PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024, 51(12): 3073~3083
www.pibb.ac.cn

AlphaFold结构预测的重大突破及其对蛋白质研究的影响与挑战*

宫维斌**
（中国科学院生物物理研究所，生物大分子重点实验室，北京 100101）

摘要 近年来，基于深度学习的方法研究在蛋白质结构预测领域实现了重大突破。AlphaFold 2（AF2）于2021年开源发布，实现了蛋白质暨蛋白质复合物三维结构的高精度预测，使得研究人员能够快速获取可靠的三维结构信息，显著加速了蛋白质结构与功能研究的进展。2024年发布的AlphaFold 3（AF3）更进一步，能对蛋白质-核酸、蛋白质-小分子等生物复合物的三维结构进行精准预测。AF3采用改进的算法与更高效的模型，大幅提升了预测准确度，特别是在抗原-抗体复合物、蛋白质-小分子复合物等方面展现出卓越的性能。AlphaFold的成功不仅为结构生物学带来了革命性进展，还在药物研发、蛋白质设计、分子功能机制研究等领域展示了巨大的应用潜力，推动了生物医学研究的革新。本文将回顾AlphaFold及相关蛋白质结构预测方法的研发历史，概述其关键技术和当下应用，并结合其局限性，展望未来的研究方向和应用。

关键词 AlphaFold，深度学习，大数据，蛋白质结构预测，生物分子复合物结构预测，蛋白质设计
中图分类号 Q518 **DOI**：10.16476/j.pibb.2024.0374

蛋白质是碳基生命活动的主要执行者。蛋白质由20种天然氨基酸通过肽键首尾串连形成，氨基酸的排列顺序构成其一级序列。根据生物学中心法则，DNA通过三联体密码编码了蛋白质的一级序列，通过RNA聚合酶转录为mRNA，再经核糖体翻译为蛋白质。蛋白质需要自发或在分子伴侣辅助下折叠成为特定的三维结构，以发挥其功能。那么蛋白质的一级序列和三维结构之间存在什么关系呢？一般认为，蛋白质一级序列决定了其稳定状态的三维结构，该三维结构往往处于能量最低状态，这就是所谓的“序列决定结构”理论[1-2]。

解析蛋白质及其复合物的三维结构，是理解蛋白质酶活性和相互作用等功能机制的一个必要条件。如果能完全了解生物分子之间如何相互作用，可以说我们就在相当程度上掌握了生命最深层的秘密，因为每一个生物现象，包括我们如何感知世界，最终都源于生物分子之间的相互作用。然而蛋白质三维结构的解析是一个需要高费用和高人力的工作，一个样品的获得和结构解析往往需要耗费数月乃至数年的时间。因而，根据蛋白质的一级序列快速获得准确的三维结构是一个重大科学问题。

在过去几年中，蛋白质三维结构预测的深度学习方法AlphaFold2（AF2）取得了巨大的成功，在蛋白质结构与功能研究领域，乃至生物学研究领域中掀起了一场革命。第三代AlphaFold（AF3）的研究论文已于2024年5月在《自然》（*Nature*）期刊上发表[3]，正在进一步推动这场革命。AF2和AF3不仅大大加速了重组蛋白质结构解析以及原位结构解析的过程，甚至在一定程度上通过提供高精度三维结构模型，使得蛋白质功能研究不再需要进行耗时耗力的实验结构解析。此外，基于AlphaFold等方法开发的蛋白质设计方法和流程将设计成功率提高了多个数量级[4]，使其应用快速普及和增长。本文试图对AlphaFold的研发历史作一次回顾，并总结、展望其当下和将来的发展和应用。

* 国家自然科学基金（32371273）资助项目。
** 通讯联系人。
Tel: 86-10-64888000, E-mail: gongweibin@ibp.ac.cn
收稿日期：2024-08-16，接受日期：2024-10-15

1　AF2的过去和现在

1.1　CASP竞赛推动AF2的研发

在早期，人们曾使用分子模拟方法模拟蛋白质的折叠过程得到预测的三级结构。然而该方法的效率和局限性较大，只能对几十个残基的短肽进行模拟预测。Moult等[5]自1994年开始组织了面向全球科研学者的蛋白质结构预测竞赛，即CASP（Critical Assessment of Techniques for Protein Structure Prediction），希望集全世界科研工作者们的努力，解决上述问题。CASP竞赛主要分为自由建模（free modelling，FM）和同源建模（template-based modelling，TBM）两类。后者以已解析的同源序列蛋白质结构为模板，对所发布蛋白质序列进行结构建模，而前者的靶标蛋白质缺乏序列同源的结构模板，需要进行从头结构建模，最具挑战性，对于解析一级序列与三级结构之间的关系也最为重要。

在TBM竞赛中，密歇根大学ZHANG Yang团队[6]开发的I-TASSER（Iterative Threading ASSEmbly Refinement）方法表现优异，在CASP7~CASP15九次竞赛中摘得服务器组的桂冠。华盛顿大学David BAKER团队开发的Rosetta软件参加了第三届CASP的FM竞赛（CASP3），被用于21个包含150个残基以下的蛋白质结构的从头预测，其中14个预测结构与实验结构的主链原子均方根差（root mean squared deviation，RMSD）约3.0~6.0 Å，取得当时最好成绩。图1a展示了其中一个较为成功的预测案例。Rosetta首先将蛋白质一级序列分割成3~9个残基片段，基于模板搜索和能量优化构建其局部构象，再将各片段结构进行组装，获得完整的三级结构[7]。结合循环突变和能量函数优化，Rosetta也被高效地用于蛋白质设计研究中。

