观点

DOI: 10.12211/2096-8280.2026-007

人工智能驱动合成生物学研究

兰晶岗[1]，傅雄飞[2]，汪小我[3]，张先恩[1,4]
（[1] 深圳理工大学合成生物学院，广东 深圳 518107；[2] 中国科学院深圳先进技术研究院，定量合成生物学全国重点实验室，深圳合成生物学创新研究院，广东 深圳 518055；[3] 清华大学自动化系，合成与系统生物学研究中心，教育部生物信息学重点实验室，北京信息科学与技术国家研究中心，北京 100084；[4] 中国科学院生物物理研究所，生物大分子全国重点实验室，北京 100101）

摘要：人工智能（AI）正在深刻重塑合成生物学的研究范式，使生命系统的设计从经验驱动迈向模型驱动。传统合成生物学依赖突变筛选与试错式优化，难以应对多尺度、高维度、强耦合的生命过程。随着组学数据的爆发式增长、自动化实验平台的普及以及深度学习技术的快速发展，AI 为揭示序列-结构-功能规律、构建可预测生物模型和实现大规模生命设计提供了全新路径。AI 合成生物学已在四个关键层级形成系统化框架：在生物大分子层面，蛋白语言模型与生成式结构模型使从头设计酶、受体与自组装材料成为可能；在基因组层面，深度学习推动突变机制建模、大片段序列生成与谱系动力学推断，为可编程基因组构建奠定基础；在细胞层面，AI 与机理模型结合加速虚拟细胞构建，使细胞行为实现可量化预测；在平台层面，多智能体与自动化“设计-建造-测试-学习”（DBTL）循环支持路径规划、酶功能预测与实验调度的全流程自动化。总体而言，AI 使合成生物学从局部优化走向系统生成，从经验探索走向预测设计，为生命系统的可控重构和生物制造创新提供了核心动力。

关键词：人工智能；合成生物学；生物制造；生物大分子设计；基因组设计；细胞设计

中图分类号：Q81　　**文献标志码：**A

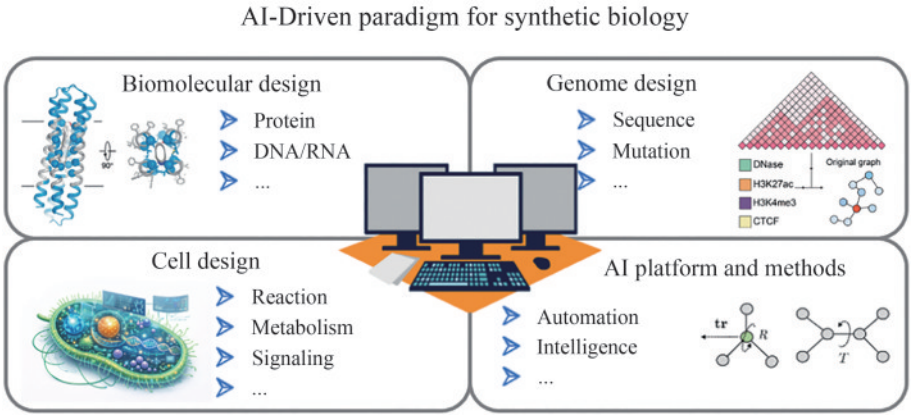
AI + synthetic biology：a paradigm shift in biomanufacturing

LAN Jinggang[1]，FU Xiongfei[2]，WANG Xiaowo[3]，ZHANG Xian-En[1,4]
（[1]*Faculty of Synthetic Biology，Shenzhen University of Advanced Technology，Shenzhen 518107，Guangdong，China*；[2]*State Key Laboratory for Quantitative Synthetic Biology，Shenzhen Institute of Synthetic Biology，Shenzhen Institutes of Advanced Technology，Chinese Academy of Sciences，Shenzhen 518055，Guangdong，China*；[3]*Department of Automation，Tsinghua University，Center for Synthetic and Systems Biology，Ministry of Education Key Laboratory of Bioinformatics，Beijing National Research Center for Information Science and Technology，Beijing 100084，China*；[4]*National Laboratory of Biomacromolecules，Institute of Biophysics，Chinese Academy of Sciences，Beijing 100101，China*）

Abstract: Artificial intelligence (AI) is profoundly reshaping the research paradigm of synthetic biology, shifting the design of living systems from empirically driven approaches to model-driven ones. Traditional synthetic biology relies

收稿日期：2026-02-25 修回日期：2026-04-21
引用本文：兰晶岗，傅雄飞，汪小我，张先恩. 人工智能驱动合成生物学研究[J]. 合成生物学，2026，7(2)：279-292
Citation：LAN Jinggang，FU Xiongfei，WANG Xiaowo，ZHANG Xian-En. AI + synthetic biology：a paradigm shift in biomanufacturing[J]. Synthetic Biology Journal，2026，7(2)：279-292

on screening mutants for trial-and-error optimization, making it difficult to address multiscale, high-dimensional, and strongly coupled biological processes. With the explosive growth of omics data, the widespread adoption of automated experimental platforms, and the rapid development of deep learning technologies, AI provides a new pathway to uncover sequence-structure-function relationships, build predictive biological models, and enable large-scale design of living systems. So far, AI-driven synthetic biology has established a systematic framework at four levels: the biomacromolecular level with protein language models and generative structural models to make *de novo* design of enzymes, receptors, and self-assembling materials possible; the genomic level with deep learning to advance the modeling of mutational mechanisms, large-fragment sequence generation, and inference of phylogenetic dynamics, laying foundation for programmable genome construction; the cellular level with the integration of AI with mechanistic models to accelerate virtual cell development, enabling quantitatively predictive descriptions of cellular behavior; the platform level with multi-agent systems and automated "design-build-test-learn" (DBTL) cycles to support the end-to-end automation of pathway planning, enzyme function prediction, and experimental scheduling. Overall, AI is revolutionizing synthetic biology from local optimization to system-level generation, and from empirical exploration to predictive design as well, providing a core driving force for the controllable reprogramming of living systems and innovation on biomanufacturing.



Keywords: artificial intelligence; synthetic biology; biomanufacturing; biomacromolecular design; genome design; cell design

21世纪以来，合成生物学秉持"造物致知、造物致用"的理念，其使能技术已拓展至从生物大分子、基因组、细胞到器官等多层次系统的设计、编辑与合成。这一领域不仅革新了生命科学的研究范式，也驱动了生物技术的快速迭代与生物制造的深刻变革，被广泛视为引领未来生物经济发展的关键驱动力[1-2]。然而，生物体系本质上是具有时空动态特征、多网络互作的高度复杂系统，其理性设计与功能预测仍受困于"维数灾难"这一根本性挑战。

2024年，诺贝尔物理学奖、化学奖分别授予了人工智能（AI）的底层原理技术和在蛋白质结构预测及从头设计中的应用，正式拉开了一场将席卷科学界乃至人类文明的人工智能革命的序幕。AI，尤其是深度学习，逐渐成为继实验、理论和模拟之后的"第四科学范式"。2024年诺贝尔化学奖的两个代表性模型分别是AlphaFold和Rosetta。AlphaFold系列的核心突破是实现了高精度的蛋白质结构预测[3]，Rosetta则侧重基于物理原理的蛋白质建模与理性设计[4]。二者为合成生物学赋能，前者提供了前所未有的结构数据库，后者提供了定制化设计工具，共同夯实了"设计-构建-测试-学习"循环的基础。这一融合方向亦在2024年发布的《合成生物学路线图2030》"理论框架"中被

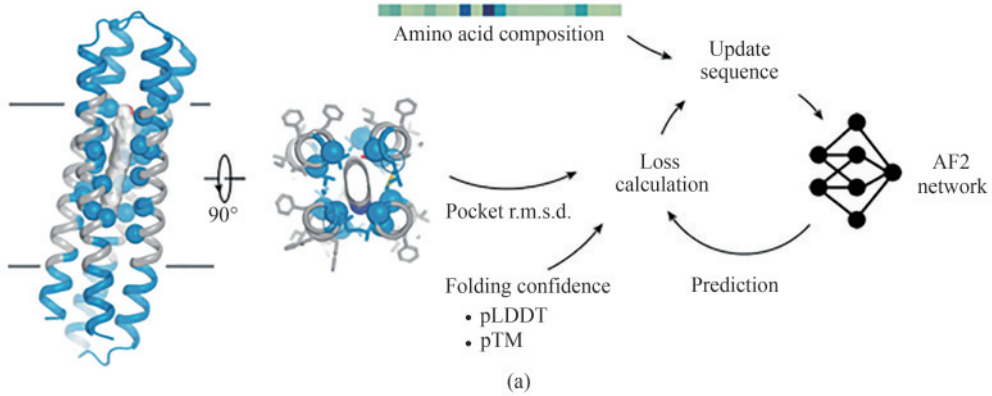
重点提及，系统阐述了合成生物学多尺度理论框架及其与人工智能的深度融合[5-6]。

