大 学 化 学
Univ. Chem. **2025,** *40* (1), 75

•今日化学•　　doi: 10.12461/PKU.DXHX202411061　　www.dxhx.pku.edu.cn

浅谈2024年诺贝尔化学奖
——通过计算模拟和人工智能破译蛋白质结构的奥秘

王冯璋[1]，刘源[1]，王初[1,2,*]

[1]北京大学化学与分子工程学院，北京 100871

[2]北京大学前沿交叉学科研究院，北大清华生命科学联合中心，北京 100871

摘要：蛋白质是生命活动的主要承担者和执行者，蛋白质的三维结构与其生物学功能息息相关。2024年诺贝尔化学奖的一半奖项授予了DeepMind公司的Demis Hassabis博士和John Jumper博士，以表彰他们在蛋白质结构预测领域的突破性贡献；另一半奖项则授予了华盛顿大学的David Baker教授，以表彰他在蛋白质计算设计领域的系统性研究。本文简要回顾了蛋白质结构预测与设计这两个互逆问题研究历程的重点事件，并对它们的前沿应用进行了展望。

关键词：诺贝尔化学奖；蛋白质折叠；蛋白质结构预测；蛋白质计算设计；Rosetta；AlphaFold

中图分类号：G64；O6

A Brief Introduction to Nobel Prize in Chemistry 2024: Deciphering the Mysteries of Protein Structures by Computational Modelling and Artificial Intelligence

Fengzhang Wang [1], Yuan Liu [1], Chu Wang [1,2,*]

[1] College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

[2] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China.

Abstract: Proteins play major roles in all life activities, and the three-dimensional (3D) structures of proteins are closely related to their biological functions. In 2024, half of the Nobel Prize in Chemistry was awarded to Dr. Demis Hassabis and Dr. John Jumper from DeepMind, for their breakthrough contributions to protein structure prediction. The other half was awarded to Prof. David Baker from the University of Washington, for his systematic research in computational protein design. Here, we briefly review the research milestones of protein structure prediction and design, and prospect their cutting-edge applications.

Key Words: Nobel Prize in Chemistry; Protein folding; Protein structure prediction; Computational protein design; Rosetta; AlphaFold

2024年10月9日，瑞典皇家科学院宣布将2024年诺贝尔化学奖颁发给华盛顿大学的David Baker教授和DeepMind公司的Demis Hassabis博士和John Jumper博士(图1)，以分别表彰他们在蛋白质计算设计和结构预测领域的卓越贡献。

通常来说，研究者们将给定氨基酸序列、预测其三维结构的过程称为蛋白质的“结构预测”，

收稿：2024-11-19；录用：2024-12-04；网络发表：2024-12-16

*通讯作者，Email: chuwang@pku.edu.cn

基金资助：科技部重点研发计划专项(2022YFA1304700)；科技部重点研发计划专项(2023YFA1506500)

而将希望得到特定蛋白质结构和功能，并找到其对应骨架结构和氨基酸序列的过程称为蛋白质的“计算设计”。不难看出，蛋白质的结构预测与计算设计互为逆问题，分别对应了本次诺贝尔化学奖的两部分工作。