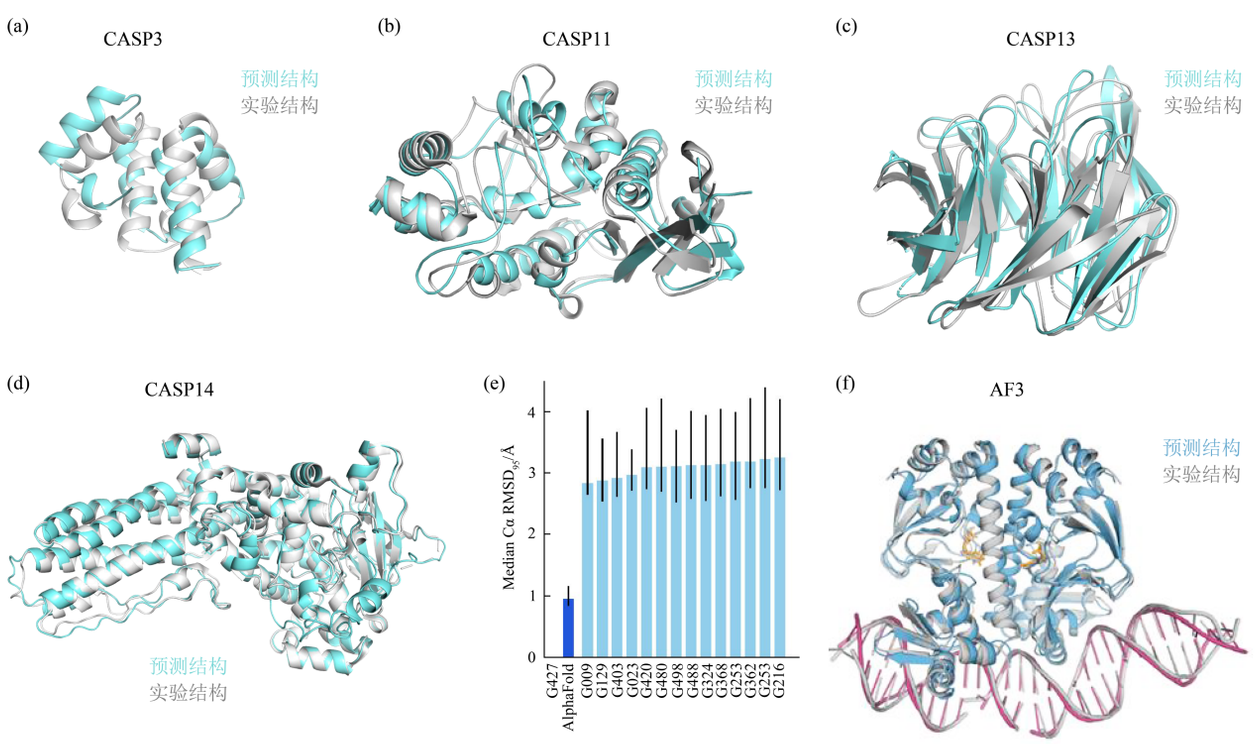


Fig. 1　The best results of protein structure prediction at different periods

图1　不同时期的蛋白质结构预测最佳结果

（a）CASP3竞赛T0056靶标，Rosetta预测结构模型含77个残基，与实验结构之间的主链RMSD约4 Å；（b）CASP11竞赛T0806靶标，Rosetta预测结构模型含223个残基，与实验结构之间的主链Cα原子RMSD约3 Å；（c）CASP13竞赛T0954靶标，AlphaFold预测结构模型含342个残基，与实验结构之间的主链Cα原子RMSD约2 Å；（d）CASP14竞赛T1053靶标，AF2预测结构模型含580个残基，与实验结构之间的主链RMSD约0.9 Å；（e）AF2与CASP14竞赛中其他方法预测结果的对比[8]，其中RMSD值越小越好；（f）AF3预测的Clr-cAMP-DNA结构模型[3]，其中含有包含同源二聚Clr蛋白（244个残基），2个cAMP小分子配体和DNA双链。

然而，FM预测的蛋白质大小一直存在较大的限制，精确度也较为有限。直到2014年CASP11竞赛中，Ovchinnikov等[9]通过多序列比对（multiple sequence alignment，MSA）获得序列中氨基酸残基的共进化信息，将其使用在结构预测中，预测的准确度大大提高，其预测获得的超过200个残基蛋白质的结构与实验结构之间的RMSD小于3 Å（图1b）。CASP12竞赛中，结合氨基酸残基的共进化信息，RaptorX-Contact方法第一次将深度学习引入蛋白质结构预测中，取得了重大进展[10]。在2018年CASP13竞赛中，基于MSA中的共进化信息和深度神经网络的使用，AlphaFold（AF1）开始崭露头角（图1c），显著地超越其他方法。深度学习方法trRosetta受RaptorX-Contact和AF1启发，在模型训练中将残基间的取向和距离测量转换为约束，结合物理方法中的能量优化，进行从头建模，获得了比AF1更高的模板建模分数（TM-score），同时却使用了较少的计算资源[11]。2020年的CASP14竞赛中，AF2更新了AF1的各个组件，以巨大的优势遥遥领先于所有其他的方法，并且对于无同源序列获得结构解析的困难蛋白质序列的结构预测也达到了CASP竞赛的最终准确度目标（图1d，e）[12]。AlphaFold由此开启了一个变革性的新时代。

AF2的相关论文于2021年在著名期刊《自然》（*Nature*）上发表[8]，并于开源网站github上公布了源代码和模型参数。AF2使用了一种称为Evoformer的transformer算法来促进两种关键数据（MSA和结构模板）之间的多次循环相互作用。其中，模板中残基对之间的空间邻近性增强了对MSA的推断分析，而从MSA提取的信息又反过来帮助更新残基对之间的空间邻近性。接着，AF2使用了一种新颖的结构模块，将蛋白质表征为“残基气体”，即每个氨基酸由其主链原子构成的三角框架集合，这些框架可以在空间中自由旋转和平移。这种表征法结合了一种新颖的旋转和平移不变的注意力机制，利用前述残基对空间邻近数据，使网络能够高效地学习将残基框架转换为预测的蛋白质结构。

AF2最初只能预测单体蛋白质的结构。为了预测蛋白质复合物的结构，可以将两个蛋白质序列中加入一段柔性连接序列，使其成为一条单链序列，从而可以使用AF2进行预测。然而这种方法会干扰不同蛋白质亚基的MSA，其准确度并不理想。2021年，RoseTTAFold发布，其单体结构预测能力接近AF2，并且其多轨神经网络实现了蛋白质复合物结构的高预测精度[13]。基于该思路，DeepMind团队也调整AF2，开发了AF2-multimer方法，将复合物的各亚基进行独立的MSA比对，并提取亚基之间的共进化信息，成为当时蛋白质复合物结构预测的最佳方法。相关论文以预印本形式发表于2021年12月，并整合进github网站上AF2的官方版本中[14]。

1.2 其他基于深度学习的结构预测方法

AF2获得巨大的成功后，各种各样的基于深度学习方法进行蛋白质结构预测的方法纷纷涌现。其中，引人关注的方法除了AF2还包括上述RoseTTAFold以及ESMFold[13，15]。二者的准确度相比AF2略有差距，但各具特色。其中，RoseTTAFold的模型参数被用来开发生成式蛋白质设计方法RFdiffusion，与序列设计方法ProteinMPNN的联用已被广泛应用于各种条件下的蛋白质设计[16-17]。ESMFold的训练集同时包括UniRef50、Pfam等蛋白质序列聚类数据库，以及PDB结构数据库，使得其模型参数中包含了序列共进化信息，在结构预测时不再需要进行序列比对以获取MSA信息，预测速度大大提升。事实上，对于目标蛋白质序列，如果不加入MSA信息，AF2预测的准确性大大低于ESMFold。此外，DMFold将AF2中MSA比对分析流程升级到最新的DeepMSA2，进一步提升了蛋白质复合物预测的准确性，并在CASP15的相关竞赛中摘得桂冠[18]。