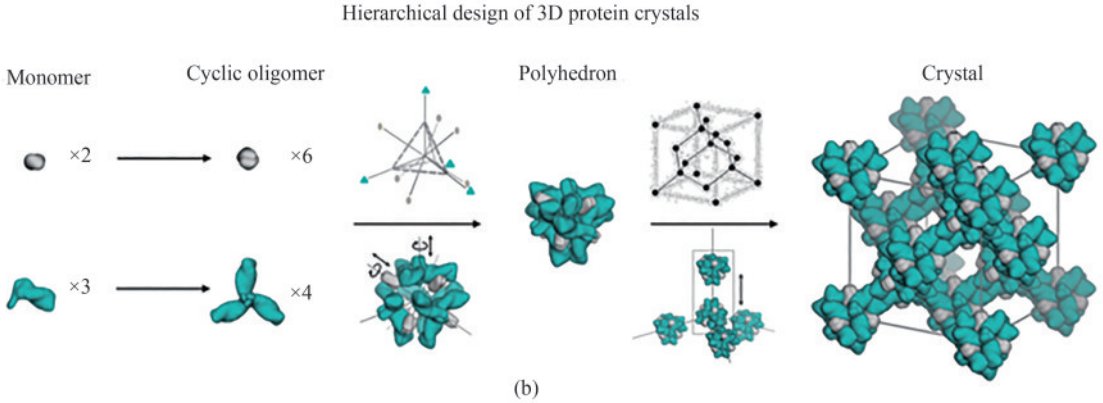
在这一背景下，2025年11月，首届人工智能合成生物学研讨会（AI-Synbio Workshop 2025）成功举办。会议不设大会报告，以圆桌会议形式，围绕AI与生物大分子设计、AI与基因组设计、AI与细胞设计、AI与合成生物学平台等四个方向展开深入讨论。与会专家认为，随着人工智能的持续推进，这四个层级的能力正在迭代提升并加速汇聚，形成统一的生命设计体系，使合成生物学真正迈向“可预测、可设计、可验证”的智能化生命工程时代。

1 生物大分子设计：从序列-结构-功能规律到智能化生成的体系化进展

生物大分子设计是合成生物学迈向工程化、可预测与可编程体系的关键科学问题之一。作为生命活动的核心执行元件，蛋白质、核酸及其组装体系承担着生命过程中的诸多关键功能，其结构-动力学-功能关系构成理解生命规律与人工重构生命功能的理论基础。传统蛋白质工程主要依赖定向进化、突变库筛选以及理性设计等策略，虽然在分子酶催化、抗体优化和材料蛋白等领域获得重要进展，但仍面临探索效率偏低、设计空间受限和功能可预测性不足等瓶颈。深度学习、分子模拟与自动化实验技术的融合，推动生物大分子设计从经验主导逐步转向智能算法驱动的系统化生成。其重点是在高维序列-结构-功能空间中建立可泛化的因果关联，突破传统试错模式的内在局限。

跨膜蛋白设计代表了从静态结构优化向动态信号响应体系构建的重要跨越。跨膜蛋白不仅承担细胞内外物质、能量和信号传递的功能，也是药物靶标与生物传感器领域的核心组件。在从头设计领域，过去的研究主要集中于稳定折叠结构的设计，而面向生物工程需求的设计任务要求蛋白能够对物理或化学信号产生可控响应[7-8]。西湖大学卢培龙团队[9]通过精准调控跨膜螺旋的电荷分布与相互作用界面，构建具备电压门控行为的人工离子通道，通过嵌入可识别小分子的口袋，实现了对特定化学激活信号的可编程荧光响应[图1(a)]。HBC599蛋白展示了人工跨膜蛋白在特





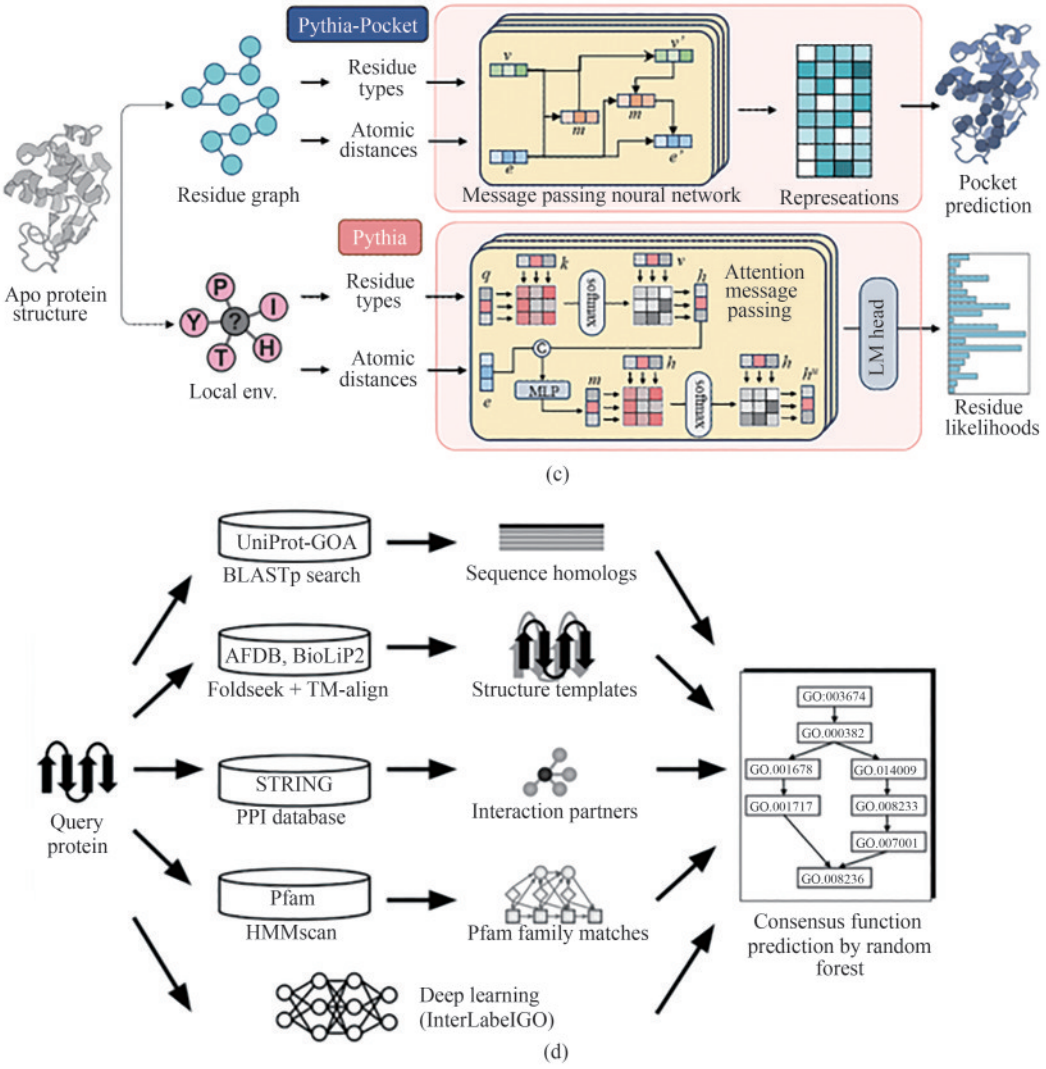


图1 生物大分子设计中的人工智能方法[9-12]

