DOI: 10.13376/j.cbls/2021161
文章编号：1004-0374(2021)12-1436-09

·技术·

吴边，北京大学药学本科、荷兰格罗宁根大学博士毕业。现就职于中国科学院微生物研究所，博士生导师、研究员，获国家自然科学基金委优青项目资助。主要工作致力于微生物酶相关的机制解析与功能设计。近年来，研究团队将蛋白质计算的前沿技术引入微生物研究领域中，解析了数类微生物碳氮成键酶的进化机制与反应机理，通过人工改造将其应用于生物分子的精准合成与定向修饰，促进了微生物大分子元件设计的发展。相关工作发表在 *Nature Catalysis*、*Nature Chemical Biology*、*National Science Review*、*ACS Catalysis*、*Advanced Science* 等学术刊物上。

合成生物学中蛋白质计算预测设计的应用与发展

孙璡原[1,2]，崔颖璐[1]，吴　边[1*]
(1 中国科学院微生物研究所，中国科学院微生物生理与代谢工程重点实验室，微生物资源前期开发国家重点实验室，北京 100101；2 中国科学院大学生命科学学院，北京 100049)

摘　要：合成生物学以“工程化”为核心指导思想，自下而上开发生物技术来解决人类社会面临的重大挑战。蛋白质作为生命活动的直接执行者和合成生物学中关键的底层元件，对其定量认识和工程改造的能力直接影响合成生物学的上层建筑。通过蛋白质计算设计技术可实现功能空间跳跃，为合成生物学提供全新元件，使序列 - 结构 - 功能的研究从“格物致知”转化为“建物致知”的新范式。该文介绍了在蛋白质序列 - 结构 - 功能的预测和设计中采用的前沿技术和应用进展，讨论了蛋白质计算设计面临的科学挑战，并提出了应对挑战应优先发展的研究方向。

关键词：合成生物学；元件工程；蛋白质计算预测；蛋白质计算设计
中图分类号：Q811.4　　**文献标志码：**A

Application and development trend of computational protein prediction and design in synthetic biology

SUN Jin-Yuan[1,2], CUI Ying-Lu[1], WU Bian[1*]
(1 CAS Key Laboratory of Microbial Physiological and Metabolic Engineering, State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China; 2 University of Chinese Academy of Sciences, Beijing 100101, China)

Abstract: Synthetic biology relies on core principles of engineering and offers a bottom-up approach to use and build upon our vast leaps in the molecular understanding of biological systems to address challenges facing today's society. As the direct executor of life activities, proteins are essential building blocks in synthetic biology. The quantitative understanding and engineering capabilities of them directly affect the superstructure of synthetic biology. Through computational protein design, sequence jump along the fitness landscape would be achieved and

收稿日期：2021-10-12
基金项目：国家重点研发计划 (2018YFA0901600)
***通信作者：**E-mail: wub@im.ac.cn

provide new building blocks for synthetic biology. A paradigm shift of the study of sequence-structure-function occurs from "observe to understand" to "build to understand". Here, we reviewed the studies following the concept of “build to understand” to predict and design protein structure and function by computational protein design and the applications of this state-of-the-art method. The challenges faced by computational protein design and the proposed solutions are also discussed.

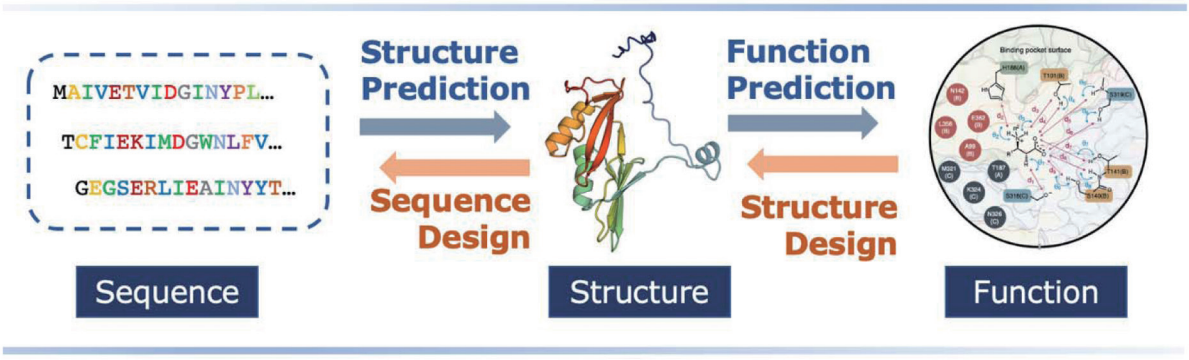
Key words: synthetic biology; building block engineering; *in silico* protein prediction; computational protein design

合成生物学致力于在理解生物学的基础上引入工程学，用生物元件构筑具有新功能的系统。蛋白质是执行生物功能的主要大分子，也是构筑生物系统的基本元件，实现其高效的设计是合成生物学的发展核心目标之一 [1]。