Mirdita等[19]基于AF2开发了ColabFold。该方法使用AF2预测的模型参数和推理过程，但是在序列比对过程中，没有使用jackhammer和HHblits方法，而是使用了速度更快的mmseq2，并且默认使用在线服务器进行序列比对。这使得其不仅显著加速了结构预测速度，而且节省了存储资源，因为AF2的数据库需要配置近3 T存储空间的固态硬盘。对于大多数蛋白质单体及复合物预测任务，其效果与未精简版的AF2基本类似。

1.3 AF2在结构预测领域的拓展应用

科研人员善于将一个好的工具用到极致。AF2发布后，其作为当时蛋白质结构预测最佳工具也得到了充分的扩展。CombFold将多结构域蛋白质序列切割为不同片段，并对不同亚基的片段进行排列组合，然后利用AF2分别预测并将置信度最高的片段组合进行组装，以最终提高复合物预测准确

度[20]；AlphaLink和AlphaLink2将化学交联产生的残基间距离信息整合进AF2中，以预测单体蛋白质和蛋白质复合物的结构[21-22]；ColabDock则将核磁共振（nuclear magnetic resonance，NMR）滴定数据、化学交联数据整合进AF2中[23]，等等。此外，AF2_conformations方法试图通过调整MSA的参数来获得不同深度的多序列比对结果，从而进行蛋白质多构象的预测[24]。但在多构象预测方法中更引人瞩目的是AF-Cluster方法[25]，它通过将MSA进行分簇，各簇内的序列具有较高的序列同源性，不同簇之间序列具有更大的不同，潜在地包含了该蛋白质不同构象的信息。针对每一簇的MSA比对，通过AF2的结构预测和打分，从而遴选出该序列潜在的不同构象。该方法在多个蛋白质的多构象解析中得到了验证。

在AF2划时代的突破后，2022年CASP15竞赛中，侧重点是蛋白质复合物预测的准确度和RNA结构预测。其中，较好的复合物预测方法基本上都是基于AF2-multimer，更有效地使用结构数据库模板，使用不同的MSA算法以增加采样，并使用dropouts技术（即训练过程中随机忽略部分参数），进行多次预测，最终选取得分最高的预测结构[26]。RNA结构预测的桂冠由国内团队摘取，其方法AIchemy_RNA2使用了类似AF2方法[27]。

2 AF3的重大升级：从蛋白质结构预测到生物分子复合物结构预测

AF2在蛋白质结构预测中取得了巨大成功，但无法预测蛋白质-核酸和蛋白质-小分子/离子复合物的三维结构。此外，抗原-抗体复合物预测准确度非常低。这三类复合物的三维结构对于蛋白质与其他生物大分子的相互作用研究以及医药开发，至关重要。

为了预测蛋白质-DNA/RNA复合物结构，David BAKER团队开发了RoseTTAFoldNA[28]。对于蛋白质-小分子复合物结构预测，因其直接关系到医药研发，基于深度学习的预测方法先后涌现。Umol和NeuralPlexer直接使用靶标蛋白质的三维结构或一级序列，以及小分子的结构式或三维坐标输入，通过指定结合位点或不指定任何位点，对靶标蛋白质和小分子的复合物进行结构预测，同时给出pLDDT等可信度预测[29-30]。David Baker团队[31]基于RoseTTAFold进一步开发了RoseTTAFold All-Atom，可以预测蛋白质与离子、小分子以及DNA/RNA等各种生物分子相互作用的复合物结构。其团队所开发的原限于蛋白质复合物设计的ProteinMPNN和RFdiffusion方法也分别有对应的升级版本，即LigandMPNN和RFdiffusion All-Atom[31-32]。而DeepMind团队与其合作公司Isomorphic Labs紧接着公布了其最新模型AlphaFold-latest-version的一些预测结果，显著地优于RoseTTAFold All-Atom，但相关算法没有披露。2024年4月，RoseTTAFold All-Atom相关论文发表于《科学》（*Science*）期刊并在github上开源[31]。2024年5月AF3文章发表于《自然》（*Nature*）期刊[3]。

AF3在算法上相比AF2出现了重大的变化。AF2使用MSA表征（MSA representation）和目标蛋白质序列“残基对表征”（residue pairwise representation）作为输入，利用Evoformer模块提取序列中的共进化信息。在AF3中，该模块被分别替换为Pairformer，其对蛋白质和核酸的MSA处理被大幅简化并提高了速度，并且其信息被传递给“标记水平表征”（token-level representation），只有后者被用作Pairformer的输入。AF3的标记水平表征中，一个标记（token）为蛋白质、核酸的单个残基或配体的单个原子。此外，AF3也利用Transformer算法在原子水平进行表征，并将其信息与标记水平表征进行交互通信。然后，AF3利用扩散模块（即diffusion module）进行结构预测。该模块与AF2的结构模块存在更大的差别，不再使用“残基气体”进行旋转、平移等计算，而是直接作用于原子水平，使用扩散模型进行训练。在结构推理时，接收原子的初始随机“噪声”坐标，然后通过模型进行反复“去噪”，产生最终结构。

相比AF2只能预测蛋白质及其复合物的结构，AF3可以根据输入的蛋白质一级序列、残基修饰、小分子配体和离子进行结构预测，其预测范围扩展到整个生物分子领域，速度相比AF2也得到显著提升，其准确性更是惊人。图1f展示了同源二聚的cAMP受体蛋白Clr与DNA及cAMP复合物的AF3结构预测结果，与晶体结构完全一致。相比AF2，AF3在蛋白质单体结构预测中略有提升，在蛋白质复合物结构预测中成功率（Cα原子RMSD小于2 Å）从68%提升到77%，尤其在抗原-抗体复合物结构预测中，成功率从30%提升到63%（图2）。在蛋白质-小分子-离子和蛋白质-DNA/RNA结构预测中，AF3的表现皆大幅领先于其他方法，仅