（a） 基于幻觉的 tmFAP 跨膜区设计流程示意图：对面向膜的残基进行序列重设计，同时固定配体结合口袋的主链坐标，其余残基序列保持不变；在迭代过程中，序列不断输入 AF2 进行结构预测，并依据预测结构、pTM 等指标计算损失以指导序列更新 [9]。（b） 分级晶体设计策略示意图：通过“单体 → 循环寡聚体 → 两组分笼状结构 → 三维晶格”的三级设计层级，构建由 C_2 二聚体与 C_3 三聚体组成的四面体单元并形成金刚石型晶格；不同层级的界面依次驱动寡聚体组装、笼体形成及晶体装配，界面埋藏面积与结合能逐级降低，以提高整体组装协同性 [10]。（c） Pythia-Pocket 与 Pythia 模型结构：两种模型均将 apo 蛋白结构抽象为图结构；Pythia-Pocket 以全蛋白图作为输入，通过 MPNN 生成残基表示并预测各残基属于配体结合口袋的概率，而 Pythia 以局部结构环境为输入，利用基于注意力的消息传递更新残基相互作用，最终预测中心残基的氨基酸类型概率分布 [11]。（d） StarFunc 方法流程图：StarFunc 通过整合五类功能预测模块（基于序列同源性、结构模板、蛋白－蛋白相互作用伙伴、Pfam 家族匹配以及深度学习模型）对蛋白质功能进行综合推断 [12]

Fig. 1 Artificial intelligence methods for designing biomolecules[9-12]

(a) Schematic illustration of the hallucination-based design workflow for the transmembrane region of tmFAP: membrane-facing residues are redesigned at the sequence level while the backbone of the ligand-binding pocket is fixed, and the sequences of the remaining residues are also kept unchanged. During the iterative process, the sequence is repeatedly fed into AF2 for structure prediction, and loss functions are calculated based on the predicted structure, pTM, and related metrics to guide the sequence optimization [9]. (b) Schematic illustration of a hierarchical crystal design strategy: a three-level design hierarchy, from monomers to cyclic oligomers, two-component cage-like assemblies, and finally to three-dimensional lattices, is used to construct tetrahedral units composed of C_2 dimers and C_3 trimers and to form a diamond-like lattice. Interfaces at different levels sequentially drive the oligomer assembly, cage formation, and crystal packing, with progressively reduced buried interface area and binding energy to enhance a cooperative assembly [10]. (c) Architecture of the Pythia-Pocket and Pythia models: both models visualize apo protein structures as graphs. Pythia-Pocket takes the whole-protein graph as input, uses MPNN to generate residue representations, and predicts the probability that each residue belongs to a ligand-binding pocket. Pythia takes the local structural environment as input, updates residue interactions through attention-based message passing, and ultimately predicts the amino acid distribution probability of the central residue [11]. (d) Workflow of the StarFunc method: StarFunc performs integrated protein function inference by combining five functional prediction modules, including sequence homology, structural templates, protein-protein interaction partners, Pfam family matching, and deep learning models [12]

定小分子激活条件下触发构象重排并输出荧光信号的可行性。这类设计表明，跨膜蛋白工程正从“结构稳定性为先”向“行为可控性优先”的范式转型，这对于人工信号传感、药物传递以及细胞级信息处理都具有重要应用价值。

相比于跨膜蛋白的动态功能化设计，可自组装蛋白材料的构建则更侧重于空间结构的可编程性。自然界中大量蛋白质结构（如病毒衣壳、细胞骨架和多酶复合体）展现出精细几何规则性与高度对称性，这启发研究者开发人工蛋白晶体、自组装笼状体及多尺度材料。自组装过程的核心在于蛋白单体的刚性设计以及界面相互作用的精细调控。近期的研究通过构建具有正四面体或八面体排列的蛋白笼，展示了从单体设计到周期性晶格生成的可行路径[13]。此外，引入金属有机框架（metal organic framework， MOF），南方科技大学李喆团队[10]将蛋白界面拆分为不同强度的相互作用模块，可分别调控晶体成核和晶体生长过程。结合高熵合金的“成分多样性-结构稳定性增强”机制，多种蛋白构件可共同组装为高熵蛋白晶体，极大拓展了材料空间［图1(b)］。这一类研究表明，蛋白材料的设计正在形成一套基于“几何基元-相互作用语法-自组装动力学”的工程规则体系，为生物材料科学开辟新的方向。

功能蛋白质设计领域的技术演进具有显著的阶段性特点。从早期基于PCR技术与突变筛选的局部探索，到定向进化框架下的系统优化，再到当前由人工智能驱动的序列空间全局搜索，功能蛋白设计的能力在持续提升。尽管AlphaFold等模型在结构预测上取得突破，但功能性蛋白质的设计仍然停留在初级阶段，其难点在于：功能往往源自动态构象集合而非单一静态结构，因此构效关系在高维空间中呈现复杂的非线性特征。当前的大模型方法，如蛋白语言模型开始展现跨越局部能量景观、直接预测功能相关特征的潜力。国内的全蛋白质骨架与序列设计系统的代表性工作包括中国科学技术大学刘海燕团队[14-15]的SCUBA2022在蛋白质设计上的重要进展，以及北京大学姜长涛与来鲁华团队[16]基于AI的代谢酶大规模发现。此外，西湖大学曹龙兴团队[17]通过从头设计可诱导聚集的蛋白多聚体体系展示了生成式功能蛋白质设计的可拓展性。在产业转化层面，新一代工业酶设计工具如北京化工大学吴边团队[18]开发的Pythia平台已显著提升工业酶活性与稳定性的工程化能力，并在免疫激动剂的合成元件计算[19]与塑料降解酶设计[11]中取得重要突破［图1(c)］。

蛋白质功能注释、免疫受体工程和酶催化机理分析等研究方向进一步拓展了大分子设计的应用边界。在大规模序列爆炸时代，功能注释工具如中国科学院深圳先进技术研究院的张成辛团队[12]开发的StarFunc可通过结合模板建模与语言模型表征［图1(d)］，实现低样本条件下的高精度功能预测，尤其适用于微生物和病毒蛋白序列的快速分析。在免疫治疗方向，对TCR–抗原识别的研究已从传统解离常数指标拓展至能量景观与构象动力学层面的分析，分子动力学结合机器学习的混合策略正在提升表位筛选与亲和力重构的效率。在酶催化方向，基于QM/MM、Cluster模型的物理机理解析仍然是功能设计不可或缺的核心手段，而人工智能可用于加速反应路径搜索、识别关键态构象并预测非天然底物的可行反应路径，形成“物理-数据”混合机制的酶设计新模式。

在蛋白质设计方面，人工智能已在结构预测与序列生成等任务中取得显著进展，展示了“结构-序列协同建模”的巨大潜力。在此基础上，相关方法正逐步拓展至RNA等其他生物大分子的设计研究中。相比蛋白质，RNA分子具有更高的构象灵活性，其设计过程不仅需要满足三维空间结构的稳定性，还需兼顾核苷酸序列与结构之间的匹配关系。然而，现有方法往往难以对结构与序列进行协同优化，导致生成序列与目标结构之间存在偏差。在数据层面，高质量的RNA–小分子相互作用数据相对匮乏，限制了模型的有效训练。同时，当前多数设计策略对靶分子的空间形状利用不足，难以实现面向特定分子的定向RNA生成。针对上述问题，上海交通大学溥渊未来技术学院郑双佳团队[20]提出了RiboFlow模型。该方法引入协同流匹配（synergistic flow matching）策略，实现了RNA三维结构与序列的联合建模，从而在生成过程中保持二者的高度一致性，并提升与目标分子的结合能力。此外，研究团队构建并发布了

目前规模较大的RNA-小分子相互作用数据集RiboBind，在一定程度上缓解了数据不足的瓶颈。实验结果表明，RiboFlow在结构合理性及靶向结合性能方面均优于现有方法，为RNA药物与功能分子的理性设计提供了新的技术路径。

在生物大分子设计层级，与会专家在关于序列-结构-功能关系的圆桌讨论中，提出了两个相对但具有互补性的观点。一方面，结构提供了理解功能机制的必要物理基础，是功能预测的关键中介变量；另一方面，蛋白语言模型的成功表明，序列中蕴含的进化统计规律可以直接映射到功能空间，使得“结构是否必须显式建模”成为新的科学议题。从机制角度看，功能必然高度依赖三维结构的存在，但在表征学习层面，结构可能被隐式编码于序列统计中。因此，未来蛋白质设计的趋势很可能是结合物理模型的因果性与语言模型的统计性，形成跨尺度、跨表征的混合设计框架。