随着人们对蛋白质构效关系的逐步认识和基因操纵技术的出现，对蛋白质进行功能调控的蛋白质工程应运而生，深刻影响了当代生物学的发展，对生物产业也产生了巨大的推动作用。蛋白质工程体系中的第一代理性设计技术与定向进化技术因此于1993 年和 2018 年两获诺贝尔奖。但是，传统理性设计完全依赖于研究人员的经验，这不仅导致设计成功率低，而且获得的成功难以复制。定向进化在一定程度上规避了先验知识的限定，通过随机突变和筛选来引导蛋白质向特定的预设方向进化，但定向进化筛选系统的适用性与通量限制往往会成为复杂蛋白质工程的瓶颈。因此，学界长期以来一直在探索能够系统实现蛋白质功能空间大幅度跃迁的设计方法，为合成生物学高效地提供崭新的功能元件。

依据“序列 - 结构 - 功能”这一黄金法则，蛋白质计算领域可分为四大重要问题：分别是从序列预测结构、从结构预测功能、根据结构设计序列以及根据功能设计结构 (图 1)。蛋白质系统的序列 - 结构 - 功能空间是非常庞大的。在利文索尔佯谬 (Levinthal’s paradox) 中，一个长度为 100 个氨基酸的小型蛋白质，不考虑氨基酸组合和侧链构象的变化，仅主链构象空间即高达 10^{143} 种变化。从数学角度考虑，蛋白质的计算预测和设计的复杂度导致相关问题几乎无法被精确求解。因此，开发有效的近似算法，以牺牲可接受范围内的精度来大幅度压缩搜索空间，是蛋白质计算的核心任务。其中，蛋白质的计算预测致力于解决根据序列预测结构和根据结构预测功能，着眼于解决一个客观存在的生物序列向结构和功能空间映射的问题。蛋白质计算设计则致力于解决根据结构设计序列以及根据功能设计结构这两个重大问题，其终极目标是：利用计算机算法，设计具有所需功能且能够折叠成特定结构的蛋白质 [2]。

蛋白质计算领域的萌芽可以溯源到 20 世纪 80 年代，早期 DeGrado 进行了蛋白质设计的初步尝试，使用基于规则的启发式设计方法成功构建出稳定的四股螺旋束 [3]。随后，基于大分子力场和侧链旋转异构体 (rotamer) 库，出现了通过自动优化能量函数进行序列设计的计算方法 [4]。相比于单纯启发式的设计方法，基于能量函数的自动设计方法不受主链结构类型的限制；此外，定量计算残基之间特异性



(1)根据序列预测结构；(2)根据结构预测功能；(3)根据结构设计序列；(4)根据功能设计结构。

图1　蛋白质计算领域的四大问题

的空间堆积和氢键等相互作用，提高了设计的成功率。进入 21 世纪，Baker 首先设计出了自然界中不存在的折叠类型，引领了蛋白质骨架从头设计的先河。2008 年，Baker 提出了由内而外的策略，通过计算设计人工创造出 Kemp 消除酶 [5]、Diels-Alder 合成酶 [6] 和缩醛酶 [7] 等数个非天然酶，蛋白质计算设计从此开始对主流生物学研究产生影响。近年来，蛋白质从头设计中出现的算法被应用于天然蛋白结构的功能重塑，出现了蛋白质计算重塑这一方向。Baker 课题组利用从数据库中发掘的特殊苯甲醛裂解酶 (BAL)，利用 Foldit 和 RosettaDesign，重新设计出甲醛聚合酶 (FLS) 催化甲醛聚合 [8]。2021 年，马延和团队对 FLS 进行设计提高其活性，在利用二氧化碳合成淀粉的体外通路中实现了无机碳到有机碳的关键转化步骤 [9]。

伴随大数据和人工智能发展的浪潮以及测序数据的积累，近期还出现了数据驱动型的蛋白质计算设计方法，成功实现了多样的腺相关病毒衣壳蛋白 [10]、蛋白质传感器 [11]、蛋白质逻辑门 [12]、跨膜蛋白 [13] 和结合新冠病毒的小蛋白 [14] 等设计案例。在蛋白质结构计算预测获得突破，蛋白质计算设计算法不断涌现的背景下，发展支撑合成生物学的蛋白质计算设计平台的相应条件已然成熟 (图 2)。

1 蛋白质计算的发展与现状

1.1 蛋白质计算结构预测

蛋白质结构预测是蛋白质计算领域的重要基础问题。结构预测可以上溯到 1972 年诺贝尔化学奖得主 Christian B. Anfinsen 的著名论断，即蛋白质折叠的全部信息蕴含于其序列中 [15]。然而，如何解读氨基酸序列中蕴含的结构信息却困扰了生物学家 50 年之久。为了解决这个问题，学界从 1994 年起开始举办蛋白质结构预测的关键评估(Critical Assessment of protein Structure Prediction, CASP) 比赛，极大推动了蛋白质结构计算预测方法的发展。

蛋白质结构预测按输入信息可以分为有模版和无模版两类。目前较为成熟的有模版建模以 I-TASSER 为代表 [16]，使用穿线法进行结构预测，可以有效地提取来自多个模版的结构信息，生成模型“诱饵”(decoy)，聚类后再利用动力学模拟等方法进行优化，在多届 CASP 比赛中拿下了冠军。而无模版方法首先要利用目标蛋白与同源蛋白的多序列比对 (multiple sequence alignment, MSA) 提取共进化信息，预测残基接触图，再反推结构的原子坐标，并使用大分子力场进行侧链构象的优化。在 2018 年的第 13 届 CASP 中，AlphaFold 横空出世 [17]，