科学的发展往往经历“理解-创造”的过程，正如精准的结构预测是正确进行计算设计的基础。因此在下文中，我们将首先回顾蛋白质结构预测领域的重大进展。

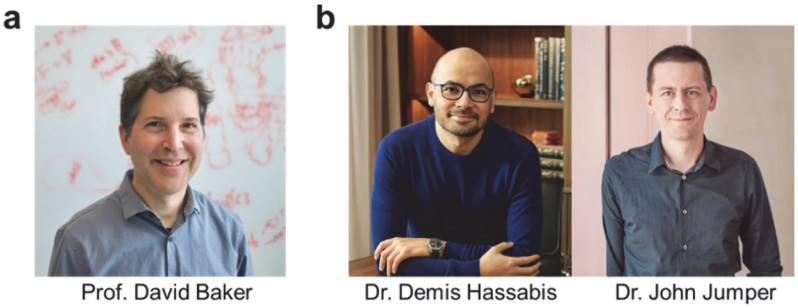

图1　2024年诺贝尔化学奖得主

(a) 来源于David Baker实验室主页；(b) 来源于Google DeepMind官网介绍

1　蛋白质的结构预测

1.1　Anfinsen热力学假说

20世纪70年代，诺贝尔化学奖得主Christian Anfinsen教授通过精巧的实验设计，描述了久负盛名的蛋白质折叠热力学假说(Thermodynamic Hypothesis)：天然蛋白质在正常生理环境(溶剂、pH、离子强度、金属或其他辅因子、温度)下的三维结构是整个系统吉布斯自由能(G)最低的结构[1]。换言之，给定环境下的蛋白质天然构象完全由其氨基酸序列决定。从那时起，理解蛋白质的序列-结构关系一直是化学和生命科学研究的重要科学问题。

1.2　CASP比赛的历史与传统结构预测方法

尽管X射线晶体衍射、核磁共振(NMR)和冷冻电镜(Cryo-EM)技术已被广泛用于蛋白质结构的实验测定，但由于周期长、成本高等原因，实验解析蛋白质结构的速度至今仍无法满足蛋白质序列爆炸式增长的需要，对高精度计算预测方法的需求是不言自明的。因此在1994年，马里兰大学的John Moult教授发起了全球蛋白质结构预测比赛(CASP)，至2024年已举办至第16届，旨在通过社区范围的评估推动领域进步。

CASP的30年历史见证了许多传统结构预测方法的发展，这其中也包括David Baker教授的Rosetta。实际上，David Baker教授早期的主要研究方向是蛋白质折叠动力学。在研究过程中，他展现出了对折叠机理的深刻洞见。David Baker教授基于一些实验事实和统计分析，推测蛋白质的折叠图景是多肽链的每个片段都采样了一系列满足其氨基酸序列的局部构象，而非局部相互作用可以优先稳定这些原本是瞬态的局部结构的近天然排列，从而使得这些片段同时满足正确的结构和取向，并推动疏水氨基酸埋藏在核心中时，折叠过程随之发生(图2)。基于这一理解，David Baker课题组开发了一种从头结构预测策略：将氨基酸侧链简化为一个球体，对于待折叠序列的局部片段，使用相似序列的已知结构来构建主链扭转角，随后通过模拟退火(Simulated annealing)方法不断优化结构，并开发了一套评分函数用于评估结构的质量[2]。这种结构预测方法在CASP3中首次报道，被David Baker教授的研究生Kim Simons命名为“Rosetta”[3]。

除用于从头建模的Rosetta以外，加州大学旧金山分校Andrej Sali等人开发的MODELLER [4]方法可用于同源建模任务，密歇根大学张阳等人开发的I-TASSER [5]算法融合了从头预测、同源比对和“穿线”(threading)等多种策略，有效地提升了表现。然而，在历年CASP比赛的结果中我们可以看出，