RNA结构预测略低于AIchemy_RNA2（图2）。AF3还可以较高的成功率预测生物大分子共价修饰后的三维结构。相比AF2及RoseTTAFold All-Atom等方法，AF3显著的性能提升，可能与利用Transformer进行原子水平表征和扩散模型应用有密切的关系。扩散模型具有高质量生成效果、训练高稳定性等特点[33-34]。AF3将其与Transformer连用，并直接在原子水平进行空间结构预测，具有一种彻底的端到端（end-to-end）的意味。在已有高质量、大规模的蛋白质序列和结构数据库的条件下，AF3取得优异的表现符合预期。

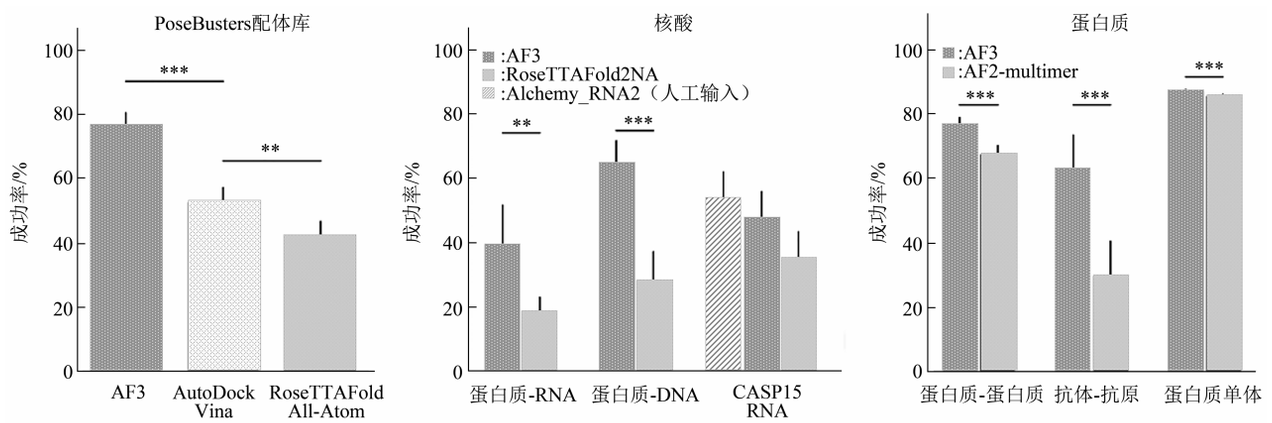


Fig. 2　Comparison of the success rate of AF3 in predicting protein−small molecule and other biomolecular complexes with other methods[3]

图2　AF3预测蛋白质−小分子等生物分子复合物的成功率与其他方法的比较[3]

其中，PoseBusters配体库为2023年8月更新版；**P<0.01，***P<0.001。

值得注意的是，AF3模型仍存在一些问题：a. 少量的建模分子手性错误；b. 在预测大蛋白质-核酸复合物（超过2 000个残基）时，偶尔会出现严重的空间冲突；c. 将蛋白质无序区域预测为紧凑的折叠构象。最后一个问题在很大程度上已通过利用AF2预测的无序结构数据集进行“交叉蒸馏”训练获得了解决。

到目前为止，AF3具体的代码和模型参数并未开源。谷歌DeepMind团队于论文在《自然》（*Nature*）期刊发布时提供了一个在线预测网站，研究人员注册后可以每天免费进行10次的预测任务。问题在于，其中的小分子组分只能选择其列表中的ATP、GTP等19种配体，离子则只能选择列表中的Mg^{2+}等9种阳离子和一种阴离子（Cl^-）。这事实上屏蔽了小分子药物相关的预测和评估需求。同时，虽然AF3在抗原-抗体复合物预测任务中，其准确率高达63%，但是其单次预测的准确性并不高，需要设置不同初始随机种子进行成百上千次的预测才能够达到最佳效果，使得其抗原-抗体预测所声称的精准度无法通过在线网站达到。

AF3研究论文发表后，加州大学旧金山分校的计算结构生物学家Stephanie Wankowicz等牵头起草了一封致《自然》（*Nature*）期刊编辑的公开信（https://zenodo.org/records/11391920），称赞AF2的发布和代码、模型参数的公开，促进了蛋白质设计和蛋白质复合物预测新方法和新应用的开发。同时，公开信批评DeepMind及其合作团队虽然高调宣传AF3的能力，却没有提供以高通量方式进行软件测试和使用的方法。这违背了学术传统，即学术论文需公开其研究方法，以被他人测试和验证。公开信作者们认为虽然公司有权对他们的创新保密，但是既要利用学术出版物的权威进行宣传，又要进行保密，使人无法重现其结果，这“颠覆了学术事业”（subverts the enterprise）。公开信也批评《自然》（*Nature*）期刊在AF3论文的审核中，实行了双重标准，也不符合科学共同体的需求。最后，作者们总结：“虽然科学的执行和传播方式在不断变化，但不论论文的作者是谁，期刊应通过确保科学在传播时可重复，来维持他们在科学社区中的角色。”，还强调：“AI方法现在直接影响生物发现和人类健康。充分实现其潜力不仅需要技术突破，还需要开放和合作的努力，建立在他人的发现基础上，这是所有科学研究的基础。”

截至2024年7月中旬，实名支持这封公开信的

全球学术界人士已超1 000人。这在一定程度上为学术界与工业界之间的互动提供了一个范例。学术界的研究主要接受政府的专项拨款和社会捐助，其初衷是为了人类的利益探究自然，开疆拓土，即信中所谓的“学术事业”（enterprise）。该事业是自文艺复兴时期以来科学共同体的一个理想。Francis Bacon（英国，1561—1626）在其《新大西岛》一书中如此描述该理想：“我们协会的目的是了解事物的因果关系和秘密运动；扩展人类帝国的疆界，实现一切可能的事情。”事实上，DeepMind团队所开发的AF2，其使用的序列数据库、PDB数据库、AI算法以及提取序列共进化信息的思路和方法都源自于学术界多年以来共同的积累，而AF2的开源也极大促进了蛋白质预测和设计领域的发展。这是学术界和工业界联手推动了学术事业发展的一个典型范例。另一方面，工业界资本的投入和利润追求也应得到尊重。AF3因其在药物研发中的潜力，研发者选择不公开其方法，本无可厚非。然而，当其团队选择在学术出版物上发表相关论文，同时不共享其方法时，就违背了学术界的传统和准则。在公开信发表之后，《自然》（*Nature*）期刊发表声明，称接受不公开代码的论文是出于生物安全的考虑。这种说法显然无法让人满意。令人鼓舞的是，AF3的作者在社交媒体上回应了学界的质疑和不满，将在线服务器每人每天的免费任务数量从10提高为20，并宣布将在6个月内为学术界提供开源模型。这一承诺已于2024年11月获得兑现（https://github.com/google-deepmind/alphafold3）