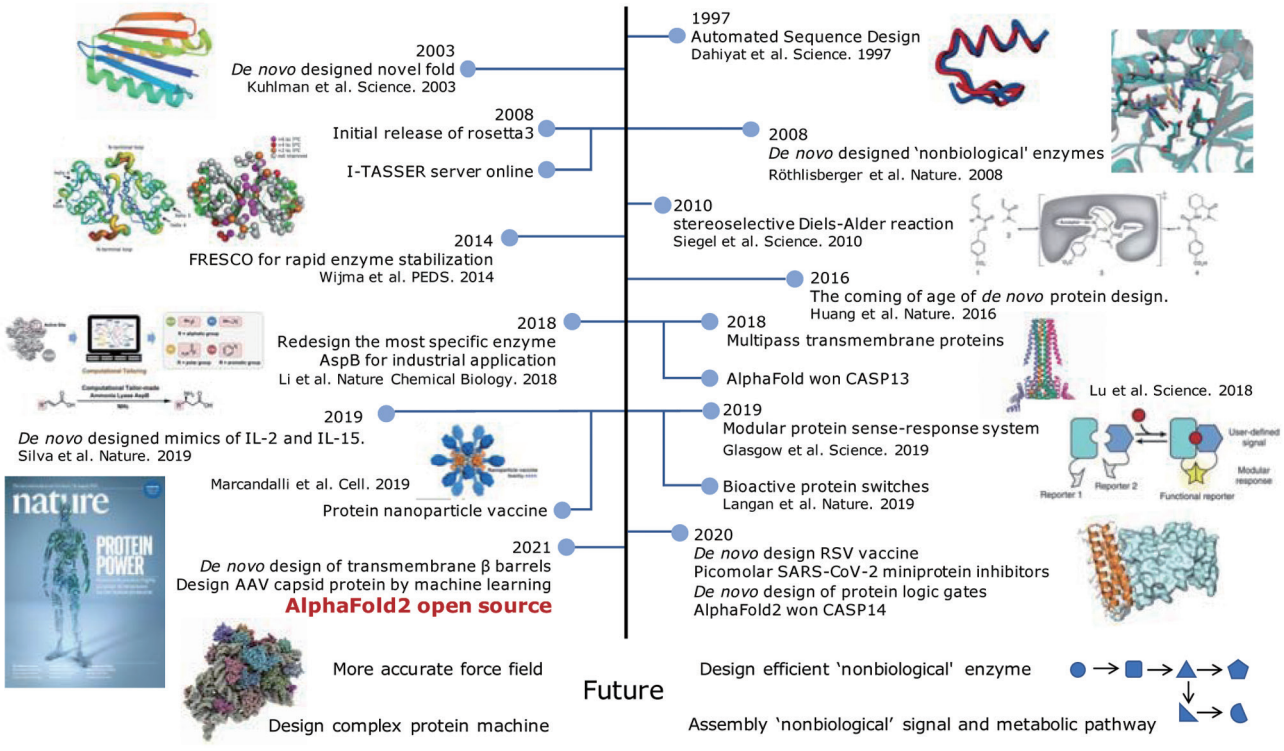


图2 蛋白质计算预测与设计发展历程与未来方向

从 MSA 中使用 Potts model 提取特征，使用卷积神经网络 (CNN) 预测残基接触，进一步利用约束建模结构，将预测准确性提高了约 15%。随后，trRosetta 进一步优化了这种方法，加入了残基间朝向，提高了计算速度和预测准确性 [18]。2020 年的第 14 届 CASP 比赛中，经过全面架构调整的 AlphaFold2 不再预测残基接触图，转而提出了 Evoformer+structure module 的网络架构，结合循环迭代等工程手段，完成了端到端的结构预测，使用神经网络从一维序列直接预测原子坐标，将预测结果与实验模型的全原子均方根偏差降低到了约 1.5 Å，同时其预测结果可以给出预测的局部距离差异测试，提供预测置信度信息 [19]。2021 年，AlphaFold2 正式开源，公开了 21 个物种的全基因组蛋白质结构预测数据库 AlphaFold DB，并且进一步训练了有能力预测多聚体的 AlphaFold-Multimer[20]。至此，对于有天然稳定结构的蛋白质序列 - 结构预测问题基本得到了解决。

1.2　蛋白质计算功能预测

长期以来，由于高精度的蛋白质结构难以获得，蛋白质功能计算预测长期停留在利用序列同源性进行推断的范式中，基于序列比对的 BLAST 和基于隐马尔可夫模型 (HMM) 的 HMMER，都对早期蛋白质功能计算预测的发展起到了重要的作用 [21]。随着蛋白质结构数据的积累，出现了结构比对方法 COATH 和 COFACTOR，作为 I-TASSER 在线服务器的一部分，解决了部分有同源结构的蛋白质功能预测，包括小分子结合口袋、基因本体 (GO)、酶学分类和催化活性位点的预测 [16]。但是，此种预测方法严重依赖于实验的先验知识，难以自动推断新功能，也难以标注无显著同源序列的新序列功能。