从CASP1到CASP12，传统方法的性能逐渐达到瓶颈(图3)。也正是从CASP12之后，人工智能(AI)、特别是深度学习(DL)技术开始崭露头角。

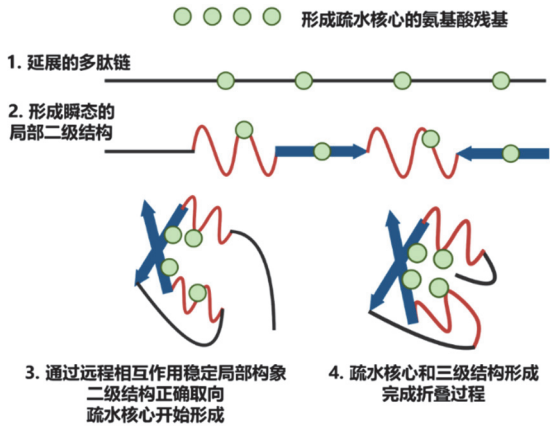

图2　David Baker教授提出的蛋白质折叠图景示意

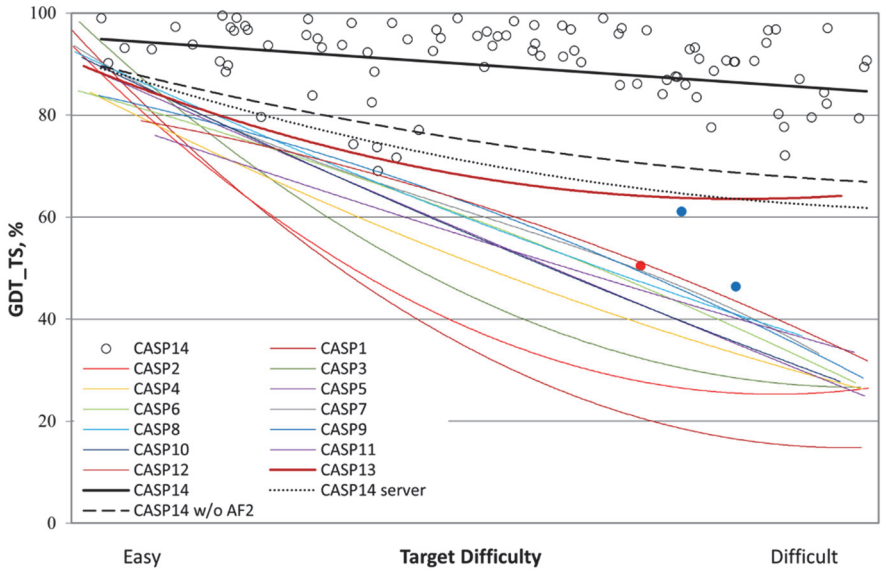

图3　历年CASP比赛结果[6]

电子版为彩图

1.3　AI技术的变革性力量

AI技术带来的变革起始于接触图(contact maps)预测。长度为L个残基的蛋白质接触图可表示为二进制的L*L矩阵，用于指示残基之间是否形成接触(图4a)。尽管使用较为简单的AI模型(如浅层神经网络)预测接触图[7]的想法早已出现，但芝加哥丰田计算技术研究所许锦波团队[8]于2017年开发的RaptorX-Contact方法率先使用深度残差卷积神经网络(ResNet [9])学习多序列比对(MSA)中的进化信息，从而大幅提高了接触图预测的性能。2018年，DeepMind团队在初代AlphaFold [10] (或称AlphaFold1)的开发中亦延续了这一思想，但将接触图扩展为信息更丰富的残基距离图(distogram)和主链扭转角分布(图4b)，并将AI模型输出的概率分布转换为势函数，从而能够使用简单的梯度下降来折叠蛋白质。2019年，David Baker课题组的杨建益等人进一步将距离图扩展至取向图(orientogram，图4c)，据此开发的trRosetta [11]额外纳入了残基之间的方向信息，并类似地将深度神经网络的输出转化为

