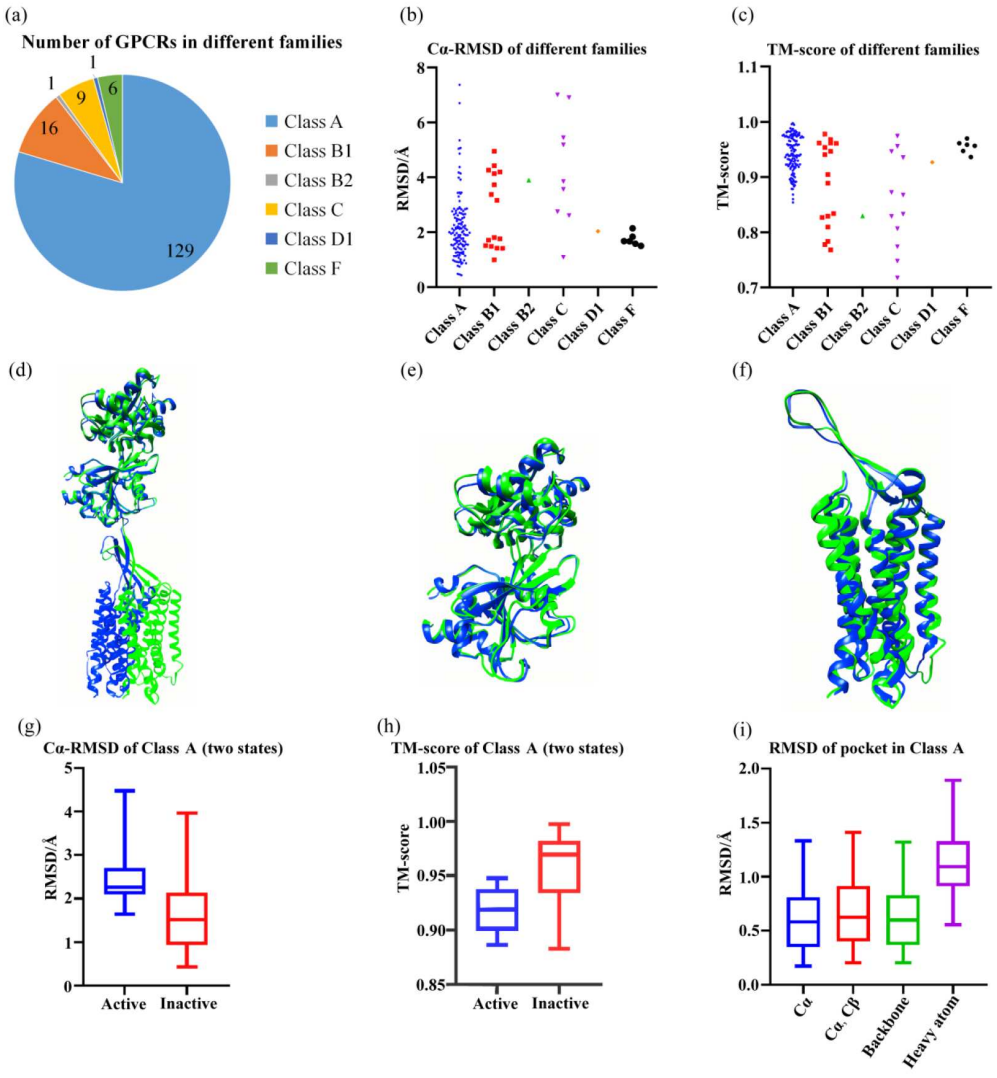


图 1 对AlphaFold2预测的GPCR结构与实验测定结构的比较和评估. (a) 参与评估的不同GPCR亚家族的受体数量分布; (b) 参与评估GPCR的AlphaFold2预测结构与实验测定结构的Cα-RMSD值; (c) 参与评估GPCR的AlphaFold2预测结构与实验测定结构的TM-Score值; (d) 人源GABA (B) receptor (PDB_I D: 7EB2)结构(绿色)与AlphaFold2预测结构(蓝色)的叠合; (e) 图1(d)中两种结构的胞外结构域的叠合; (f) 图1(d)中两种结构的跨膜结构域的叠合; (g) 参与评估GPCR的AlphaFold2预测结构与实验测定的激活态结构的Cα-RMSD值(蓝色), 及与实验测定的拮抗态结构的Cα-RMSD值(红色); (h) 参与评估GPCR的AlphaFold2预测结构与实验测定的激活态结构的TM-score值(蓝色), 及与实验测定的拮抗态结构的TM-score值(红色); (i) AlphaFold2预测结构与实验测定结构正构口袋残基的Cα-RMSD、Cα-Cβ-RMSD、主链-RMSD及组成正构口袋所有残基原子的RMSD值