随着人工智能技术发展和结构预测精度的提高，近年出现了一些不依赖于先验知识转移注释的蛋白质功能预测方法。为了解决蛋白质序列的酶学分类号 (EC number) 的注释，DEEPre 从目标序列的 MSA 结合 HMM 提取特征，利用 CNN 结合长短期记忆 (long short-term memory, LSTM) 网络，预测 EC number 的准确性超过了以 COFACTOR 为代表的转移注释方法 [22]。为了蛋白质结构的小分子结合口袋预测问题，Tuffery 课题组基于 Voronoi 镶嵌模型和 alpha 球体概念，开发了 fpocket 软件，可在预测小分子云的同时预测口袋的可药性 (Drugability)；作为开源软件，该方法在速度、精度和预测信息丰富程度上较为优秀 [23]。DeepFRI 基于蛋白质结构，使用具有语言模型特征的图卷积网络，实现了基于结构的蛋白质功能预测和功能残基识别，可给出 GO 注释、酶学分类号注释和例如金属结合位点的关键残基 [24]。谷歌开发的 ProteInfer 使用扩张的卷积神经网络，能够从单序列直接预测 GO 和酶学分类号注释，在与序列比对方法结合时可以产生较好的效果 [25]。

1.3　蛋白质序列计算设计

根据蛋白质结构设计序列通常在蛋白质计算设计领域内被称为固定主链的序列设计，因为相比于一个给定的狭小结构空间，其对应的蛋白质序列空间是庞大的，而且负作用的上位效应会急剧损害设计出的蛋白质的可折叠性，因此序列设计并不是结构预测的等价逆问题，需要开发针对性的算法、软件和策略。蛋白质计算设计中广泛使用的方法和算法可被大致分为以下几类 [26]：

(1) 侧链放置。即在给定一个蛋白质骨架的结构的基础上，选择一组合适的氨基酸侧链构象，使之能够满足主链结构的要求。由于这实际上设计了序列，也称为蛋白质序列设计。

(2) 主链生成。即根据设计的需要生成一个主链构象的模型，在这个模型的基础上进行序列设计。

(3) 刚体放置。即固定蛋白质与蛋白质或蛋白质与小分子之间相对的空间位置和朝向。这通常用于设计具有结合活性的蛋白质或者酶。

(4) 负设计。即设计时提高非目标状态的能量以实现更好的折叠，可以看作是侧链放置算法中的优化与补充。

主链生成、刚体放置和负设计通常根据设计的目标会有所区别，但是最终都会生成骨架的模型，使用序列设计方法设计合适的序列。

蛋白质计算设计通常采取三步走的策略。第一步，将离散的侧链构象放置于主链上；第二步，使用能量函数计算被放置的侧链与侧链、侧链与主链之间的能量；第三步，使用搜索算法优化序列和构象的组合。整个过程涉及一系列的序列组合及其对应结构的优化，主链骨架是被事先给定的 (如来源于天然蛋白质结构)，且被假设为固定不变。设计中需要通过计算来确定的未知量包括每个主链位置上的氨基酸残基类型以及其侧链构象。不同位置的残基选择及其构象状态的可能组合构成了氨基酸序列和侧链构象空间。定义在该空间上的能量函数则被用于评估特定序列和构象组合的好坏。通过搜索算法在序列和侧链构象的未知量空间中自动搜索，找出能量尽可能低的解，得到设计结果。针对已有

的结构进行再设计，正确模拟突变后侧链构象至关重要，这一步通常使用“主链依赖的侧链旋转异构体库”(backbone-dependent rotamer library)，随后的侧链优化又依赖于力场与能量函数。

能量函数是对各序列组合的不同构象结构打分时的主要依据。不同软件使用的能量函数不尽相同，主要的能量项包括了物理能量项 (主要为非共价的范德华相互作用、静电能、氢键能、溶剂化自由能) 和统计能量项 (主要为主链二面角、侧链扭转)。目前国际上应用最广的能量函数包括 Baker 课题组开发的 Rosetta 能量函数 [27](以物理能量项为主) 和刘海燕课题组开发的 ABACUS 能量函数 [28](以统计能量项为主)。在固定主链的蛋白质设计中，共价键的键长键角一般设置为定值，主要考虑的相互作用是非共价的。在 Rosetta 能量函数中，使用 Lennard-Jones 势来计算范德华相互作用能量。使用最初来自 CHARMM 分子力场的原子电荷分布来计算静电能，并通过组优化进行了调整。使用静电模型和特殊的氢键模型来计算氢键的能量，并且该能量被细分为长距离主链氢键、短距离主链氢键、主链和侧链原子之间的氢键、侧链之间的氢键四个不同的类型分别计算。使用 Lazaridis-Karplus 隐式高斯排除模型，能包括各向同性和各向异性两种溶剂化自由能来刻画溶剂化效应。统计能量项是对数据库中出现的概率分布进行转化后得到的能量。一方面，从统计热力学角度来看，在平衡态，系统的不同微观状态的能量与概率服从玻尔兹曼分布；另一方面，从纯统计学角度出发，假设给定主链结构后氨基酸序列分布可记为条件概率，序列设计要解决的问题是寻找让该条件概率最大的序列。ABACUS 把不同的结构特征结合了起来，包括氨基酸所在位置的结构类型、主链二面角、溶剂可及性、残基间相对位置和统计得到的侧链旋转异构体和原子堆积能量。

