序

DOI: 10.12211/2096-8280.2023-037

人工智能：开启生物体系计算设计的新篇章

刘海燕
（中国科学技术大学生命科学与医学部，安徽　合肥　230022）

中图分类号：Q81　　**文献标志码：**A

合成生物学基于对天然生命体系机制和规律认识，开发工程化的使能技术和工具，通过“设计-构造-测试”的闭环打造人工生物体系，实现生命科学研究和生物技术应用中的“建物致知”“建物致用”。从传统生物工程过渡到合成生物学，既包含了渐变性的演化，也包含了突变性的跃迁：“渐变性演化”体现在合成生物学在技术层面继承、集成生物分子、网络、细胞乃至有机体等不同层次的生物工程方法；而“突变性跃迁”的主要体现之一，是“设计”环节在合成生物学研究中的重要性显著增加。后者必然需要相关“设计”能力的大幅提升。

在传统生物工程中，设计问题主要通过利用特定体系的特点和专家的经验来解决，缺乏半定量、定量模型的支撑，难以形成系统的、能够方便地在不同应用场景或不同研究团队之间迁移推广的设计方法和设计技术。要提升合成生物学设计能力，需要针对不同层次的生物学问题发展基于计算的定量方法和模型；这类以计算为基础的模型较少依赖于研究者个人经验，可以迁移应用于不同场景，从而让我们能够以更趋近于现代工程学的方式来设计人工生物系统。

合成生物学中的设计问题面临不同的尺度。纵向来看：要对蛋白质等分子元件自身结构功能进行设计，核心模型可以只考虑单个分子；要设计分子识别和组装，需要使用考虑分子间复合物或多分子聚集体的模型；要优化设计细胞代谢网络、信号调控网络等，则需要能处理多节点分子互作网络的模型。横向来看：有的计算模型只适用于特定的靶标分子；有的模型则考虑了某一家族或具有某种特定功能的同类生物大分子（如催化某种类型化学反应的酶）；更通用的模型涵盖的分子类型则更广泛，如固有无序蛋白、非编码RNA等。

本专辑中，多篇文章讨论的计算模型属于分子或分子间层次。来自赵国屏团队的王晟等[1]聚焦合成生物学分子元件，从设计原理、计算方法、应用等角度，介绍了催化元件、调控元件、传感元件的计算设计前沿进展。本专辑中另外几篇论文则从不同角度综述了酶催化元件的计算设计进展。巫瑞波团队[2]长期从事酶反应机制的理论模拟；他们的综述聚焦于酶催化底物、产物的预测，以及酶设计改造。他们汇总比较了酶反应相关数据库、数据驱动的酶反应设计工具等，着重介绍了深度学习在该领域的发展和应用前景。洪亮团队[3]专长于人工智能与生命科学的交叉研究，他们的综述重点关注了应用于酶工程的人工智能方法。基于对酶工程的发展历程和现状的分析，他们综述了可被用于预测有益突变、优化蛋白质稳定性、提高催化活性等的深度学习方法进展。孟巧珍和郭菲[4]则以AlphaFold2为例，对把蛋白质结构预测方法作为结构“分析器”、突变“筛选器”或者折叠“监督器”应用于酶智能设计进行了总结。

生物体系中最重要的一类分子元件是蛋白质。可靠的蛋白质功能预测方法对合成生物学元件挖掘具有重要意义。杨跃东团队[5]长期从事疾病机制阐明和药物靶点发现等领域的蛋白质功能预测。他们综述

收稿日期：2023-05-31 修回日期：2023-06-02
引用本文：刘海燕. 人工智能:开启生物体系计算设计的新篇章[J]. 合成生物学, 2023, 4(3): 419-421

了残基水平的结合位点预测和蛋白水平的基因本体论（gene ontology）预测等蛋白质功能预测的最新方法，比较了不同方法的优劣并展望了未来可能的发展方向。

蛋白质功能往往建立在三维结构基础之上。戚逸飞等[6]的综述侧重介绍了蛋白质结构设计的人工智能算法。他们从固定骨架设计、可变骨架设计和序列结构生成三个方面总结了最新算法进展。可以预期，单体蛋白质结构设计问题基本解决以后，具有形成特异性复合物等功能的蛋白的设计将成为方法研究的重点。