Figure 1 Comparison and evaluation of AlphaFold2 predicted GPCR structures against experimentally determined structures. (a) Distribution of the number of receptors across different GPCR subfamilies which are involved in the evaluation. (b) The Cα-RMSD values between AlphaFold2 predicted structures and experimentally determined structures of GPCRs. (c) The TM-Score values between AlphaFold2 predicted structures and experimentally determined structures of GPCRs. (d) The superposition between human GABA(B) receptor structure (PDB_ID: 7EB2) (green) and AlphaFold2 predicted structure (blue). (e) The superposition between the extracellular domains of the two structures shown in Figure 1(d). (f) The superposition between the transmembrane domains of the two structures shown in Figure 1(d). (g) Cα-RMSD values between AlphaFold2 predicted structures and experimentally determined active-state structures of GPCRs (blue), and with experimentally determined antagonist-state structures (red). (h) TM-score values between AlphaFold2 predicted structures and experimentally determined active-state structures of GPCRs (blue), and with experimentally determined antagonist-state structures (red). (i) Cα-RMSD, Cα-Cβ-RMSD, backbone-RMSD, and RMSD of all atomic residues composing the orthosteric pocket between AlphaFold2 predicted structures and experimentally determined structures

开发了一种新的蛋白质设计算法AlphaProteo[46], 该算法专门用于生成高亲和力的binder蛋白, 并且在七个目标蛋白质binder设计的湿实验验证中实现了9%~88%的成功率, 为许多相关研究应用提供了更高效的解决方案.

5 结语

AlphaFold2的诞生是蛋白质结构预测领域的一次革命性突破，它展现了人工智能在解决生物学问题上的巨大潜力，通过与蛋白质设计的结合，使得我们能够不再仅仅依赖自然界中已有的蛋白质，而是有望通过计算与实验相结合，创造出功能更加多样和优化的蛋白质，从而在生物医学、药物开发、材料科学等多个领域取得更大进展.

尽管AlphaFold2在蛋白质结构预测领域取得了令人瞩目的成果，但它的出现也揭示了一些潜在的挑战和局限性. 首先，即使新一代AlphaFold3[47]的出现使得预测蛋白质的三维结构变得更加精准，但对于蛋白配体结合，蛋白质动力学，以及构象多样蛋白分子的准确预测能力仍有待提升. 此外，AlphaFold2代表了其模拟端到端生物过程的成功，在某种程度上忽略了生物学本质的复杂性，我们仍然需要从深层机理上探寻蛋白质折叠的奥秘. 通过本年度诺奖的颁布，我们看到了人工智能与生物科学深度融合的光明前景. 未来的研究将朝着更智能、更精准的方向发展，期待人类在破解生命奥秘、疾病治疗、健康管理以及推动合成生物学的道路上取得更多里程碑式的成就.