搜索算法对于蛋白质序列设计同样是至关重要的，考虑到巨大的序列空间和更大的构象空间，遍历所有的构象组合实际上是不可能的。因此，Rosetta 被设计为一个采用蒙特卡洛方法的随机软件，通过对多次模拟产生的大量构象进行统计分析，然后给出数值解。Rosetta 首先利用随机数生成器生成随机的构象，随机微扰此构象后对新构象打分，接受所有打分变好的构象，以一定的概率接受打分变差的构象，直至在给定的循环次数内挑选出打分最好的结果。但是，这种迭代算法容易陷入局部最小值。为了得到全局能量最小的构象，除了借助分子动力学模拟的方法，Rosetta 还利用物理学中的动量概念 (想象一个小球从能量函数高处滚下，动量足够大时小球就不会被卡在小坑里而是会冲向最后的峡谷)，在迭代时不仅考虑这一次的能量变化，还兼顾上一次的能量变化。

除了基于物理学原理的算法，还有基于统计学和机器学习的算法。由于结构预测上 trRosetta 获得成功，Baker 课题组进一步开发了幻想 (Hallucination) 蛋白质从头设计方法 [29]。首先，给出一条随机的序列输入 trRosetta，预测其残基接触图；然后，对该序列附近的氨基酸序列空间利用蒙特卡洛方法采样，并计算序列之间的 KL 散度；最终，给出一个可以折叠的序列和预测的结构。Hallucination 借鉴了谷歌公司提出的“深梦”(DeepDream) 算法，该算法使用卷积神经网络，尽全力将输入改造为它曾在训练中见过的东西，产生了如梦幻般的幻觉外观。Hallucination 本质上是制造出与输入序列距离较近且符合 trRosetta 学习到的序列 - 结构关系的序列。因此，用该方法可以快速设计出与天然序列差距较大的蛋白质序列。

1.4 功能导向的蛋白质设计

根据蛋白质结构设计序列并不能直接解决合成生物学对新功能蛋白质的需求，从合成生物学的需求出发，蛋白质计算设计主要包括蛋白质自体骨架设计、蛋白质与大分子相互作用设计以及蛋白质与小分子相互作用设计。通过设计这些相互作用可以有效优化天然蛋白质作为合成生物学元件的功能，同时创造具有所需功能的生物传感器、生物催化剂和疫苗等。

蛋白质骨架设计主要用于提升天然蛋白质的鲁棒性或者设计额外的支撑骨架稳定疫苗的抗原表位，还可以改变蛋白质在特定条件下的稳定性。结合基于物理能量项、统计能量项的算法和生物信息学分析，吴边课题组开发 GRAPE 策略对 PET 塑料水解酶进行了计算重塑 [30]，通过融合单点预测算法并结合贪婪算法叠加单点突变，将最终突变体的热熔融温度提升了 31°C。为了开发新型冠状病毒抑制剂，在新冠病毒 S 蛋白与人血管紧张素转化酶 2 (ACE2) 复合物结构确定的基础上，Baker 课题组使用 ACE2 与 S 蛋白受体结合区域结合的螺旋片段为起点，尝试增加两股螺旋使之稳定；另外，在微蛋白库中使用蛋白质分子对接和蛋白质相互作用界面设计方法，最终设计出的小蛋白在皮摩尔浓度下即

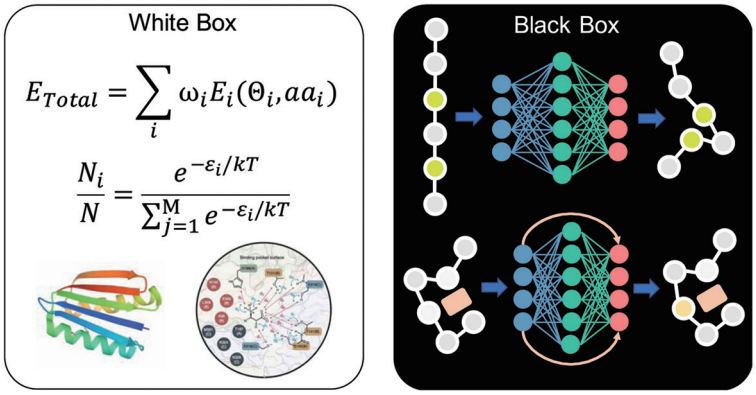
可对新冠病毒产生抑制作用[14]。瑞士 Correia 课题组开发了 TopoBuilder 系统来从头设计能够稳定复杂预定义结构单元的蛋白质[31]。针对不同的抗原表位，首先枚举二维空间上合适的蛋白质拓扑结构，并使用理想二级结构构建三级结构模型。利用此方法，他们设计出了可以同时呈递三种抗原的蛋白。膜蛋白在生命活动中具有重大意义，起到空间上传递信号交换物质的作用，西湖大学卢培龙课题组成功设计了多种不同的膜蛋白，实现了钾离子跨膜功能[32]。Rama 课题组使用直接耦合分析，提取 MSA 中隐含的序列 - 结构 - 功能空间的统计学约束，设计出了与天然酶活性相当的分支酸变位酶[33]。