RoseTTAFold研发团队负责人David Baker曾预言因工业界强大的资源，基于学术界已开源的各种丰富资源，最强大的模型很可能由类似谷歌公司DeepMind这样的团队开发出来，并且表示其研究团队在RoseTTAFold和RoseTTAFold All-Atom等系列软件上的开发和开源发布在一定程度上就是为了促进AF2和AF3等方法的开源，并由此促进学术事业的发展[35]。值得一提的是，目前github代码开源网站上已发布了借鉴AF3文献中伪代码（pseudocode）实现的类AF3方法，HelixFold3（https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/protein_folding/helixfold3）和Chai-1（https://github.com/chaidiscovery/chai-lab），后者已引发极大关注。

3 AF2和AF3的深远影响和未来展望

到目前为止，DeepMind团队不仅开源了AF2的源程序和模型参数，同时建立了预测结构的数据库（https://alphafold.ebi.ac.uk/），包含UniProt数据库中超过2亿个蛋白质序列的预测结构。这极大地推动了结构生物学的发展，并推动了蛋白质设计等领域的加速发展。

3.1 辅助实验结构解析，推动结构生物学的普及

AF2和AF3预测的结构模型有助于重要蛋白质的基因构建优化，如界定分子内和分子间相互作用的序列，删除不必要的柔性区域，插入增强稳定性的设计片段，等等。所获得的高置信度的蛋白质复合物结构可用于晶体结构解析过程中的分子置换环节，并可作为模板进行低温冷冻电镜三维重构，从而加速传统的蛋白质复合物结构解析。其中的典型案例是《科学》（*Science*）期刊于2022年背靠背刊登五篇被称为结构生物学圣杯的核孔复合体的三维结构解析，其中Fontana等[36]利用AF2预测的亚复合物结构作为模板和较低分辨率的单颗粒冷冻电镜数据，成功地进行三维重构，获得了与实验结构一致的核孔复合体三维结构。

目前，冷冻电镜原位结构解析备受瞩目，提供了近生理状态下蛋白质和其他生物分子的三维电子密度信息，然而如何快速、准确将所处理数据中的电子密度归属于特定的蛋白质或其他生物分子，还需要进行大量的工作和努力。这一步骤也限制了原位结构方法的应用。可以预期，建立人的所有蛋白质相互作用高精度结构预测数据库，并通过算法的自动匹配和分析，将加速其中的电子密度匹配与归属过程，并加速原位电镜结构解析的应用。最新的AI辅助方法ModelAngelo已经在单颗粒冷冻电镜三维重构研究中，不仅实现了与人类专家相似水平的蛋白质原子模型构建，并且在识别未知序列蛋白质的能力方面甚至优于人类专家[37]。

传统上，为了获得一个蛋白质或其复合物样品，解析其高精度三维结构，往往需要花费数月甚至数年的时间。这也限制了结构生物学的应用。因为AF2等蛋白质预测方法的巨大成功，生物学家们越来越多直接使用置信度较高的预测结构，通过突变验证其准确性，即可进行功能和分子机制研究，而不需再解析其实验结构。这正在大大地推动结构

生物学的普及。AF2预测结构已被成功地用于小分子药物虚拟筛选[38]。在本文作者的合作研究中，生物化学和细胞生物学研究发现，阿米巴盘基网柄菌（*Dictyostelium discoideum*）中高度同源的Leep2A和Leep2B蛋白质构成了异源二聚复合体，其中Leep2A具有Leep2B依赖的小GTP酶RasD激活功能[39]。通过AF2的预测，可获得高置信度的Leep2A-Leep2B的复合物预测结构模型（图3a）。加入RasD进行三元复合物结构预测发现RasD与Leep2A GAP结构域发生相互作用，而非Leep2B GAP结构域（图3b）。另一方面，Leep2B与Leep2A GAP结构域及其附近区域形成了广泛的相互作用界面，对于Leep2A GAP结构域的构象具有显著的骨架支撑和空间精细定位的作用（图3c）。该结构模型较深入地揭示了Leep2A-Leep2B复合物功能的结构机制。

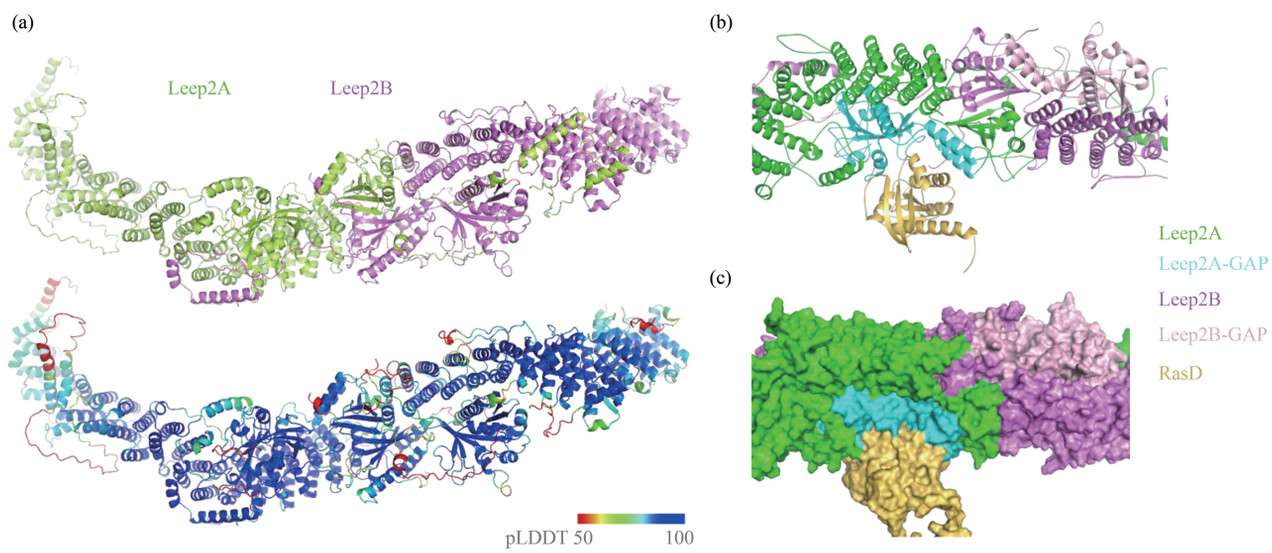


Fig. 3　Predicted structural models of the Leep2A−Leep2B complex

图3　Leep2A和Leep2B复合物的预测结构模型

（a）AF2预测的Leep2A-Leep2B复合结构（上）及其pLDDT着色图（下）；（b）Leep2A-Leep2B-RasD复合物的预测结构，其中Leep2A和Leep2B只显示了部分结构；（c）Leep2A-Leep2B-RasD复合物预测结构的表面展示。（b）和（c）的颜色含义如右侧标识。