总体而言，生物大分子设计的发展趋势体现为：设计目标从稳定结构设计扩展到可控功能生成，从局部序列修改转向全局序列空间探索，从单尺度建模迈向多尺度动力学表征，从经验驱动的改造走向智能驱动的生成。自动化实验、计算模拟与AI模型的结合，使分子元件层面的生命系统工程逐步具备可预测、可设计、可验证的能力。

2 基因组设计：基于深度学习的序列生成、功能重写与细胞演化建模

如果说生物大分子设计聚焦于功能元件，那么基因组设计则旨在从整体遗传信息层面对生命系统进行可预测、可编程与可验证的重构，是推动合成生物学向系统级工程迈进的核心支撑。相较于生物大分子设计侧重局部结构与分子尺度功能，基因组设计直接面向序列层级的复杂调控逻辑、跨尺度信息流动和动态细胞命运过程，因此其设计难度显著更高，传统经验驱动的基因编辑和元件组装策略难以满足大规模、精准化与功能导向的工程需求。随着深度学习、基因组大模型和单细胞动态测量技术的发展，基因组设计正在从线性编辑范式转向“规则学习-序列生成-动态预测”的系统化框架，该框架包括三个典型组成部分——突变机制的高精度建模、大片段序列的从头生成、细胞谱系与演化动力学的定量重构，三者共同构成了未来基因组可工程化的基础理论与工具体系。

基因组突变是进化的重要驱动力，也是许多疾病、生物多样性与遗传结构形成的基础。然而，长期以来，由于自然突变事件极其稀疏，传统基于统计模型的突变率估计只能捕捉粗略趋势，难以解析突变在基因组不同区域之间的显著差异性。中山大学李彩团队[21]通过构建深度学习模型MuRaL［图2（a）］，将突变率预测转化为“序列到突变倾向”的局部学习任务。MuRaL以突变位点上下游序列作为输入，不依赖大规模突变观测数据即可生成高分辨率突变率图谱，显著优于经典模型。更为关键的是，该模型学习到的突变机制具有跨物种与跨背景的泛化能力，使得其在稀疏数据条件下仍能保持稳定的预测性能。突变率图谱不仅能用于识别高突变区域，也能定位突变不耐受的功能关键片段，从而为疾病基因筛查、调控元件鉴定与基因线路稳健性评估提供重要依据。随着深度学习方法不断进化，突变模型正在为基因组设计提供一种从“随机演化”到“可预测演化”的理论基础，使得序列功能与演化约束能够以更定量的方式纳入设计框架。

在掌握突变规律的基础上，基因组工程的另一挑战是实现大片段乃至整条基因组的从头设计。随着各类组学数据在不同物种间不断积累，如何在复杂、异质且高维的跨尺度数据中识别功能规律，成为大模型介入基因组设计的突破口。聚焦于“有限数据条件下的大片段生成”，陈河兵等[22]提出通过组学信息整合与深度序列生成模型结合的策略，构建跨物种、跨尺度的统一表征空间［图2（b）］。在这一体系中，大模型不仅学习局部序列语法（如密码子偏好、调控基序），还同时理解更高层级的系统特性（如表达稳态、代谢负担、调控模块化结构）。通过整合转录组、代谢组和调控网络数据，模型能够生成满足生物物理可实现性与系统功能要求的长序列片段。更重要的是，该研究建立了高通量验证体系，使得数千条候选序列可以在实验条件下快速筛选并反馈至模型中，实现“生成-筛选-再生成”的闭环优化。这标志着基因线路设计已从人工组合模块的局部试错，迈

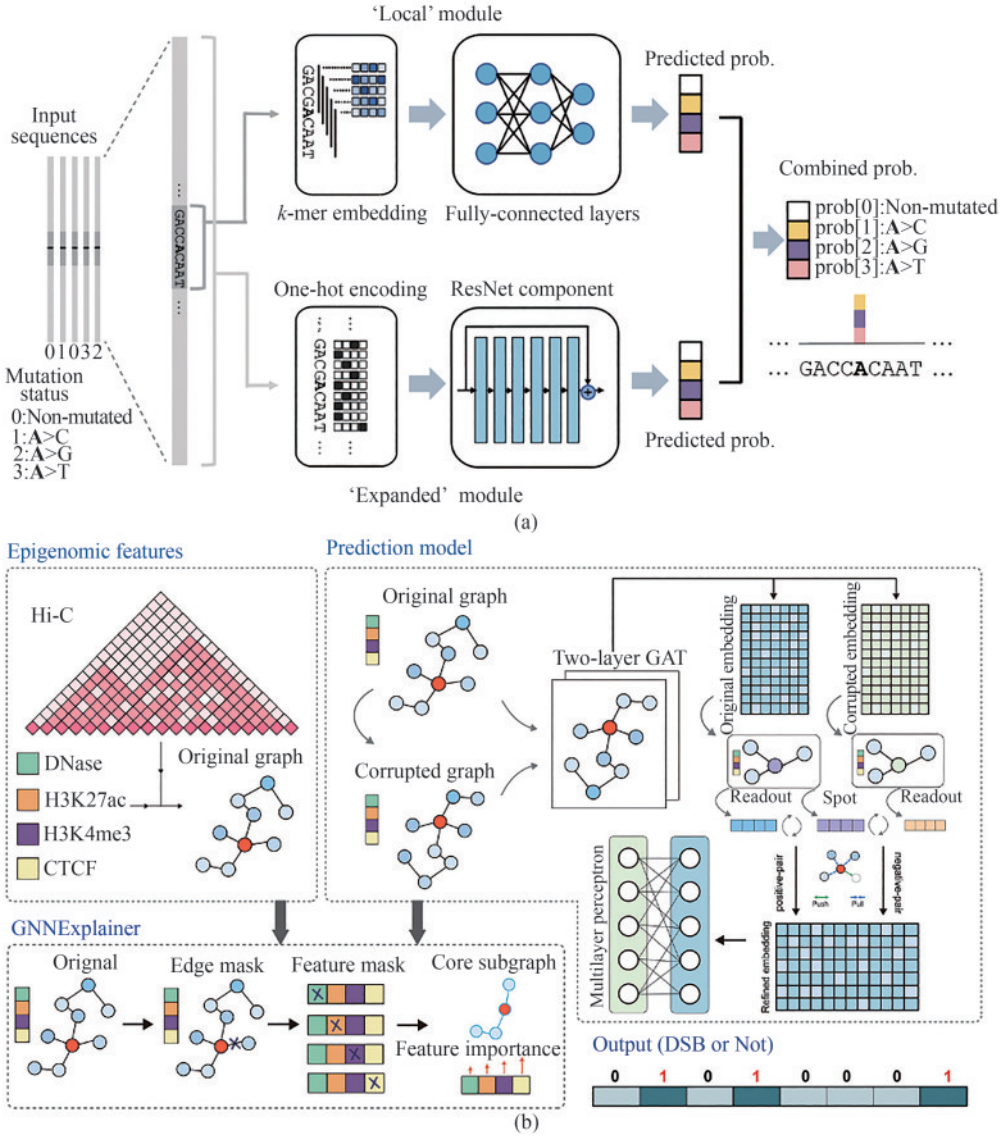


图 2　基因组设计中的人工智能方法 [21-22]

（a） MuRaL 模型整体结构与训练流程示意图：模型由“局部（local）”与“扩展（expanded）”两个模块组成：局部模块将中心核苷酸周围序列拆分为重叠的 *k*-mer，经嵌入层和全连接网络后预测非突变或三种碱基替换的概率；扩展模块对更大范围序列进行 one-hot 编码，并通过 ResNet 学习长程序列的突变倾向，同样输出四类突变概率；两模块的预测结果以等权重融合得到最终概率，模型训练时同时输入每条序列对应的真实突变状态 [21]。（b） 一种基于图对比学习的可解释 DNA 双链断裂预测模型，融合表观遗传特征与 Hi-C 信息构建图网络，通过结合图注意力网络与多层次对比学习实现节点表示的精细化学习，并由输入层、嵌入生成、对比学习和输出层四个模块协同完成预测，同时利用 GNNExplainer 解析三维染色质结构与双链断裂之间的关联 [22]

Fig. 2　Artificial intelligence methods for designing genomes [21-22]