设计蛋白质与大分子的相互作用可以用于合成细胞中的信号转导与调控。Baker 课题组通过计算设计了可以利用信号通路中天然存在的相互作用蛋白的生物传感器。在没有检测对象时，传感器的 lucCage 蛋白的锁扣结构域与笼结构域结合；有检测对象时，锁扣结构域的末端区域与检测对象结合，lucCage 蛋白打开并与传感器的 lucKey 蛋白结合，激活荧光素酶发出荧光[34]。Baker 课题组还设计了可调节蛋白质结合的逻辑门[12]，通过从头构建主链螺旋骨架，建立氢键网络进行序列优化，设计了多对可特异性二聚化的蛋白质，使用单体或连接的单体作为输入，并通过设计的氢键网络编码特异性结合，构建出能够接受不同输入的门控单元。借助于高通量的实验技术，Church 课题组利用机器学习算法设计了由 60 个单体组成的复杂球状蛋白质复合物 (AAV 衣壳蛋白)，并且发现即使经过有限的数据训练，深度神经网络模型也可以准确预测各种突变体中的衣壳活力[10]。

蛋白质与小分子的相互作用设计，可以用于获得新的酶催化元件、转录因子、小分子传感器等。美国 Kortemme 课题组参考了天然蛋白质结合法尼基焦磷酸 (FPP) 的结构，筛选了结合 FPP 的四残基结合模体，然后通过与大量骨架界面的对接和进一步的优化，设计了可被 FPP 调节的生物传感器[11]。设计酶的底物选择性可以产生新的生化反应，不仅可以设计新路径，也可以直接用于生物工业催化。但是，酶的活性中心具有一定的柔性且有复杂的氢键网络，细小的偏差都会导致设计的直接失败，吴边课题组使用固定主链设计的方法，结合多次平行的短时间动力学模拟弥补固定主链和侧链采样不均匀的缺陷，设计天冬氨酸裂解酶催化氢胺化反应，实现了非天然氨基酸的工业生产[35]。

2　蛋白质计算设计的挑战

为了解决蛋白质设计问题，基于蛋白质本身的物理化学原理，学界开发了以 Rosetta、ABAUCS 为代表的基础软件，发展出一系列的计算策略，已有经过实验验证的系列成功案例，展示了广阔的应用前景；同时，作为有明确意义的“白箱模型”，对验证学界对蛋白质折叠与序列选择原理的理解，具有深远的科学意义。由于体系的复杂性，科研人员尝试借助深度学习来解决蛋白质的计算设计，这一类缺乏明确物理意义和可解释性的“黑箱模型”(图 3) 也获得了以 AlphaFold 为代表的巨大成功。



“白箱”指代利用从头计算或基于统计的能量函数，利用搜索算法在能量分数和先验的化学知识的指导下进行的计算设计，这一过程的理论是充足的、形式是美观的。“黑箱”指代利用数据去训练难以解释明确物理意义的神经网络中的参数，进而预测该空间中其他数据点的映射。

图3　蛋白质计算的“白箱”与“黑箱”

但是，蛋白质计算设计仍然是一个新兴的前沿交叉领域，发展过程中仍存在诸多瓶颈，未来应用的道路仍然面临挑战。

在进行蛋白质计算设计时，主链结构一般会被假设为固定不变。如果主链结构也被作为未知量与序列、侧链同时被优化，尽管直觉上更为合理，但一方面对主链没有较好的离散表示，在计算层面上，变量空间维度会过高，使得计算无法完成；另一方面，力场的误差被进一步放大，甚至可能降低准确性。固定主链的主流蛋白质设计方法虽然取得了一定的成果，但不能掩盖这一权宜之计的不合理性。在部分酶活性中心的计算设计案例中，晶体结构分析表明实验所得的活性中心 loop 区域实际比预期设计结构有较大的变化，直接导致了设计的失败 [36]；而在突变引起的能量变化的计算中，完全主链柔性大大增加了计算量，不仅没有显著提高预测准确性，甚至在某些案例中还会导致准确性的下降 [37]。

在侧链放置与优化过程中，当前的大分子力场采用各向同性的小球模型来描述原子，忽略了电子云的实际状态，从而引入系统性的误差，限制了分子设计的成功率。同时，目前使用的隐式水模型难以捕捉由水分子传递的氢键相互作用。最终，使用人为规则从头设计的蛋白质往往看似非常理想，有较短的 loop 区域和完美的表面电荷，同时没有可能导致稳定性损失的空腔等，但是这些特征与自然界中真正执行功能的蛋白质差异较大 [38]。其中只有占比非常少的结构具有可延展设计性，能够作为合成生物学所需的功能生物大分子骨架。但在现实中，蛋白质作为生命活动的直接承担者并不具有物理学上完美的结构，相反，大部分蛋白质仅仅维持在稳定的边缘，只有约 5~10 kcal/mol 的富余能量 [39]。但是，漫长进化所获得的天然蛋白质能够高效精准且受控制地执行功能，这需要学界继续深入地去理解和思考结构与功能的关系。