Rosetta中使用的约束条件(restraints)，使用能量最小化即可得到最终的预测模型。

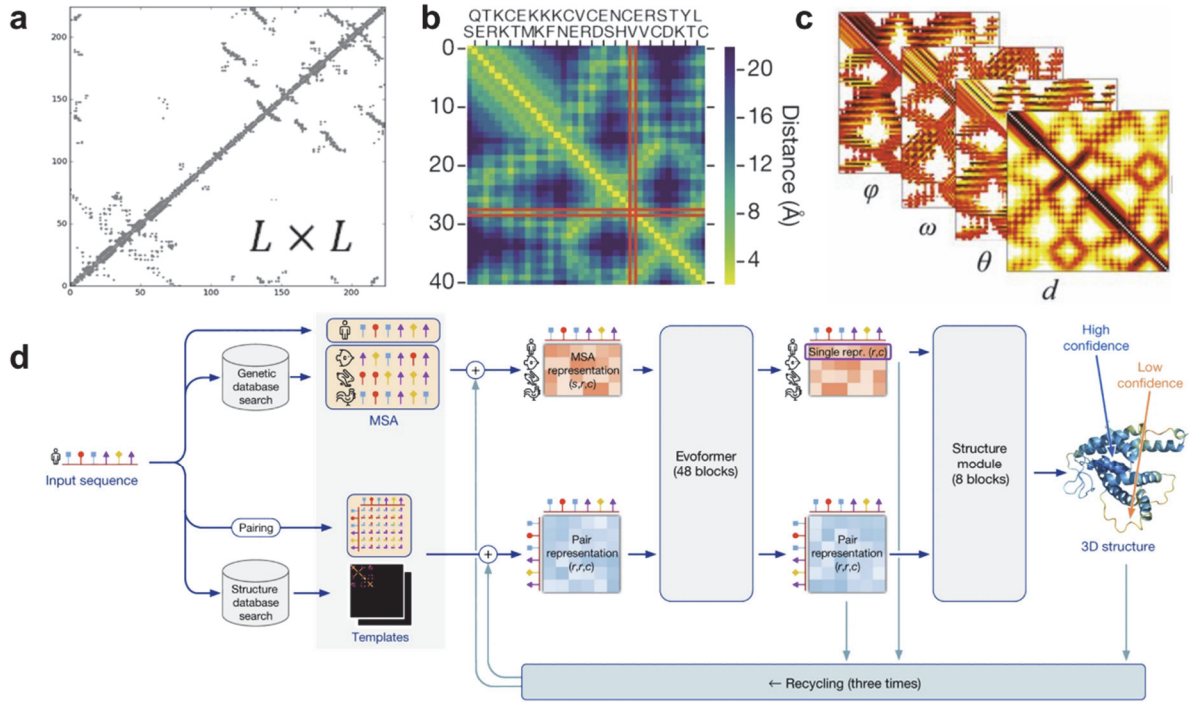

图4 (a) 蛋白质的接触图示意[8]；(b) 蛋白质的距离图示意[10]；(c) 蛋白质的取向图示意[11]；(d) AlphaFold2模型架构[12]

尽管AlphaFold1在CASP13上相较于其他方法已经取得了较大的优势，但其核心思想和所使用的计算策略仍建立在学界的研究基础上。然而在两年后的CASP14中，DeepMind凭借AlphaFold2 [12]再次震惊了整个领域，其对单链蛋白的预测性能已达到原子级，可认为在大部分情况下，预测结构可与实验解析结果媲美。相比于AlphaFold1，AlphaFold2并不是一次简单的升级，而是引入了全新的、原创性的神经网络模块和模型训练技术(图4d)。其中，Evoformer模块接受从输入序列中构建的多序列比对表示和成对特征，使序列信息和几何信息可以充分流动交互。随后，与此前接触预测-能量最小化的两步策略不同，新颖的结构模块(structure module)通过不变点注意力机制进一步处理Evoformer模块的输出，实现了“端到端”(end-to-end)的预测，即从网络中直接输出预测结构。此外，损失函数、自蒸馏等训练技术上的优化也是AlphaFold2取得成功的关键因素。

AlphaFold2论文发表当年即获评*Nature* 2021年度方法[13]。随后，AlphaFold2预测了人类蛋白质组的所有结构[14]，并迅速推进至更多常见物种。如今，AlphaFold蛋白质结构数据库(AlphaFold DB)中已经存储了超过2亿个蛋白质结构[15]。而Demis Hassabis和John Jumper两位科学家获得2024年诺贝尔化学奖距离AlphaFold2原始论文于2021年发表在*Nature*杂志仅仅过去了三年时间，足以窥见AlphaFold2的巨大影响力。时至今日，AlphaFold2原始论文已被引用逾17000次。可以认为，AlphaFold2已经对蛋白质相关的生物学研究产生了变革性的影响。

2 蛋白质的计算设计

2.1 蛋白质计算设计的曙光

蛋白质结构预测问题是基于给定序列预测三维结构，而蛋白质设计问题可认为是结构预测的

逆问题，旨在找到能够折叠至给定结构的潜在骨架和序列。蛋白质设计同样有着数十年研究历史，如David Baker教授在获奖感言中所提及的，加州大学旧金山分校的William F. DeGrado教授通过一系列螺旋束蛋白证明蛋白质理性设计的可能性[16–18]，而加州理工学院的Stephen L. Mayo教授率先开发了自动化的序列选择算法用于蛋白质的从头设计[19]。这些工作证明了蛋白质的计算设计是可实现的，启发了David Baker教授的研究。

2.2 Rosetta: 从结构预测到Top7的设计

如前文所述，在Rosetta被成功用于蛋白质结构预测之后，David Baker实验室也在不断对Rosetta进行迭代更新。随着侧链全原子模型的加入和评分函数精度的逐渐提高。David Baker授敏锐地意识到，Rosetta工具亦可以用于蛋白质设计问题。他的博士后Brain Kuhlman从蛋白质结构拓扑服务器中选择了一种PDB数据库中不存在的拓扑结构，并开发了一种序列优化-结构预测迭代的从头设计算法。该方法通过片段组装生成骨架结构，随后通过蒙特卡洛优化程序尝试序列替换，最终于2003年成功从头设计了具有全新拓扑结构的蛋白Top7 [20] (图5a)，成为蛋白质设计历史上的里程碑事件。