使用AF2只能获得蛋白质或者蛋白质-蛋白质复合物的预测结构模型，而AF3则可以进一步获得蛋白质与DNA、小分子或离子的复合物结构，从而扩展上述研究对象范围。这些方法在加速生物大分子的功能机制研究的同时，也使研究人员可以将更多的时间和精力聚焦于蛋白质多构象、变构调控以及酶催化机理等更为艰深的研究中。

3.2　加速蛋白质设计发展

AF2和AF3等方法所带来的另一个革命性维度在于它们大大促进了蛋白质设计的发展。蛋白质设计主要包含骨架设计和序列设计两大方面。ProteinMPNN根据已有的骨架从头设计氨基酸残基，使用AF2对其设计的序列进行结构预测和筛选，该流程显示出极高的成功率，而且所设计的蛋白质展现出优秀的热稳定性[16]。结合ProteinMPNN和AF2，Sumida等[40]成功地优化TEV蛋白酶的氨基酸序列，使其具有更高的表达量、稳定性和酶活性，而Goverde等[41]则改造GPCR蛋白的跨膜区序列，使其具有良好的水溶性，同时维持了GPCR跨膜螺旋簇的构象以及螺旋簇两侧的信号通讯能力。AFdesign则综合序列采样、ProteinMPNN序列设计和AF2预测筛选，可以从头设计微结合蛋白（mini-binder，序列长度一般小于100个残基）的骨架和序列（https://github.com/sokrypton/ColabDesign）。在更为强大的骨架设计软件RFdiffusion被开发出来以后，RFdiffusion、ProteinMPNN和AF2的组合显示出强大的设计能力，在微结合蛋白的设计中，其成功率相较于传统的RoseTTA设计方法提高了三个数量级[17]。结合酵母展示和饱和突变筛选等亲和力成熟方法，已经具备工业应用的能力。这大大降低了蛋白质设计的难度，使得多数研究团队都可以从事相关研究。

AF3对蛋白质与DNA、小分子配体和离子等生物活性分子复合物结构的强大预测能力将显著扩展蛋白质设计的范围。特别是在酶的相关设计中，其体系往往包含金属离子及小分子等辅酶，其底物不仅包含蛋白质，也包含核酸以及脂类等其他生物分子，AF3所提供的复合物结构预测模型及其对重设计酶序列的筛选能力，都将为酶的设计优化提供更强大的支持力。这将有力地推动蛋白质研究的工业应用。

3.3 AFDB助力深度学习方法的开发

AF2提供了一个庞大的、规模超PDB数据库数百倍的蛋白质结构数据库——AlphaFold Database（AFDB），包括了所有UniProt数据库中约2.1亿条蛋白质序列的预测结构[42]。其中，pLDDT分数超过90的序列部分，其预测结构已被验证具有很高的准确度，甚至其侧链原子的构象也达到了实验结构的精度。这不仅为已有结构生物学研究提供助力，也为新的预测方法的开发提供了庞大的、高质量的训练集，而这对于基于大数据的深度学习方法的开发至关重要。该数据库已被用于DMFold、AF3等方法的训练集[3, 18]。此外，在蛋白质设计研究中，除了通过结构预测进行设计所得的备选序列的筛选，AF2和AF3提供的大规模、高质量数据库也推动了蛋白质设计方法的开发。基于AF2的数据库AFDB也被用于类似Pallatom等蛋白质设计方法的训练集[43]，而DeepMind团队利用基于AF3的蛋白质、核酸和配体复合物结构数据库，开发了最新一代蛋白质设计方法AlphaProteo（https://deepmind.google/discover/blog/alphaproteo-generates-novel-proteins-for-biology-and-health-research/）。据报道，AlphaProteo可同时设计主链骨架和氨基酸序列，其性能显著地超越RFdiffusion。毫无疑问，与AF3类似，AlphaProteo是否会发表于学术期刊，是否会开源，将同样引人关注。

3.4 加速药物和试剂研发

抗体主要利用其重链和轻链的互补决定区（complementarity-determining region，CDR）进行底物识别，然而CDR因其序列变化大，不保守，导致此前的AF2和RoseTTAFold等结构预测模型对抗体-抗原复合物的结构预测表现不够理想。利用SAbDab和VDJdb抗体结构数据库，David Baker团队[44]对RFdiffusion和RoseTTAFold2进行了针对抗体蛋白质复合物的微调。此外，模仿抗体利用CDR的柔性环进行蛋白质互作，二者的微调在训练集中也使用了PDB库中柔性环（loop）介导蛋白质互作的三维结构。微调后的RFdiffusion可针对抗原特定位点设计较高亲和力的抗体序列，而RoseTTAFold2可较准确预测所设计抗体与抗原的复合物结构，但对天然抗原-抗体结构预测的准确性仍然较低[44]。考虑到AF3在抗原、抗体复合物结构预测中所显示出的优越性能，可以预期基于AF3开发抗体设计模型和结构预测模型，将会取得更好的效果。如前所述，AlphaProteo在微结合蛋白的设计中已经体现出显著的性能提升。

微结合蛋白是一类潜在的类抗体药物，主要基于中和靶标蛋白质的策略来发挥功能。结合RFdiffusion、ProteinMPNN以及AF2等方法所设计的一种靶向蛇毒的微结合蛋白，作为中和药物已在动物模型中表现出优异的疗效[45]。针对细胞因子IL-23R和IL-17所设计的微结合蛋白药物对小鼠的结肠炎疗效优于临床抗IL-23的抗体，并展现出可以口服用药的特点和良好的药代动力学特征[46]。AF3因其更高的准确性和更广的生物分子适应性，将在这类微结合蛋白药物研发中发挥更重要的作用。

AF3在小分子药物研发领域中的潜力更为引人注意。AF3在无结合位点提示时进行蛋白质-小分子复合物预测的准确度已近80%，显著高于AutoDock vina等传统分子对接方法[3]，这将显著提高小分子药物虚拟筛选的准确性和效率。同时，基于AF3进行针对特定蛋白质靶标的小分子药物从头生成和优化设计，将具有生成高亲和力、高质量前导小分子药物的潜在应用前景，并极大地推动工具小分子研发和小分子医药研发。已有报道称，跨国药企礼来等公司已与AF3开发团队之一Isomorphic Labs达成战略合作协议，进行巨额投资以利用AF3进行药物研发（https://news.cnstock.com/industry,rdjj-202401-5173731.htm）。