利用深度学习这一类数据驱动的“黑箱模型”，面临标准化数据缺失和数据“偏见”问题。生物体系复杂且历史较长，数据表征方法不统一，缺乏大规模的专业性数据集。针对蛋白质结构进行计算预测，过去 50 年学术界积累了 18 万个实验测定的结构，由此构成的 PDB 数据库是一个相对标准的数据库，因此可以用来训练出 AlphaFold，即使如此依然使用了数据蒸馏等数据增强方法 [19]。另一方面，长期以来学术界不发表负面数据，这导致了蛋白质功能数据的不真实分布，产生错误的统计图景，因此难以用于大规模的模型训练。公开的蛋白质 (突变) 功能实验数据集在实验选择上具有偏好性，例如在稳定性单点突变数据集中，丙氨酸扫描的实验结果占了主导地位。大数据不是仅仅指数据量大，更是要求大量数据呈现某种形式的、可归纳的“特征”。

酶是合成生物学中最重要的元件之一，设计新反应才能创造新生化途径，进而创造出新生命。随着计算算力的提高，新酶设计已突破传统理性设计方法仅对结构进行微小扰动的桎梏，逐步迈向活性位点大尺度协同突变的计算重设计，乃至全局序列空间搜索的从头计算设计方法。但是，在很多案例中，新设计的酶活性极低，难以满足合成生物学需求，更无法进一步应用到工业生产中，需要后续借助定向进化等手段提升酶的活性。酶的催化依赖于活性中心氨基酸侧链的构象，需要对结构进行非常准确的建模，而既有力场存在蛋白质和小分子描述精度不足、难以刻画复杂相互作用等缺点，成为了制约新酶设计发展的主要因素。

最后，限制蛋白质计算设计成为合成生物重要的使能技术的重大限制还在于过高的门槛。现有蛋白质计算设计工作大量依赖于华盛顿大学 Baker 课题组主导开发的 Rosetta 软件包，软件学习成本极高，运行代码晦涩难懂，需要使用者拥有熟练的计算机编程能力和扎实的结构生物学知识才能够运行计算并合理解读结果。在解决具体的问题时，基础预测软件通常并不能直接输出预期结果，研究人员还需要根据所针对的生物学问题，将其合理翻译为计算问题，再进行设计范畴策划，开发特定的计算策略。例如酶的计算设计通常需要使用量子化学计算确定设计尺度，再使用蛋白质预测算法推算设计文库，最后使用分子动力学模拟进行虚拟筛选。而现有的学科培养体系难以为领域提供大量的交叉复合性研究人员。

3 蛋白质计算设计的未来发展方向

由于蛋白质科学对于认识生命的重要性和蛋白质作为疫苗、药物、催化剂等合成生物学元件的重要应用意义，可以预言，蛋白质计算设计是未来我国必须要抢占的科学和技术高地。针对目前蛋白质计算设计所需解决的瓶颈，未来一段时间主要有以下几个方面需要着重发展。

(1) 开发恰当描述主链运动和更加精确描述侧链构象的表示方法，解决目前固定主链这一不合理

假设带来的困难，和侧链构象离散描述带来的误差，补全采样时的空隙，进而提高序列设计的准确性。在计算设计框架里加入对引入功能可能需要承受的“结构不完美”的容忍，批量设计带有潜在的小分子结合口袋和蛋白质互作疏水区域的蛋白这一类能量上不完美，或者有多个不同折叠构象能量最小值的蛋白，进而设计出一个骨架上具有多个活性中心的酶和可控的变构蛋白。

(2) 提高能量函数的准确性和通用性。对于结构依赖型设计方法，提高对化学机制的解析能力和描述精度，发展能够计算多位点协同进化的设计策略；对于序列依赖型设计方法，改善二维空间向三维空间的投影能力，拓展库容较小数据集的训练能力。融合结构依赖型及非结构依赖型序列空间搜索策略的计算优势，增强蛋白质从头设计或重设计的迭代能力，搭建蛋白质结构预测和蛋白质计算设计的统一闭环框架。在提高对蛋白质的打分精确度的同时，发展非天然氨基酸力场及计算预测策略，将更多类型的非天然氨基酸引入蛋白质分子设计领域，并发展能够携带非天然氨基酸的高兼容性蛋白质合成方法，实现“非天然”催化反应设计，拓展现有的分子结构空间。解决目前小分子化合物、大分子的糖和核酸与蛋白质的相互作用打分问题，提高蛋白质与其他分子相互作用计算设计的准确率和效率，实现多个蛋白质单体组成的复杂分子机器设计。

(3) 构建高质量蛋白质标注数据集。在构建数据集时必须充分考虑蛋白质科学的特性，将生化实验或计算预测结果的置信度纳入考量。参考目前已有的蛋白质语言模型，进一步利用无功能标注的结构数据训练蛋白质结构的自监督语言模型。在人工智能本身的发展中，深度学习对数据标注也做出了巨大的贡献。在完成高精度的“黑箱模型”训练后，制造大量高精度的标注数据，进一步利用可解释的统计模型进行知识发现来把“黑箱”变成“白箱”。

(4) 推进蛋白质计算设计软件的国产化，摆脱长期以来对外国软件的依赖，构建自主可控的蛋白质计算设计平台。在新的平台中完善国产蛋白质设计软件中的力场、采样方法和对生物学问题到计算问题的模块化拆分，提高设计方法精准度，降低使用的难度和门槛，积极推进蛋白质计算设计技术在合成生物学领域的应用拓展。在基础合成生物学领域，开发作为感受器、逻辑门等调控元件的非天然蛋白质。在工业生物领域，利用新型生物催化反应改造和优化现有自然生物体系，从头创建合成可控、功能特定的人工生物体系。在医学应用领域，拓展抗体、疫苗以及药物蛋白的设计等。