(a) Schematic illustration of the overall architecture and training workflow of the MuRaL model. The model consists of two modules, a local module and an expanded module: The local module splits the sequence surrounding the central nucleotide into overlapping *k*-mers, which are then processed through an embedding layer and fully connected networks to predict the probabilities of no mutation or three possible base substitutions. The expanded module one-hot encodes a broader sequence context and uses ResNet to learn long-range sequence-dependent mutation propensities, also outputting four-class mutation probabilities. The predictions from the two modules are fused with equal weights to obtain the final probability. During the model training, the true mutation status corresponding to each input sequence is provided simultaneously [21]. (b) An interpretable DNA double-strand break prediction model based on the graph contrastive learning. The model integrates epigenetic features and Hi-C information to construct a graph network, combines graph attention networks with multi-level contrastive learning to refine node representatives, and performs prediction through four coordinated modules: the input layer, embedding generation, contrastive learning, and output layer. GNNExplainer is further used to interpret the association between three-dimensional chromatin architecture and DNA double-strand breaks [22]

向可预测、可扩展且具备泛化能力的系统级序列生成，为未来从头构建功能性基因簇乃至合成基因组提供了重要支撑。

尽管静态序列生成为基因组工程构建了设计空间，但生命系统的功能最终在细胞命运与动态演化过程中得以体现。因此，细胞谱系追踪与动态建模成为基因组设计中的关键环节[22]。中国科学院深圳先进技术研究院胡政团队[23]开发基于碱基编辑的条形码系统，使得细胞之间的谱系关系能够在单细胞尺度被完整记录。该系统利用具有高多样性突变能力的工程化胞嘧啶脱氨酶，在细胞分裂过程中不断积累可识别的遗传标记，从而形成细胞历史轨迹的“时间印记”。结合单细胞转录组测量，该体系可以同时获得细胞来源、分化路径及其分子状态，实现谱系-状态的多模态整合。在此基础上，进一步提出“未扰动条件的谱系动力学可用于预测扰动实验的潜在结果”的概念，意味着细胞命运不仅受当前状态影响，还受其谱系背景所决定的动态潜势约束。这一思想为构建虚拟细胞模型提供了理论基础，使得对基因扰动、基因线路构建或药物干预的结果可以通过动力学模型进行预测，为基因组设计提供贯穿“序列→状态→命运”的因果链条。

这三类研究虽聚焦不同层级，但共同构成了基因组设计的系统结构：以突变率图谱刻画序列演化的“可能性空间”，以大模型生成方法构建功能片段的“可实现空间”，以谱系与动力学建模揭示细胞行为的“因果空间”。从演化约束到功能生成，再到动态预测，它们形成了贯穿序列、调控与细胞命运的连续框架，使基因组工程从传统线性编辑走向定量化、系统化与可泛化的设计科学。在未来，随着大规模组学数据进一步积累、深度模型的表征能力不断增强以及自动化高通量平台的普及，基因组设计将逐步实现从局部优化向系统生成的转变，提供面向细胞工厂、治疗细胞与合成活体系统的核心技术支撑。

3 细胞设计：AI与机理模型驱动的动态生命系统构建

细胞是生命系统的基本功能单元。围绕细胞的生长、代谢、响应与命运调控建立可预测、可设计的模型体系，是推动合成生物学从基因与分子层面迈向系统级工程的重要内容。相较于蛋白质或基因组的静态设计，细胞设计面临着更复杂的动态过程、多尺度耦合机制与跨层级反馈调控。细胞内部存在数量庞大的酶促反应网络、代谢通路、物质传输过程与信号调控机制，而这些过程与环境、营养、能量状态以及进化压力紧密耦合，使得细胞行为具有高度的非线性、随机性与多稳态特征。近年来，人工智能与机理建模的结合为细胞设计提供新的理论工具与工程能力，使研究者能够从静态组分解析逐步迈向对细胞整体动态行为的可预测建模。

虚拟细胞模型（virtual cell model）是实现可编程生命单元的核心。基于机理的细胞模型长期以来面临的主要瓶颈在于参数数量庞大而难以测定，包括酶动力学常数、底物亲和力、调控强度及细胞内分子相互作用的定量描述。马红武提出的思路强调通过人工智能弥补机理模型中“参数未知”的结构性难题。在传统机制模型中，从蛋白质结构推导功能，从功能推导动力学规律，是一条逐级放大的不确定链条。深度学习的引入为这条链条提供了补全机制：从蛋白质的序列与结构预测功能定量特征，从反应体系的数据驱动推断酶动力学参数，并利用自动化数据解析与模型质控体系实现参数的系统化校正。在这一框架中，细胞建模由静态反应网络扩展至动态反应体系，由单一代谢通路拓展至多通路耦合的整细胞模型，体现出虚拟细胞工程从“自上而下的近似推断”向“自下而上的系统重构”的方法学演进。然而，细胞系统的复杂性远超反应网络本身。细胞是一个典型的跨尺度复杂系统，在分子层面受到酶及反应动力学约束，在中尺度受到代谢通量与细胞能量状态影响，在宏观层面则由环境、营养与时间变化共同塑造其状态空间。马红武指出，尽管人工智能在预测胞内分子相互作用方面取得显著进展，但数据是否足以支撑建立全细胞模型仍是开放问题。系统生物学强调的“宏观规律”与“分子机制”的叠加，使得细胞行为呈现既可预测又带有不可测随机性的特征。因此，虚拟细胞模型的构建不仅依赖参数推断与反应式建模，更依赖对细胞行为的统计规律、能量约束和稳态性质

的系统性理解。

在虚拟细胞研究逐渐成型的同时，数字生命（digital life）概念正推动细胞设计迈向更加统一的工程框架。清华大学国际深圳研究院李斐然[24]的研究将数字孪生技术引入细胞模拟中，提出AI与物理机理融合的细胞工厂设计模型D2Cell［图3（a）］。数字孪生概念起源于工程学，旨在将物理规律、传感器数据、历史数据和实时行为整合为可预测的计算模型。将其引入细胞建模，使得基于机理的框架得以与深度学习进行耦合，从而克服纯AI模型可解释性差、纯机理模型可预测性有限的双重局限。对于酶催化行为，深度学习能够预测多种环境变量下的活性参数；对于代谢通路，机理模型能够保证质量守恒、热力学一致性与稳态特性。D2Cell通过将这两类建模方法统一到可泛化的大模型框架中，使细胞工厂能够设计出训练集中未出现过的新型产物路径，进而拓宽了传统代谢工程的设计边界。数字孪生的思想也正在向更大尺度延伸，例如构建从器官到细胞的多尺度数字人体，为细胞疗法设计与疾病建模提供新工具。

细胞设计不仅关注单个细胞，还需关注细胞群体层面的多细胞与多物种动态行为。深圳湾实验室胡脊梁团队的研究展示了微生物群落中复杂行为的涌现特征，强调多物种系统的稳定性与多样性在生态动力学中的关键作用。在其提出的生态网络模型[27]中，物种之间的生长、互作与竞争可以通过耦合的微分方程进行描述。随着物种数量的增加，系统的动力学行为会显著复杂化，在一定参数条件下可能出现物种灭绝、种群振荡以及混沌等多种动力学态。值得注意的是，即使初始条件非常接近，系统也可能演化至不同的稳态或周期行为，体现出多物种系统的强非线性特征及对初始条件的敏感性。这一理论对细胞设计具有直接启示——细胞工厂、组织工程或肿瘤微环境中的多细胞系统都可能遵循类似的动力学规律。理解何时系统趋于稳态、何时进入振荡、何时发生崩溃，不仅是生态学问题，也是构建稳定可控细胞系统的重要理论基础。

圆桌讨论围绕生命过程中的“第一性原理”可建模性展开。部分观点认为，基因表达、表观遗传调控和细胞状态转换的核心规律尚未被现有模型完全解析，人工智能的大规模数据驱动策略仍是当前的重要途径。另一些观点则强调，进化规律、能量约束和物理机制已构成生命系统的“基本框架”，但大模型多停留在统计关联层面，未能充分学习这些规律。生命系统的复杂性与偶然性，也极大程度限制了设计的可预测性。综合不同观点可以形成明确方向：未来细胞设计的重要任务，是融合数据驱动与机理约束两类范式——用AI捕捉高维复制模式，用物理与进化原理保证合理性与因果性，从而实现数据可靠、理论可解释的细胞行为预测。