本专辑中另外两篇聚焦蛋白质结构计算的综述都是关于分子间复合物的。环肽用作蛋白-蛋白互作的调控分子具有独特优势。王凡灏、来鲁华和张长胜[7]的综述分析了环肽与蛋白结合的结构数据，介绍了基于分子对接的虚拟筛选、借助于动力学模拟的设计、从头生成设计以及跨膜环肽设计等环肽计算方法，展望了人工智能在环肽设计中的应用前景。相对单体蛋白结构预测，目前对蛋白质复合物的结构预测精度仍然不高，在算法方面有较大的进步空间。龚新奇团队[8]长期从事该方向的研究。他们的综述侧重于总结蛋白质复合物结构预测的相关算法以及介绍最新进展。

除通过稳定的三维结构形成分子间复合物外，细胞内还有大量固有无序蛋白或蛋白固有无序区。它们可以通过由多价分子间互作介导的液-液相分离来调控生物功能。无序蛋白聚集失调被认为是引发神经退行性疾病等的可能机制。韦广红团队[9]长期开展基于多尺度分子力场等物理模型的无序蛋白聚集机制研究。他们的综述重点介绍了神经退行性疾病相关蛋白聚集和液-液相分离的方法和前沿进展。他们还讨论了相关微观机理的理论和计算研究结果，以及预测蛋白相分离能力的机器学习方法。

除了以上关于分子和分子间层次问题的计算模型外，本专辑另外两篇综述则分别关注通路层次和网络层次的问题。生物合成基因簇包含了特定天然产物合成的完整通路，是合成生物学极具潜力的元件来源。宁康团队[10]在他们的综述中讨论了基于微生物组数据发现新生物合成基因簇的问题。他们总结了相关数据资源和挖掘方法，特别是人工智能方法，展示了新发掘的生物合成基因簇的多样性和应用潜力。汤超、杨晓静等[11]则指出，完整的生物功能依赖于能执行各种各样复杂功能、高精度、可靠、鲁棒的分子网络，发现网络的拓扑结构、动力学与功能之间关系，找到生物网络的底层设计规律是系统生物学和合成生物学的巨大挑战。他们归纳了天然网络中的拓扑-功能关系，介绍了系统生物学的相关理论成果，进而总结了近年来合成生物学功能网络拓扑设计的研究进展。

综上，可用于合成生物学设计问题的计算生物学模型纷繁多样，难以在简短篇幅内逐一介绍。为了概括不同模型的原理，我们可以考虑根据建立计算模型的主要依据类型，对不同计算生物学模型进行粗略分类。计算模型建立的依据可以包括物理原理（基于物理原理的模型）、假设或经验规则（基于规则的模型）、实验数据（数据驱动的模型）等。以蛋白质结构预测、设计为例：描述分子能量与分子结构依赖关系的分子力场属于典型的基于物理原理的经验模型；用深度学习预测蛋白质结构的AlphaFold2则是典型的数据驱动模型。对于生物体系，目前基于物理原理或基于规则的模型类型相对比较有限，而数据驱动模型类型最多。数据驱动模型覆盖的问题范围也十分广泛，如前述综述中提到的数据驱动的元件设计优化、基于组学数据的分子元件发现和结构功能预测等。

作为目前最前沿的数据驱动建模技术，人工智能（artificial intelligence或AI）在生物计算中的应用非常广泛。尽管目前对“人工智能“一词所涵盖技术的范围并没有公认的明确定义，美国食品药品管理局2023年5月发布的关于人工智能/机器学习与药物开发的讨论文件中的定义可作为有价值的参考。在该文件中，人工智能被定义为“用算法或模型来执行任务并表现出如学习、做出决策、做出预测等行为的一个计算机科学、统计学、工程学的分支”。同时，“机器学习”（machine learning或ML）被定义为“人工智能的一个子集”，“用数据和算法不通过显式编程地去模拟人类怎样学习”。进一步地，深度学习（deep learning或DL）被归为人工智能/机器学习的子领域。无可争议，深度学习是最近十余年人工智能领域最重大的突破。正如我们从本专辑中多篇关于蛋白质结构预测和蛋白质设计问题的综述所看到的，在数据充分、算法恰当的情况下，最新的深度学习技术能够以前有未有的方式提升我们对复杂生物大分子序列、

结构、功能的预测和设计能力。元件层次预测、设计能力的提升将会很快被传递应用于对网络、细胞的设计。与此同时，在计算机和信息科学领域内部，人工智能技术本身仍在快速迭代发展之中。我们预期，人工智能技术与计算生物学方法以及合成生物学应用问题的融合将越来越广泛、越来越紧密，从而不仅在分子元件层次，还会在网络、细胞等层次带来算法能力的大幅提升。可以说，与深度学习等人工智能技术的结合，正在开启生物计算设计的新篇章。

参 考 文 献

[1] 王晟，王泽琛，陈威华，等. 基于人工智能和计算生物学的合成生物学元件设计[J]. 合成生物学，2023，4(3)：422-443.
WANG Sheng, WANG Zechen, CHEN Weihua,et al. Design of synthetic biology components based on artificial intelligence and computational biology[J]. Synthetic Biology Journal,2023,4(3):422-443.