[参 考 文 献]

[1] Brini E, Simmerling C, Dill K. Protein storytelling through physics. Science, 2020, 370: eaaz3041

[2] Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. Nature, 2016, 537: 320-7

[3] Regan L, DeGrado WF. Characterization of a helical protein designed from first principles. Science, 1988, 241: 976-8

[4] Dahiyat BI, Mayo SL. *De novo* protein design: fully automated sequence selection. Science, 1997, 278: 82-7

[5] Röthlisberger D, Khersonsky O, Wollacott AM, et al. Kemp elimination catalysts by computational enzyme design. Nature, 2008, 453: 190-5

[6] Siegel JB, Zanghellini A, Lovick HM, et al. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science, 2010, 329: 309-13

[7] Jiang L, Althoff EA, Clemente FR, et al. *De novo* computational design of retro-aldol enzymes. Science, 2008, 319: 1387-91

[8] Siegel JB, Smith AL, Poust S, et al. Computational protein design enables a novel one-carbon assimilation pathway. Proc Natl Acad Sci USA, 2015, 112: 3704-9

[9] Cai T, Sun H, Qiao J, et al. Cell-free chemoenzymatic starch synthesis from carbon dioxide. Science, 2021, 373: 1523-7

[10] Bryant DH, Bashir A, Sinai S, et al. Deep diversification of an AAV capsid protein by machine learning. Nat Biotechnol, 2021, 39: 691-6

[11] Glasgow AA, Huang YM, Mandell DJ, et al. Computational design of a modular protein sense-response system. Science, 2019, 366: 1024-8

[12] Chen Z, Kibler RD, Hunt A, et al. *De novo* design of protein logic gates. Science, 2020, 368: 78-84

[13] Vorobieva AA, White P, Liang B, et al. *De novo* design of transmembrane β barrels. Science, 2021, 371: eabc8182

[14] Cao L, Goreshnik I, Coventry B, et al. *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors. Science, 2020, 370: 426-31

[15] Anfinsen CB. Principles that govern the folding of protein chains. Science, 1973, 181: 223-30

[16] Yang J, Yan R, Roy A, et al. The I-TASSER Suite: protein structure and function prediction. Nat Methods, 2015, 12: 7-8

[17] Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706-10

[18] Yang J, Anishchenko I, Park H, et al. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA, 2020, 117: 1496-503

[19] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein

structure prediction with AlphaFold. Nature, 2021, 596: 583-9

[20] Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2021, doi: https://doi.org/10.1101/2021.10.04.463034

[21] Pearson WR. Protein function prediction: problems and pitfalls. Curr Protoc Bioinformatics, 2015, 51: 4.12.1-8

[22] Li Y, Wang S, Umarov R, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics, 2018, 34: 760-9

[23] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinform, 2009, 10: 168

[24] Gligorijević V, Renfrew PD, Kosciolek T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun, 2021, 12: 3168

[25] Sanderson T, Bileschi ML, Belanger D, et al. ProteInfer: deep networks for protein functional inference. bioRxiv, 2021, doi: https://doi.org/10.1101/2021.09.20.461077

[26] Richter F, Baker D. Chapter 6 - Computational protein design for synthetic biology [M]//Zhao H. Synthetic biology. Boston: Academic Press, 2013: 101-22

[27] Leman JK, Weitzner BD, Lewis SM, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods, 2020, 17: 665-80

[28] Xiong P, Wang M, Zhou X, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat Commun, 2014, 5: 5330

[29] Anishchenko I, Pellock SJ, Chidyausiku TM, et al. *De novo* protein design by deep network hallucination. Nature, 2021, 600: 547-52

[30] Cui Y, Chen Y, Liu X, et al. Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy. ACS Catalysis, 2021, 11: 1340-50

[31] Sesterhenn F, Yang C, Bonet J, et al. *De novo* protein design enables the precise induction of RSV-neutralizing antibodies. Science, 2020, 368: eaay5051

[32] Xu C, Lu P, Gamal El-Din TM, et al. Computational design of transmembrane pores. Nature, 2020, 585: 129-34

[33] Russ WP, Figliuzzi M, Stocker C, et al. An evolution-based model for designing chorismate mutase enzymes. Science, 2020, 369: 440-5

[34] Quijano-Rubio A, Yeh HW, Park J, et al. *De novo* design of modular and tunable protein biosensors. Nature, 2021, 591: 482-7

[35] Li R, Wijma HJ, Song L, et al. Computational redesign of enzymes for regio- and enantioselective hydroamination. Nat Chem Biol, 2018, 14: 664-70

[36] Huang PS, Boyken SE, Baker D. The coming of age of *de novo* protein design. Nature, 2016, 537: 320-7

[37] Nisthal A, Wang CY, Ary ML, et al. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. Proc Natl Acad Sci USA, 2019, 116: 16367-77

[38] Vriend G. Protein design: Quo Vadis? Science, 2004, 306: 1135

[39] Taverna DM, Goldstein RA. Why are proteins marginally stable? Proteins, 2002, 46: 105-9