推荐阅读文献

1 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
2 Perutz M F, Rossmann M G, Cullis A F, et al. Structure of Hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. Resolution, obtained by X-ray analysis. Nature, 1960, 185: 416–422
3 Kendrew J C, Dickerson R E, Strandberg B E, et al. Structure of Myoglobin: a three-dimensional Fourier synthesis at 2 Å. Resolution. Nature, 1960, 185: 422–427
4 Kendrew J C, Bodo G, Dintzis H M, et al. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature, 1958, 181: 662–666
5 Tabet J C, Rebuffat S. Prix Nobel de chimie 2002. Med Sci, 2003, 19: 865–872
6 Cressey D, Callaway E. Cryo-electron microscopy wins chemistry Nobel. Nature, 2017, 550: 167
7 Richards F M. The 1972 Nobel Prize for Chemistry. Science, 1972, 178: 492–493
8 Browne W J, North A C T, Phillips D C, et al. A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol, 1969, 42: 65–86
9 Šali A, Potterton L, Yuan F, et al. Evaluation of comparative protein modeling by MODELLER. Proteins, 1995, 23: 318–326
10 Xu J, Li M, Kim D, et al. Raptor: optimal protein threading by linear programming. J Bioinform Comput Biol, 2003, 01: 95–117
11 Bowie J U, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science, 1991, 253: 164–170
12 Dobson C M, Šali A, Karplus M. Protein folding: a perspective from theory and experiment. Angew Chem Int Ed, 1998, 37: 868–893
13 Hamelryck T, Kent J T, Krogh A, et al. Sampling realistic protein conformations using local structural bias. PLoS Comput Biol, 2006, 2: e131
14 Levitt M, Lifson S. Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol, 1969, 46: 269–279
15 Lindorff-Larsen K, Piana S, Dror R O, et al. How fast-folding proteins fold. Science, 2011, 334: 517–520
16 Simons K T, Kooperberg C, Huang E, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol, 1997, 268: 209–225
17 Rohl C A, Strauss C E, Misura K M, et al. Protein structure prediction using Rosetta. Methods Enzymol, 2004. 383: 66–93
18 Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA, 2020, 117: 1496–1503
19 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706–710
20 Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics, 2012, 28: 2449–2457
21 Nguyen S P, Shang Y, Xu D. DL-PRO: a novel deep learning method for protein model quality assessment. In: 2014 International Joint Conference on Neural Networks (IJCNN). New York: IEEE, 2014. 2071–2078, doi: 10.1109/IJCNN.2014.6889891
22 Spencer M, Eickholt J, Jianlin C. A deep learning network approach to *ab initio* protein secondary structure prediction. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12. New York: IEEE, 2015. 103–112
23 Wang S, Sun S, Li Z, et al. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol, 2017, 13: e1005324
24 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379: 1123–1130
25 Kong L, Ju F, Zheng W, et al. ProALIGN: directly learning alignments for protein structure prediction via exploiting context-specific alignment motifs. J Comput Biol, 2022, 29: 92–105

26 Kryshtafovych A, Schwede T, Topf M, et al. Critical assessment of methods of protein structure prediction (CASP)—Round XV. Proteins, 2023, 91: 1539–1549
27 Xu W, Wu L, Liu S, et al. Structural basis for strychnine activation of human bitter taste receptor TAS2R46. Science, 2022, 377: 1298–1304
28 Hu X, Ao W, Gao M, et al. Bitter taste TAS2R14 activation by intracellular tastants and cholesterol. Nature, 2024, 631: 459–466
29 Berman H M. The Protein Data Bank. Nucleic Acids Res, 2000, 28: 235–242
30 Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins, 2004, 57: 702–710
31 Kufareva I, Abagyan R. Methods of protein structure comparison. Methods Mol Biol, 2012, 857: 231–257
32 Li H, Zhang J, Yu Y, et al. Structural insight into the constitutive activity of human orphan receptor GPR12. Sci Bull, 2023, 68: 95–104
33 Zhou Y, Daver H, Trapkov B, et al. Molecular insights into ligand recognition and G protein coupling of the neuromodulatory orphan receptor GPR139. Cell Res, 2022, 32: 210–213
34 Yang Y, Kang H J, Gao R, et al. Structural insights into the human niacin receptor HCA2-Gi signalling complex. Nat Commun, 2023, 14: 1692
35 Kuhlman B, Dantas G, Ireton G C, et al. Design of a novel globular protein fold with atomic-level accuracy. Science, 2003, 302: 1364–1368
36 Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. Nature, 2010, 466: 756–760
37 Röthlisberger D, Khersonsky O, Wollacott A M, et al. Kemp elimination catalysts by computational enzyme design. Nature, 2008, 453: 190–195
38 Marcandalli J, Fiala B, Ols S, et al. Induction of potent neutralizing antibody responses by a designed protein nanoparticle vaccine for respiratory syncytial virus. Cell, 2019, 176: 1420–1431.e17
39 Cao L, Goreshnik I, Coventry B, et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science, 2020, 370: 426–431
40 Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373: 871–876
41 Dauparas J, Anishchenko I, Bennett N, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science, 2022, 378: 49–56
42 Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion. Nature, 2023, 620: 1089–1100
43 Eguchi R R, Choe C A, Huang P S, et al. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. PLoS Comput Biol, 2022, 18: e1010271
44 Luo S T, Su Y F, Peng X G, et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022, Article No. 709. 9754–9767, https://dl.acm.org/doi/10.5555/3600270.3600979
45 SCUBA-D: a freshly trained diffusion model generates high-quality protein structures. Nat Methods, 2024, 21: 1990–1991
46 Zambaldi V, La D, Chu A E, et al. De novo design of high-affinity protein binders with AlphaProteo. 2024, arXiv: 2409.08022
47 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493–500