[2] 曾涛，巫瑞波. 数据驱动的酶反应预测与设计[J]. 合成生物学，2023，4(3)：535-550.
ZENG Tao, WU Ruibo. Data-driven prediction and design for enzymatic reactions[J]. Synthetic Biology Journal,2023,4(3):535-550.

[3] 康里奇，谈攀，洪亮. 人工智能时代下的酶工程[J]. 合成生物学，2023，4(3)：524-534.
KANG Liqi, TAN Pan, HONG Liang. Enzyme engineering in the age of artificial intelligence[J]. Synthetic Biology Journal,2023,4(3):524-534.

[4] 孟巧珍，郭菲. “可折叠性”在酶智能设计改造中的应用研究——以AlphaFold2为例[J]. 合成生物学，2023，4(3)：571-589.
MENG Qiaozhen, GUO Fei. Applications of foldability in intelligent enzyme engineering and design: take AlphaFold2 for an example[J]. Synthetic Biology Journal,2023,4(3):571-589.

[5] 宋益东，袁乾沐，杨跃东. 深度学习在蛋白质功能预测中的应用[J]. 合成生物学，2023，4(3)：488-506.
SONG Yidong, YUAN Qianmu, YANG Yuedong. Application of deep learning in protein function prediction[J]. Synthetic Biology Journal, 2023,4(3): 488-506.

[6] 陈志航，季梦麟，戚逸飞. 人工智能蛋白质结构设计算法研究进展[J]. 合成生物学，2023，4(3)：464-487.
CHEN Zhihang, JI Menglin, QI Yifei. Research progress of artificial intelligence in desiging protein structures[J]. Synthetic Biology Journal, 2023,4(3):464-487.

[7] 王凡灏，来鲁华，张长胜. 基于靶标结构的环肽分子计算设计[J]. 合成生物学，2023，4(3)：551-570.
WANG Fanhao, LAI Luhua, ZHANG Changsheng. Target structure based computational design of cyclic peptides[J]. Synthetic Biology Journal, 2023,4(3):551-570.

[8] 黄鹤，吴桐，王闻达，等. 蛋白质复合物结构预测：方法与进展[J]. 合成生物学，2023，4(3)：507-523.
HUANG He, WU Tong, WANG Wenda, et al. Prediction of protein complex structure: methods and progress[J]. Synthetic Biology Journal, 2023,4(3):507-523.

[9] 唐一鸣，姚逸飞，杨中元，等. 神经退行性疾病相关蛋白病理性聚集和液液相分离研究进展[J]. 合成生物学，2023，4(3)：590-610.
TANG Yiming, YAO Yifei, YANG Zhongyuan,et al. Pathological aggregation and liquid-liquid phase separation of proteins associated with neurodegenerative diseases[J]. Synthetic Biology Journal,2023,4(3):590-610.

[10] 赖奇龙，姚帅，查毓国，等. 微生物组生物合成基因簇发掘方法及应用前景[J]. 合成生物学，2023，4(3)：611-627.
LAI Qilong, YAO Shuai, ZHA Yuguo, et al. Microbiome-based biosynthetic gene cluster data mining techniques and application potentials[J]. Synthetic Biology Journal,2023,4(3):611-627.

[11] 孙智，杨宁，娄春波，等. 功能拓扑的理性设计及其在合成生物学中的应用[J]. 合成生物学，2023，4(3)：444-463.
SUN Zhi, YANG Ning, LOU Chunbo, et al. Rational design for functional topology and its applications in synthetic biology[J]. Synthetic Biology Journal,2023,4(3):444-463.

刘海燕（1969—），男，中国科学技术大学讲席教授，国家杰出青年基金获得者，国家自然科学基金创新群体项目负责人。1985—1996年在中国科学技术大学生物系学习并先后获学士和博士学位，曾在苏黎世高等理工学院进行研究生联合培养、在Duke大学和UNC Chapel Hill从事博士后研究等。主要研究方向为蛋白质设计、基于计算的蛋白质结构功能研究，在*Nature*及子刊、*JACS*、*PRL*等杂志发表论文一百多篇。
E-mail：hyliu@ustc.edu.cn