因为抗体、多肽和小分子药物用于人体，需要考虑免疫原性、脱靶效应等副作用问题，其研发难度和成本都很高。与此相反，体外检测试剂和靶标生物分子功能调控试剂，需要解决的主要是亲和力和特异性，不存在副作用的问题，更适于AF3发挥其优势。可以预期，AF3将更为迅速地推动相关抗体、多肽和小分子试剂的研发，在科研领域和工业生产中具有广阔的应用前景。

4 AF3后时代的展望

虽然AF2和AF3带来了结构预测的巨大突破，但结构生物学的目标不只是预测蛋白质与其他生物分子的三维结构，还需要解析它们执行生理功能时发生的事情。例如，驱动蛋白的马达结构域在结合微管蛋白时加速释放产物ADP的速率、刺激其ATP酶活性，但我们还无法通过结构预测解析其中的变构调控机制。也无法预测转录因子与其DNA靶序列作用时的结合速率或亲和力会如何改变其行为。此外，活性位点和分子互作界面位点的点突变往往会显著地改变蛋白质的功能行为，而其他远程位点也可以远程变构调控成为蛋白质的功能开关。然而，目前AF2和AF3因其算法的内在特点还无法对点突变的影响做出成功的预测。

为了准确预测酶分子和转录因子在生理条件下如何执行特定功能，这些信息是必需的。预测蛋白质等生物分子的这些行为特征要比CASP竞赛目标远为复杂，因为后者的目标仅仅是得到一个与实验结构类似的静态构象。但在生理条件下，蛋白质及其复合物不会折叠成单一结构，而是形成数千种不同构象的集合，每种构象具有不同的概率和寿命，并且受环境因素、点突变的调控而发生改变，进而影响其功能状态。

在AF3之后，如何通过计算和预测方法研究这些构象景观以及它们在生物分子相互作用时如何变化，进而定量预测亲和力和动力学速率，变得至关重要。此外，预测点突变对蛋白质构象景观的影响，也同样重要。这意味着，我们需要新的“编译器”来解码由生物分子一维序列编码的动态信息，然后利用这种动态三维结构模型来预测其功能执行的分子机理[47]。这将是在AF2和AF3取得革命性突破之后的下一个重量级挑战。

参 考 文 献

[1] Zhuravlev P I, Papoian G A. Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. Q Rev Biophys, 2010, **43**(3): 295-332

[2] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol, 2019, **20**(11): 681-697

[3] Abramson J, Adler J, Dunger J, *et al*. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, **630**(8016): 493-500

[4] Bennett N R, Coventry B, Goreshnik I, *et al*. Improving *de novo* protein binder design with deep learning. Nat Commun, 2023, **14**(1): 2625

[5] Moult J, Pedersen J T, Judson R, *et al*. A large-scale experiment to assess protein structure prediction methods. Proteins, 1995, **23**(3): ii-v

[6] Zhou X, Zheng W, Li Y, *et al*. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc, 2022, **17**(10): 2326-2353

[7] Simons K T, Bonneau R, Ruczinski I, *et al*. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins, 1999, **Suppl 3:** 171-176

[8] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, **596**(7873): 583-589

[9] Ovchinnikov S, Kim D E, Wang R Y R, *et al*. Improved *de novo* structure prediction in CASP11 by incorporating coevolution information into Rosetta. Proteins, 2016, **84**(Suppl 1): 67-75

[10] Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins, 2018, **86**(Suppl 1): 67-77

[11] Du Z, Su H, Wang W, *et al*. The trRosetta server for fast and accurate protein structure prediction. Nat Protoc, 2021, **16**(12): 5634-5651

[12] Kryshtafovych A, Schwede T, Topf M, *et al*. Critical assessment of methods of protein structure prediction (CASP) -Round XIV. Proteins, 2021, **89**(12): 1607-1617

[13] Baek M, DiMaio F, Anishchenko I, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, **373**(6557): 871-876

[14] Evans R, O'neill M, Pritzel A, *et al*. Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2021. DOI: 10.1101/2021.10.04.463034

[15] Lin Z, Akin H, Rao R, *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, **379**(6637): 1123-1130

[16] Dauparas J, Anishchenko I, Bennett N, *et al*. Robust deep learning-based protein sequence design using ProteinMPNN. Science, 2022, **378**(6615): 49-56

[17] Watson J L, Juergens D, Bennett N R, *et al*. *De novo* design of protein structure and function with RFdiffusion. Nature, 2023, **620**(7976): 1089-1100

[18] Zheng W, Wuyun Q, Li Y, *et al*. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. Nat Methods, 2024, **21**(2): 279-289

[19] Mirdita M, Schütze K, Moriwaki Y, *et al*. ColabFold: making protein folding accessible to all. Nat Methods, 2022, **19**(6): 679-682

[20] Shor B, Schneidman-Duhovny D. CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. Nat Methods, 2024, **21**(3): 477-487

[21] Stahl K, Graziadei A, Dau T, *et al*. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning.

Nat Biotechnol, 2023, **41**(12): 1810-1819

[22] Stahl K, Warneke R, Demann L, *et al*. Modelling protein complexes with crosslinking mass spectrometry and deep learning. Nat Commun, 2024, **15**: 7866

[23] Feng S, Chen Z, Zhang C, *et al*. ColabDock: inverting AlphaFold structure prediction model for protein-protein docking with experimental restraints. bioRxiv, 2023. DOI: 10.1101/2023.07.04.547599

[24] Del Alamo D, Sala D, Mchaourab H S, *et al*. Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife, 2022, **11**: e75751

[25] Wayment-Steele H K, Ojoawo A, Otten R, *et al*. Predicting multiple conformations *via* sequence clustering and AlphaFold2. Nature, 2024, **625**(7996): 832-839

[26] Elofsson A. Progress at protein structure prediction, as seen in CASP15. Curr Opin Struct Biol, 2023, **80**: 102594

[27] Chen K, Zhou Y, Wang S, *et al*. RNA tertiary structure modeling with BRiQ potential in CASP15. Proteins, 2023, **91**(12): 1771-1778

[28] Baek M, McHugh R, Anishchenko I, *et al*. Accurate prediction of protein-nucleic acid complexes using RoseTTAFoldNA. Nat Methods, 2024, **21**(1): 117-121

[29] Bryant P, Kelkar A, Guljas A, *et al*. Structure prediction of protein-ligand complexes from sequence information with Umol. Nat Commun, 2024, **15**(1): 4536