Summary for “人工智能与蛋白质科学的融合: 2024年诺贝尔化学奖背后的蛋白质结构预测与设计革命”

The integration of artificial intelligence and protein science: the protein structure prediction and protein design revolution behind the 2024 Nobel Prize in Chemistry

Shenhui Liu[1,2] & Zhi-Jie Liu[1,2*]

[1] *iHuman Institute, ShanghaiTech University, Shanghai 201210, China*
[2] *School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China*
* Corresponding author, E-mail: liuzhj@shanghaitech.edu.cn

In recent years, groundbreaking advances in artificial intelligence (AI) have transformed various aspects of human society at an unprecedented pace, driving revolutionary progress in scientific research. The 2024 Nobel Prizes in Physics and Chemistry were both awarded for AI-related research, highlighting the immense impact of AI in reshaping scientific fields. Notably, the Nobel Prize in Chemistry was awarded to Demis Hassabis and John M. Jumper of Google DeepMind for their remarkable contributions to protein structure prediction, while the other half of the award went to Professor David Baker from the University of Washington, recognizing his leadership in the revolutionary progress of protein design. The Rosetta software suite, developed by Baker’s team, not only predicts protein structures but also designs entirely new proteins, including those that do not exist in nature, opening up infinite possibilities for pharmaceutical development, vaccine creation, and sensor technology.

Proteins, as the fundamental building blocks of life, are the direct executors of biological activities. Their structure and function are determined by the amino acid sequence encoded in their primary structure. The peptide chains forming proteins fold into three-dimensional structures as they are synthesized by ribosomes, ultimately assuming specific functional roles. However, predicting the three-dimensional structure of proteins from their amino acid sequences has proven to be an extraordinarily complex challenge due to the vast number of possible folding configurations. Despite decades of efforts, significant breakthroughs only occurred in the past few years with the advent of AI-driven methods, particularly AlphaFold2 (AF2), a deep learning algorithm developed by DeepMind that has revolutionized protein structure prediction with unprecedented accuracy.

This article traces the historical development of protein science, emphasizing key milestones, from X-ray crystallography and NMR spectroscopy to cryo-electron microscopy, all of which have advanced our understanding of protein structure. The pursuit of predicting protein structure from amino acid sequences has been a central goal in biochemistry since the 1970s, yet it remained elusive until AI breakthroughs. The early methods, such as homology modeling and threading, relied on comparing known structures to predict the unknown. In contrast, template-free modeling, which is grounded in first principles, strives to predict the structure by minimizing energy functions, but computational limitations hinder its progress.

The real breakthrough came with the introduction of deep learning techniques in the last decade. AI algorithms, including AlphaFold2, revolutionized the field by directly learning from large-scale datasets of protein sequences and their associated structures. By combining multi-sequence alignments (MSA) and template modeling, AlphaFold2 effectively captures evolutionary information and molecular interactions to predict protein structures with remarkable accuracy. Furthermore, the use of deep neural networks and novel training strategies in AlphaFold2 significantly improved the efficiency and reliability of predictions, making it a game-changer for structural biology.

This article discusses the technical innovations behind AlphaFold2, including its neural network architecture, integration of evolutionary data through MSA, and the novel use of a structure refinement module. These advancements allowed AlphaFold2 to predict protein structures with atomic-level precision, transforming the field and demonstrating its potential in drug discovery and disease research. The success of AlphaFold2 has implications beyond basic biology, offering a new paradigm for understanding protein functions, accelerating the development of therapeutic agents, and enabling the design of synthetic proteins with specific properties.

The article also evaluates AlphaFold2’s performance in predicting structures of G protein-coupled receptors (GPCRs), highlighting its accuracy in predicting single-domain structures and its potential for advancing drug discovery, especially in ligand-receptor interactions. However, challenges remain, particularly in predicting protein side-chain conformations, where improvements are still needed. Despite these challenges, AlphaFold2 represents a monumental step forward in the field, providing an unprecedented tool for structural biologists and molecular biologists alike.

In conclusion, the fusion of AI and protein science has reshaped the landscape of biomedical research, as evidenced by the 2024 Nobel Prize in Chemistry. As AI continues to evolve, its impact on protein science will expand, unlocking new avenues for therapeutic development and revolutionizing the way we approach disease treatment and drug design.

artificial intelligence (AI), protein structure prediction, AlphaFold2, Nobel Prize in Chemistry

doi: 10.1360/TB-2024-1178