综上所述，细胞设计正在经历一场由AI、机理模型与系统生物学共同推动的范式转变。虚拟细胞模型补全了细胞行为的参数化基础，数字生命系统提供了跨尺度整合的可预测框架，而生态动力学揭示了多细胞系统的非线性结构与稳态特征。实验自动化平台、多模态测量技术与物理建模工具的持续进步，带动细胞设计正从个体反应网络的参数估计迈向生命系统的整体建模，从静态解析迈向动态预测，从线性设计迈向复杂系统的可控构建。这些进展表明，未来的细胞设计将不再局限于基因组编辑或代谢工程，而将成为一种新型的“生命系统工程”，实现从分子层面到生态层面的完整设计能力。

4 AI 合成生物学平台与方法开发

大分子、基因组、细胞的三级设计，最终需要统一的工程化平台作为依托。AI技术的深度融入，使合成生物学从“以实验为中心”的科学逐步迈向“以模型驱动的工程系统科学”。这一趋势在蛋白质设计、代谢途径构建、生物催化器开发以及多尺度细胞功能调控中尤为突出，平台化与方法学的系统革新持续助力AI合成生物学走向规模化应用。

在代谢途径和天然产物合成领域，AI平台的作用从“辅助酶功能预测”扩展至“端到端的生物合成路径反演”。中国科学院深圳先进技术研究院罗小舟等[25]基于Transformer机制的天然产物语言模型能够学习化学反应转换规律，以目标产物

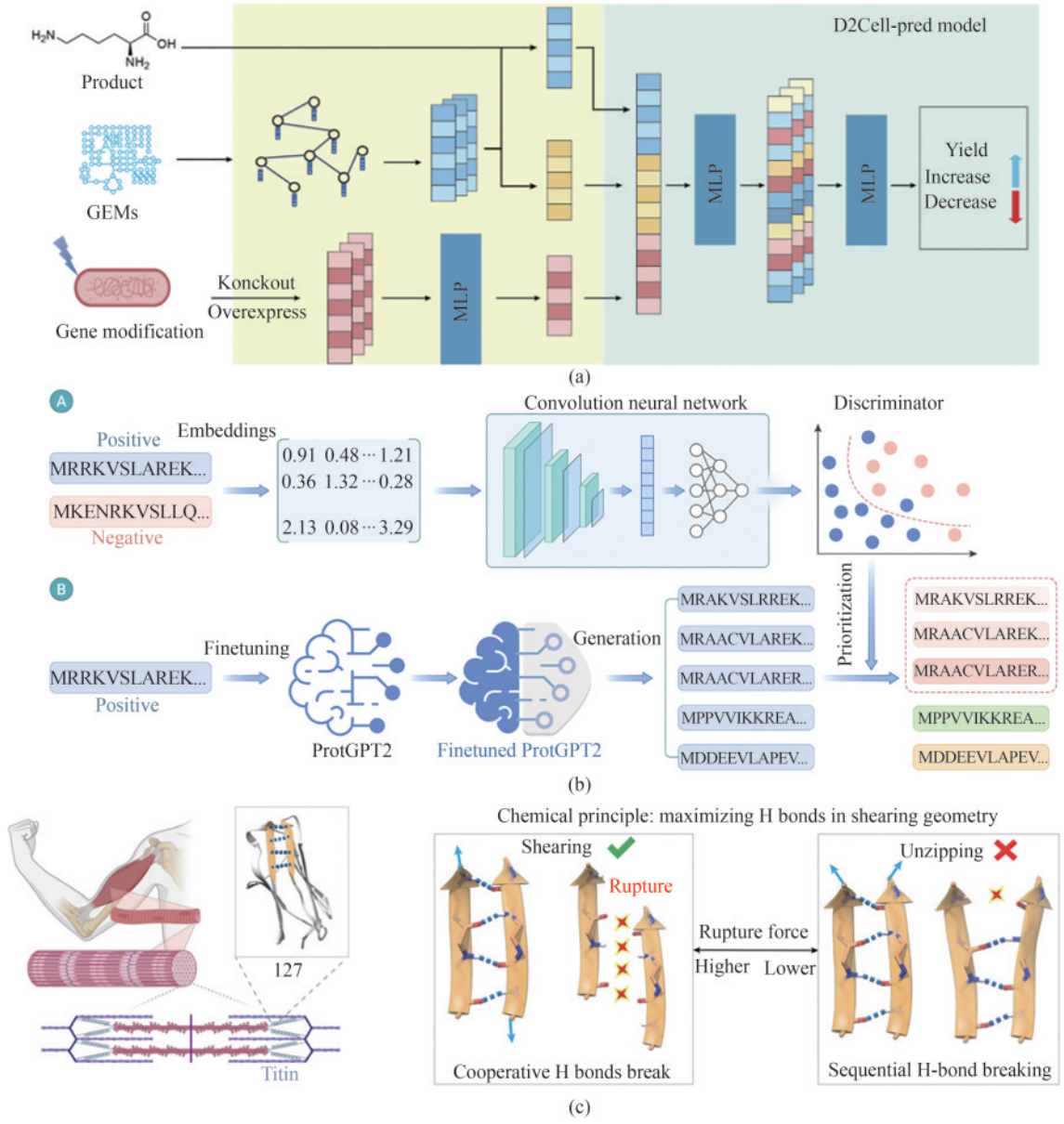


图3 细胞设计中的人工智能方法[24-26]

（a） D2Cell模型示意图：该模型用于预测基因修饰对产物得率的影响，以目标产物、基因组尺度代谢模型结构及基因修饰集合为输入，输出相应基因修饰对细胞工厂生产性能的影响[24]。（b） 在判别器A中，正、负样本序列首先通过ProtTrans进行嵌入表示，随后输入包含三层卷积的神经网络，并通过全连接层完成判别；在生成器B中，利用正样本序列对ProtGPT2进行微调以生成候选正样本序列，最终由判别器对生成序列进行筛选，选出最可能的正样本候选[25]。（c） 蛋白质力学展开机制示意图。人体肌肉中以Ig样结构域为代表的蛋白结构及其在受力条件下的力学展开过程，突出I27结构域中由主链氢键稳定的承力β链在机械拉伸中的解折叠机制[26]

Fig. 3 Artificial intelligence methods for designing cells[24-26]

(a) Schematic illustration of the D2Cell model. This model is used to predict the effects of genetic modifications on the product yield. It takes the target product, the structure of a genome-scale metabolic model, and a set of genetic modifications as inputs to predict the impact of these modifications on the production performance of the corresponding cell factory[24]. (b) In discriminator A, positive and negative sample sequences are first embedded using ProtTrans, and then fed into a neural network containing three convolutional layers to assess their contributions. In generator B, ProtGPT2 is fine-tuned using positive sample sequences to generate candidate positive sequences. The generated sequences are subsequently screened by the discriminator to identify the most likely positive candidates[25]. (c) Schematic illustration of the mechanical unfolding mechanism of proteins. The figure shows protein structures represented by Ig-like domains in human muscle and their force-induced unfolding process, highlighting the mechanical unfolding mechanism of the load-bearing β-strands in the I27 domain, which are stabilized by backbone hydrogen bonds under the mechanical stretching[26]

为输入实现高效的前体与反应路径生成［图3（b）］。同时，配合搜索策略的路径优化系统可系统化地对候选路径进行评分、扩展与更新，并在多个复杂天然产物案例中展示了接近或达到人工专家水平的路径设计质量。另外，酶检索与酶工程模块通过多模态描述（序列、结构、反应中心）整合，显著改进了酶-底物特异性预测的准确性。AI设计方法已被扩展至蛋白稳定性预测与分子力学性质调控等更具物理复杂性的体系。蛋白热稳定性对工业酶、PCR聚合酶、极端环境生物学等领域具有关键作用，传统实验筛选成本高、周期长。南京大学郑鹏团队[26]通过AI构建可预测高温结构保持能力的模型，在构象层面筛选具有高能垒特征的序列，并结合分子动力学模拟与突变扫描策略，对α-螺旋和β-折叠的力学断裂机理进行高精度建模，这种“预测-筛选-再设计”的闭环模式使得蛋白质稳定性工程能够在大规模上获得可控性，实现从自然启发到超越自然材料的跨越［图3（c）］。在蛋白质设计方向，深度生成模型及结构预测模型已推动功能性分子设计从经验驱动走向全空间搜索。以高维结构中Binder设计为例，深圳医学科学院杨为等针对CTLA-4、TGFβRII、PD-L1等重要免疫检查点靶标，基于深度学习的蛋白质生成平台能够通过对接界面几何约束与超大规模骨架库搜索，实现全新具有内凹结合界面的蛋白结合体的精确设计与构建，并在多个案例中取得远优于传统设计的命中率表现［图4（a）］[28]。