[30] Qiao Z, Nie W, Vahdat A, *et al*. State-specific protein-ligand complex structure prediction with a multiscale deep generative model. Nat Mach Intell, 2024, **6**: 195-208

[31] Krishna R, Wang J, Ahern W, *et al*. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science, 2024, **384**(6693): eadl2528

[32] Dauparas J, Lee G R, Pecoraro R, *et al*. Atomic context-conditioned protein sequence design using LigandMPNN. bioRxiv, 2023. DOI: 10.1101/2023.12.22.573103

[33] Sohl-Dickstein J, Weiss E A, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics. Pr Mach Learn Res, 2015, **37**(1): 2256-2265

[34] Ho J, Jain A, Abbeel P J a I N I P S. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst, 2020, **33**(1): 6840-6851

[35] Doerr A. Protein design: the experts speak. Nat Biotechnol, 2024, **42**(2): 175-178

[36] Fontana P, Dong Y, Pi X, *et al*. Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. Science, 2022, **376**(6598): eabm9326

[37] Jamali K, Käll L, Zhang R, *et al*. Automated model building and protein identification in cryo-EM maps. Nature, 2024, **628**(8007): 450-457

[38] Lyu J, Kapolka N, Gumpper R, *et al*. AlphaFold2 structures guide prospective ligand discovery. Science, 2024, **384**(6702): eadn6354

[39] Chao X, Yang Y, Gong W, *et al*. Leep2A and Leep2B function as a RasGAP complex to regulate macropinosome formation. J Cell Biol, 2024, **223**(9): e202401110

[40] Sumida K H, Núñez-Franco R, Kalvet I, *et al*. Improving protein expression, stability, and function with ProteinMPNN. J Am Chem Soc, 2024, **146**(3): 2054-2061

[41] Goverde C A, Pacesa M, Goldbach N, *et al*. Computational design of soluble and functional membrane protein analogues. Nature, 2024, **631**(8020): 449-458

[42] Varadi M, Bertoni D, Magana P, *et al*. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res, 2024, **52**(D1): D368-D375

[43] Qu W, Guan J, Ma R, *et al*. P (all-atom) is unlocking new path for protein design. bioRxiv, 2024. DOI: 10.1101/2024.08.16.608235

[44] Bennett N R, Watson J L, Ragotte R J, *et al*. Atomically accurate *de novo* design of single-domain antibodies. bioRxiv, 2024. DOI: 10.1101/2024.03.14.585103

[45] Torres S V, Valle M B, Mackessy S P, *et al*. *De novo* designed proteins neutralize lethal snake venom toxins. Res Sq, 2024. DOI: 10.21203/rs.3.rs-4402792/v1

[46] Berger S, Seeger F, Yu T Y, *et al*. Preclinical proof of principle for orally delivered Th17 antagonist miniproteins. Cell, 2024, **187**(16): 4305-4317.e18

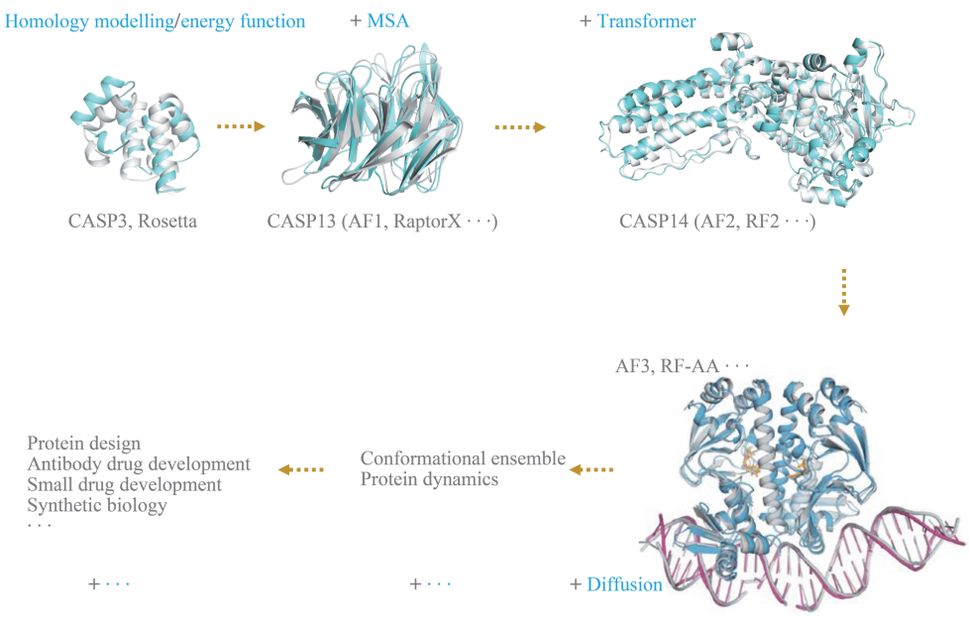
[47] Roy R, Al-Hashimi H M. AlphaFold3 takes a step toward decoding molecular behavior and biological computation. Nat Struct Mol Biol, 2024, **31**(7): 997-1000

Breakthrough of AlphaFold Structure Prediction and Its Impact and Challenges on Protein Research*

GONG Wei-Bin**

(*National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing* 100101, China)

Graphical abstract



Abstract In recent years, deep learning-based methods have achieved significant breakthroughs in protein structure prediction. The open-source release of AlphaFold 2 (AF2) in 2021 enabled high-precision prediction of three-dimensional structures for both individual proteins and protein complexes, allowing researchers to rapidly obtain reliable structural information and greatly accelerating advancements in protein structure and function studies. The release of AlphaFold 3 (AF3) in 2024 took this further by achieving accurate predictions of three-dimensional structures for protein-nucleic acid and protein-small molecule complexes. With improved algorithms and a more efficient model, AF3 significantly enhanced prediction accuracy, especially demonstrating outstanding performance in antigen-antibody and protein-small molecule complexes. The success of AlphaFold has not only brought revolutionary progress to structural biology but also showcased immense application potential in fields such as drug development, protein design, and molecular function research, driving innovation in biomedical studies. This article will review the development history of AlphaFold and related protein structure prediction methods, summarize their key technologies and current applications, and, by considering their limitations, provide an outlook on future research directions and applications.

Key words AlphaFold, deep learning, big data, protein structure prediction, biomolecular complex structure prediction, protein design

DOI：10.16476/j.pibb.2024.0374

* This work was supported by a grant from The National Natural Science Foundation of China (32371273).

** Corresponding author.

Tel：86-10-64888000，E-mail：gongweibin@ibp.ac.cn

Received：August 16，2024 Accepted：October 15，2024