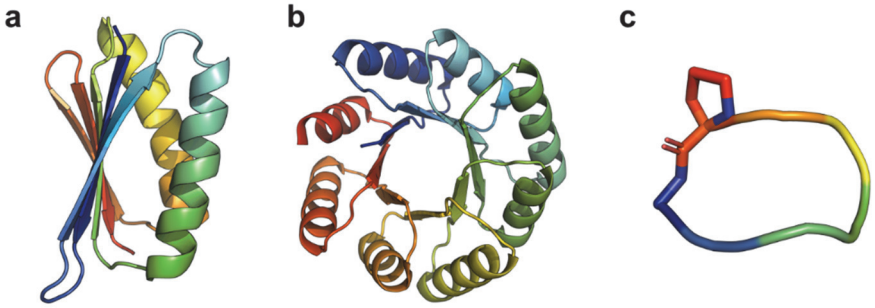

图5 (a) Top7的结构；(b) 对称*β*桶的设计；(c) 复杂环状多肽的准确设计

尽管不具有任何生物学功能，但Top7的成功无疑为更复杂的蛋白质设计铺平了道路。伴随着二十余年Rosetta的优化升级，David Baker实验室也在不断开拓蛋白质功能化设计的边界。例如正交的蛋白质相互作用对[21,22]、催化特定化学反应的酶[23,24]、稳定抗体结合表位的蛋白质疫苗[25,26]、结合关键致病蛋白的治疗性分子[27,28]、对称的蛋白质材料[29,30]、具有全新复杂拓扑结构的蛋白(图5b)和多肽链(图5c) [31,32]等等。值得指出的是，除经典的Rosetta方法外，David Baker实验室近些年也广泛将AI方法用于蛋白质结构预测和设计中。例如与AlphaFold2媲美的结构预测方法RoseTTAFold [33]、全新的主链设计方法RFdiffusion [34]和序列设计方法ProteinMPNN [35]等。这些基于AI的设计流程现已被David Baker实验室内外广泛采用，大幅提高了计算蛋白质设计的可及性、易用度和成功率。

3 展望未来

人工智能的发展已经且正在持续为蛋白质结构预测和设计领域注入新的活力。AlphaFold2已经能够很好地预测单链蛋白的结构，而最近DeepMind推出的AlphaFold3 [36]已经能够预测蛋白质与其他生物分子如核酸、辅因子、药物分子等之间的相互作用。即便如此，蛋白质的突变效应、行使功能所需的动态性、大型组装体等问题的预测尚缺乏很好的解决方案。而在蛋白质设计领域，治疗性结合剂、酶、膜蛋白、配体结合蛋白和抗体等复杂体系的设计潜力仍值得深入挖掘。此外，我们对序列-结构关系的生物物理学理解尚不足以解释所有的实验现象。但不难想见，蛋白质结构预测和设计领域将在获得2024年诺贝尔化学奖后得到更多的关注、发展和转化，并最终造福于全人类。

参 考 文 献

[1] Anfinsen, C. B. *Science* **1973,** *181*, 223.

[2] Simons, K. T.; Kooperberg, C.; Huang, E; Baker, D. *J. Mol. Biol.* **1997,** *268*, 209.

[3] Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. *Proteins* **1999,** *37*, 171.

[4] Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993,** *234*, 779.

[5] Roy, A.; Kucukural, A.; Zhang, Y. *Nat. Protoc.* **2010,** *5*, 725.

[6] Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. *Proteins* **2023,** *91*, 1539.

[7] Fariselli, P. *Protein Eng.* **1999,** *12*, 15.

[8] Wang, S.; Sun, S.; Li, Z.; Zhang, R. Xu, J. *PLOS Comput. Biol.* **2017,** *13*, e1005324.

[9] He, K.; Zhang, X.; Ren, S. Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun 27–30, 2016; IEEE: Las Vegas, NV, USA, 2016; pp. 770–778.

[10] Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; *et al. Nature* **2020,** *577*, 706.