综合来看，一个完整的AI合成生物学平台不再只是模型集合，而是贯穿数据、模型、设计、自动化验证与知识反馈的工程系统。以多智能体框架为例，每个模型代理负责特定子任务，如生物合成路径规划、酶功能识别、蛋白结构生成、功能模拟与筛选等，而总控模型则负责策略协调、信息融合与任务调度。这种体系结构使平台能够在复杂的设计空间中进行全局搜索并具备自适应优化能力。例如，在天然产物生物合成中，模型可从目标结构出发，自动选择路径节点、检索必要酶、预测功能并调控路径评分，实现从目标分子到可执行路径的全自动规划；在蛋白质工程中，模型可生成支架、调整界面几何、预测结合能并调用实验模块进行快速验证；在酶设计中，上海交通大学郑双佳团队[29]使用几何等变神经网络实现了蛋白质复合物结合能力的快速预测［图4（b）］。

此外，工程化可扩展性是平台走向产业级应用的重要条件。平台需具备多尺度数据融合能力，以整合序列、结构、动力学、表型以及组学数据；需构建统一的模型中间表示，使不同领域模型（蛋白语言模型、结构生成模型、反应预测模型等）能够互操作；还需开发高效的实验自动化接口，使设计环节产生的候选分子能够快速进入验证流程，从而形成快速收敛的设计-构建-测试-学习（design-build-test-learn，DBTL）循环。在这一过程中，AI不再只是建模工具，而是生物工程系统的“操作系统”，使研究者能够在更大尺度上操控生命体系的可设计性与可预测性。

综上所述，AI合成生物学平台正朝着“多模态大模型-自动化设计工具链-实验反馈系统-智能体协同优化”的综合架构方向演化，AI模型正在以不同方式重构合成生物学研究方式，使传统依赖经验、试错与高成本实验的过程转变为系统化可控、可预测和可扩展的工程科学。伴随模型规模的增长、数据质量的提升与物理机制嵌入的加强，未来的AI合成生物学平台有望实现真正意义上的“生物系统自动化设计”，并成为连接基础科学研究与大规模实际应用的核心基础设施。

5 总 结

当前，人工智能+合成生物学正处于全球科技竞争与战略博弈的关键交汇点，逐步成为各国争夺未来产业主导权的重要赛道。以美国和欧洲为代表的发达国家和地区，依托其在基础研究、原始创新能力及数据生态方面的长期优势，持续加码布局AI驱动的生物设计、自动化实验平台以及生物安全治理体系，力图在核心算法、关键数据与标准制定层面占据主导地位。在此背景下，相关技术不仅体现为科学问题的突破，更逐渐演化为国家层面的科技竞争与规则博弈。

中国高度重视AI与合成生物学战略方向，已在国家层面通过“十四五”规划、生物经济发展规划以及人工智能发展规划等政策文件持续推进相关布局，将合成生物学与人工智能明确列为重

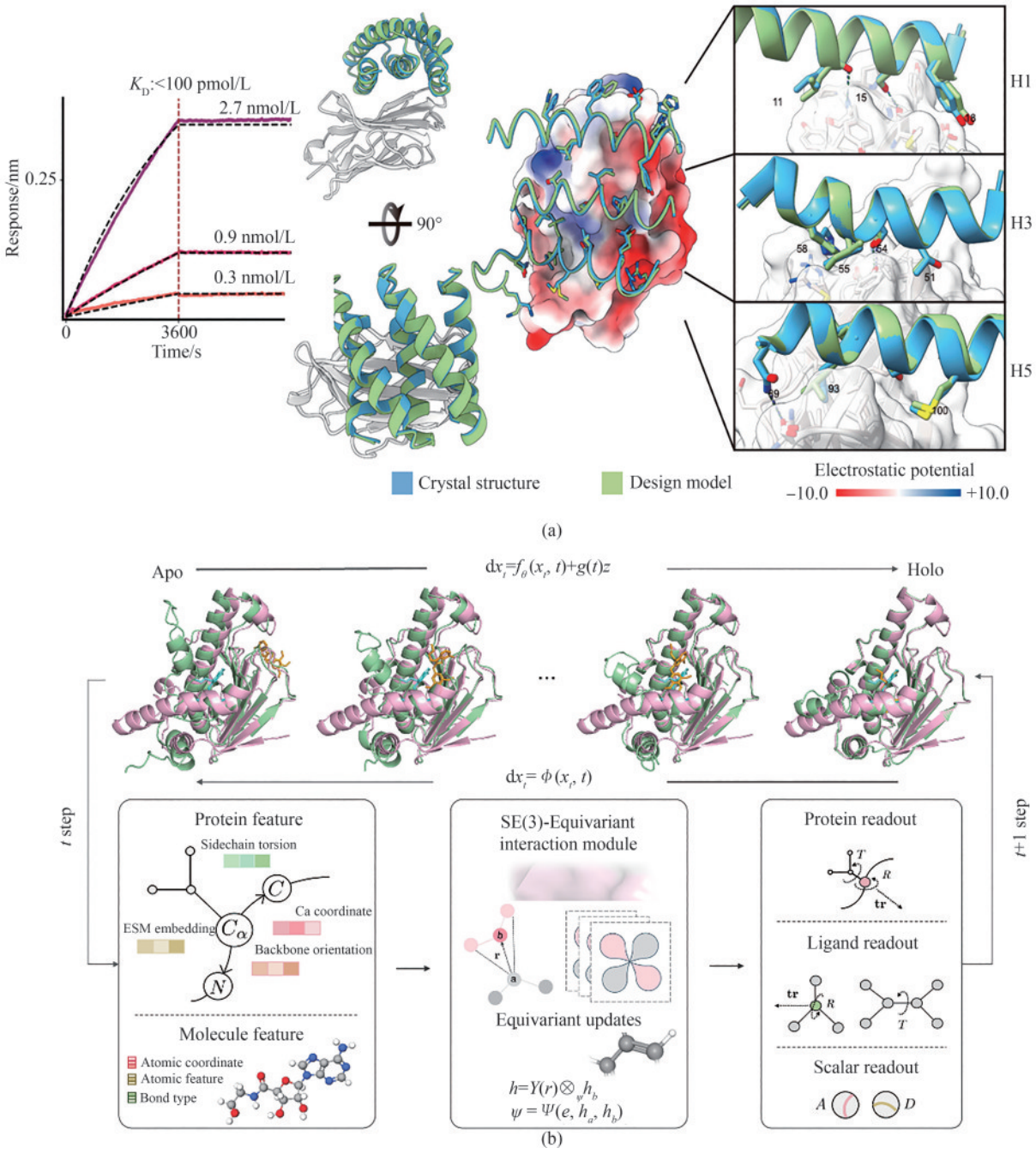


图 4 合成生物学人工智能方法的应用 [28-29]

（a） 通过生物层干涉实验测定 5HCS_CTLA4_1 与 CTLA-4 在不同纳摩尔浓度下的结合亲和力，并结合两者复合物的晶体结构解析，揭示 5HCS_CTLA4_1（绿） CTLA-4（白）之间经设计的相互作用界面 [28]。 （b） 粉色表示 holo 构象，绿色表示初始 apo 及模型预测构象，橙色为预测配体。模型以蛋白-配体的特征及当前构象为输入，输出包括蛋白与配体的平移、旋转及扭转角更新，并预测结合亲和力与置信度；训练阶段学习从 apo 向 holo 构象的转变，推理阶段通过多轮迭代逐步更新结构 [29]

Fig. 4 Applications of artificial intelligence methods in synthetic biology [28-29]

(a) Binding affinity between 5HCS_CTLA4_1 and CTLA-4 at different nanomolar concentrations measured by biolayer interferometry, together with the crystal structure of the complex reveals the designed interaction interfaces between 5HCS_CTLA4_1 (green) and CTLA-4 (white) [28]. (b) Pink, green, and orange represent the holo conformation, initial apo and model-predicted conformations, and orange predicted ligand, respectively. The model takes protein-ligand features and the current conformations as inputs to predict the updates of protein and ligand translation, rotation, and torsion angles, as well as binding affinity and confidence. During training, the model learns the transition from the apo to the holo conformation; during reasoning, the structure is progressively updated through multiple iterative rounds [29]