[11] Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. *Proc. Natl. Acad. Sci. U. S. A.* **2020,** *117*, 1496.

[12] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; *et al. Nature* **2021,** *596*, 583.

[13] Method of the Year 2021: Protein Structure Prediction. *Nat. Methods* **2022,** *19*, 1.

[14] Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; *et al. Nature* **2021,** *596*, 590.

[15] Varadi, M.; Bertoni, D.; Magana, P.; Paramval, U.; Pidruchna, I.; Radhakrishnan, M.; Tsenkov, M.; Nair, S.; Mirdita, M.; Yeo, J.; *et al. Nucleic Acids Res.* **2023,** *52*, D368.

[16] Regan, L.; DeGrado, W. F. *Science* **1988,** *241*, 976.

[17] Lovejoy, B.; Choe, S.; Cascio, D.; McRorie, D. K.; DeGrado, W. F.; Eisenberg, D. *Science* **1993,** *259*, 1288.

[18] Handel, T. M.; Williams, S. A.; DeGrado, W. F. *Science* **1993,** *261*, 879.

[19] Dahiyat, B. I.; Mayo, S. L. *Science* **1997,** *278*, 82.

[20] Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* **2003,** *302*, 1364.

[21] Kortemme, T.; Joachimiak, L. A.; Bullock, A. N.; Schuler, A. D.; Stoddard, B. L.; Baker, D. *Nat. Struct. Mol. Biol.* **2004,** *11*, 371.

[22] Karanicolas, J.; Corn, J. E.; Chen, I.; Joachimiak, L. A.; Dym, O.; Peck, S. H.; Albeck, S.; Unger, T.; Hu, W.; Liu, G. *et al. Mol. Cell* **2011,** *42*, 250.

[23] Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; *et al. Science* **2008,** *319*, 1387.

[24] Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; *et al. Nature* **2008,** *453*, 190.

[25] Correia, B. E.; Bates, J. T.; Loomis, R. J.; Baneyx, G.; Carrico, C.; Jardine, J. G.; Rupert, P.; Correnti, C.; Kalyuzhniy, O.; Vittal, V.; *et al. Nature* **2014,** *507*, 201.

[26] Correia, B. E.; Ban, Y.-E. A.; Holmes, M. A.; Xu, H.; Ellingson, K.; Kraft, Z.; Carrico, C.; Boni, E.; Sather, D. N.; Zenobia, C.; *et al. Structure* **2010,** *18*, 1116.

[27] Fleishman, S. J.; Whitehead, T. A.; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E.-M.; Wilson, I. A.; Baker, D. *Science* **2011,** *332*, 816.

[28] Silva, D.-A.; Yu, S.; Ulge, U. Y.; Spangler, J. B.; Jude, K. M.; Labão-Almeida, C.; Ali, L. R.; Quijano-Rubio, A.; Ruterbusch, M.; Leung, I.; *et al. Nature* **2019,** *565*, 186.

[29] King, N. P.; Bale, J. B.; Sheffler, W.; McNamara, D. E.; Gonen, S.; Gonen, T.; Yeates, T. O.; Baker, D. *Nature* **2014,** *510*, 103.

[30] King, N. P.; Sheffler, W.; Sawaya, M. R.; Vollmar, B. S.; Sumida, J. P.; André, I.; Gonen, T.; Yeates, T. O.; Baker, D. *Science* **2012,** *336*, 1171.

[31] Hosseinzadeh, P.; Bhardwaj, G.; Mulligan, V. K.; Shortridge, M. D.; Craven, T. W.; Pardo-Avila, F.; Rettie, S. A.; Kim, D. E.; Silva, D.-A.;

Ibrahim, Y. M.; *et al. Science* **2017,** *358*, 1461.

[32] Bhardwaj, G.; O'Connor, J.; Rettie, S.; Huang, Y.-H.; Ramelot, T. A.; Mulligan, V. K.; Alpkilic, G. G.; Palmer, J.; Bera, A. K.; Bick, M. J.; *et al. Cell* **2022,** *185*, 3520.

[33] Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; *et al. Science* **2021,** *373*, 871.

[34] Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; *et al. Nature* **2023,** *620*, 1089.

[35] Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; *et al. Science* **2022,** *378*, 49.

[36] Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; *et al. Nature* **2024,** *630*, 493.