点发展领域，并在深圳、上海等地加快建设生物制造与智能化研发平台，推动技术与产业深度融合。依托在算力基础设施、工程化能力及应用场景方面的快速发展，中国在生物制造、工艺优化及规模化应用方面展现出显著优势。

面向未来，应围绕合成生物学“设计-构建-测试-学习（DBTL）”闭环体系，进一步强化高质量生物数据资源建设与标准化积累，提升面向生物大分子设计与代谢网络优化的智能算法能力，并持续增强算力基础设施对大规模模型训练与高通量模拟的支撑作用。同时，加快自动化实验平台与智能设计模型的深度耦合，推动从分子设计到细胞工厂构建的全流程智能化升级。在此基础上，结合国家战略需求与产业发展导向，完善相关技术标准与治理框架，在开放合作与有序竞争中不断提升“AI+合成生物学”的系统创新能力与国际影响力。

致谢：首届人工智能合成生物学研讨会得到中国生物工程学会合成生物学分会、深圳理工大学合成生物学院、中国科学院深圳先进技术研究院合成生物学研究所、美国化学会广东分会(ACS Guangdong Chapter)等机构的支持。

参 考 文 献

[1] “中国学科及前沿领域发展战略研究”项目组. 中国合成生物学2035发展战略[M]. 北京: 科学出版社, 2023.
Chinese Disciplines and Frontier Fields Development Strategy Research Project Group. Development strategy of synthetic biology 2035 in China[M]. Beijing: Science Press, 2023.

[2] 张先恩. 十大措施合力探索合成生物学发展“深圳模式”[J]. 中国科学院院刊, 2024, 39(9): 1574-1582.
ZHANG X E. Ten measures to jointly explore Shenzhen model of synthetic biology development[J]. Bulletin of Chinese Academy of Sciences, 2024, 39(9): 1574-1582.

[3] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[4] BAKER D, SALI A. Protein structure prediction and structural genomics[J]. Science, 2001, 294(5540): 93-96.

[5] 合成生物学发展战略研究组. 合成生物学路线图-2030: 驱动下一代生物制造的引擎[M]. 北京: 科学出版社, 2024.
Strategic Research Group for the Development of Synthetic Biology. Roadmap of synthetic biology-2030: engine driving the next generation bio-manufacturing[M]. Beijing: Science Press, 2024.

[6] 李玉娟, 傅雄飞, 张先恩. 合成生物学发展脉络概述[J]. 中国生物工程杂志, 2024, 44(1): 52-60.
LI Y J, FU X F, ZHANG X E. A brief overview of synthetic biology[J]. China Biotechnology, 2024, 44(1): 52-60.

[7] ZHOU C, LI H C, WANG J X, et al. *De novo* designed voltage-gated anion channels suppress neuron firing[J]. Cell, 2025, 188 (26): 7495-7511.e21.

[8] XU C F, LU P L, GAMAL EL-DIN T M, et al. Computational design of transmembrane pores[J]. Nature, 2020, 585(7823): 129-134.

[9] ZHU J Y, LIANG M F, SUN K, et al. *De novo* design of transmembrane fluorescence-activating proteins[J]. Nature, 2025, 640(8057): 249-257.

[10] LI Z, WANG S Z, NATTERMANN U, et al. Accurate computational design of three-dimensional protein crystals[J]. Nature Materials, 2023, 22(12): 1556-1563.

[11] CHEN Y C, SUN J Y, SHI K L, et al. Glycolysis-compatible urethanases for polyurethane recycling[J]. Science, 2025, 390 (6772): 503-509.

[12] ZHANG C X, LIU Q C, FREDDOLINO L. StarFunc: fusing template-based and deep learning approaches for accurate protein function prediction[J]. Genomics, Proteomics & Bioinformatics, 2026: qzag018.

[13] SHEFFLER W, YANG E C, DOWLING Q, et al. Fast and versatile sequence-independent protein docking for nanomaterials design using RPXDock[J]. PLoS Computational Biology, 2023, 19(5): e1010680.

[14] LIU Y F, WANG S, DONG J X, et al. *De novo* protein design with a denoising diffusion network independent of pretrained structure prediction models[J]. Nature Methods, 2024, 21(11): 2107-2116.

[15] LIU Y F, ZHANG L, WANG W L, et al. Rotamer-free protein sequence design based on deep learning and self-consistency [J]. Nature Computational Science, 2022, 2(7): 451-462.

[16] DING Y, LUO X, GUO J S, et al. Identification of gut microbial bile acid metabolic enzymes *via* an AI-assisted pipeline[J]. Cell, 2025, 188(21): 6012-6027.e20.

[17] JIN Q H, WANG Y K, CHEN D C, et al. *De novo* design of small molecule-regulated protein oligomers[J]. Science, 2026, 391(6780): eady6017.

[18] SUN J Y, ZHU T, CUI Y L, et al. Structure-based self-supervised learning enables ultrafast protein stability prediction upon mutation[J]. Innovation, 2025, 6(1): 100750.

[19] TANG Y, TIAN X Y, WANG M, et al. The β-D-*manno*-

heptoses are immune agonists across Kingdoms[J]. Science, 2024, 385(6709): 678-684.

[20] MA R Z, ZHANG Z Y, WANG Z C, et al. RiboFlow: conditional *de novo* RNA co-design *via* synergistic flow matching[C/OL]. The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025[2026-02-26]. https://neurips.cc/virtual/2025/loc/san-diego/poster/117145.

[21] FANG Y Y, DENG S Y, LI C. A generalizable deep learning framework for inferring fine-scale germline mutation rate maps [J]. Nature Machine Intelligence, 2022, 4(12): 1209-1223.

[22] XU K, YU Z Y, SUN C Z, et al. ChromInSight: revealing DNA double-strand breaks through chromatin structural insights with an interpretable graph neural network framework[J]. Advanced Science, 2025, 12(36): e04571.

[23] LU Z L, MO S L, XIE D, et al. Polyclonal-to-monoclonal transition in colorectal precancerous evolution[J]. Nature, 2024, 636, 233-240.

[24] LI X W, LIANG Z, GUO, Z T, et al. Leveraging large language models for metabolic engineering design[PP/OL]. bioRxiv(2024-09-13)[2025-02-26]. https://doi.org/10.1101/2024.09.09.612023.

[25] ZENG Z S, XU R F, GUO J, et al. Accelerating functional protein discovery with GPT models: Antimicrobials and enzymes[J]. The Innovation Life, 2025, 3(2): 100133.

[26] ZHENG B, LU Z J, WANG S C, et al. Computational design of superstable proteins through maximized hydrogen bonding [J]. Nature Chemistry, 2026, 18(2): 364-373.

[27] HU J L, AMOR D R, BARBIER M, et al. Emergent phases of ecological diversity and dynamics mapped in microcosms[J]. Science, 2022, 378(6615): 85-89.

[28] YANG W, HICKS D R, GHOSH A, et al. Design of high-affinity binders to immune modulating receptors for cancer immunotherapy[J]. Nature Communications, 2025, 16: 2001.

[29] LU W, ZHANG J X, HUANG W F, et al. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model[J]. Nature Communications, 2024, 15: 1071.

共同通讯作者：兰晶岗，男，特聘教授。研究方向为计算化学与AI for Science，机器学习方法研发及其在生物系统与物理体系中的跨尺度应用。

E-mail：Jinggang.lan@suat-sz.edu.cn

共同通讯作者：傅雄飞，男，研究员，中国科学院深圳先进技术研究院合成生物学研究所所长，定量合成生物学全国重点实验室副主任，深圳合成生物创新研究院副院长。主要从事合成生物学、定量生物学、统计物理与复杂系统等多个交叉学科领域，聚焦随机涨落在细菌单细胞与多细胞有序空间结构形成过程中的作用，通过定量理论与合成重构结合，揭示跨尺度涌现性原理。

E-mail：xiongfei.fu@siat.ac.cn

共同通讯作者：汪小我，男，博士，教授。主要研究方向为模式识别与机器学习、生物信息学、合成生物学。

E-mail：xwwang@tsinghua.edu.cn

共同通讯作者：张先恩，男，深圳理工大学合成生物学院院长、讲席教授，中国科学院生物物理所研究员。从事合成生物学、生物传感和纳米生物学交叉创新研究，并用于解决细胞生物学、病毒学、肿瘤生物学问题。

E-mail：zhangxianen@suat-sz.edu.cn；zhangxe@ibp.ac.cn