第 10 卷 第 5 期
2021 年 9 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 10 No. 5
Sep. 2021

引文格式：
李敏, 林子杰, 廖文斌, 等. 人工智能在合成生物学的应用 [J]. 集成技术, 2021, 10(5): 43-56.
Li M, Lin ZJ, Liao WB, et al. Application of artificial intelligence in synthetic biology: a review [J]. Journal of Integration Technology, 2021, 10(5): 43-56.

人工智能在合成生物学的应用

李 敏[1,2#] 林子杰[3#] 廖文斌[3] 陈廷柏[3] 李坚强[3*] 陈 杰[3*] 肖敏凤[1,4*]

[1]（深圳华大生命科学研究院 深圳 518083）
[2]（中国科学院大学生命科学学院 北京 100049）
[3]（深圳大学计算机与软件学院 深圳 518060）
[4]（深圳市未知病原体应急检测重点实验室 深圳 518083）

摘 要 生命系统极其复杂，难以精确描述和预测，这给高效设计合成生物系统提出了挑战，故在合成生物系统构建中往往须进行海量工程试错和优化。近年来，人工智能技术快速发展，其基于海量数据的持续学习能力和在未知空间的智能探索能力有效契合了当前合成生物学工程化试错平台的需求，在复杂生物特征的挖掘与生命系统的设计方面具备巨大潜力。该文回顾并总结人工智能在合成元件工程、线路工程、代谢工程及基因组工程领域的研究进展，并分析和讨论人工智能与合成生物学交叉研究在数据标准化、平台智能化、实验自动化、预测精准化方面存在的一系列挑战。人工智能和合成生物学的融合有望给“设计—构建—测试—学习”闭环的全流程带来变革，而孕育“类合成生物学家”也将反过来引起人工智能技术的飞跃。

关键词 类合成生物学家；智能化试错；合成生物系统；生物铸造厂；海量工程平台
中图分类号 Q-1；Q 812；TP 18；TP 181 **文献标志码** A
doi: 10.12146/j.issn.2095-3135.20210510001

收稿日期：2021-05-10 **修回日期**：2021-05-20
基金项目：国家重点研发计划“合成生物学专项”项目（2020YFA0908700）；国家自然科学基金面上项目（62072315）；深圳市孔雀团队项目（KQTD2015033117210153）
作者简介：李敏（共同第一作者），硕士研究生，研究方向为微生物基因组学、生物信息学和合成生物学；林子杰（共同第一作者），硕士研究生，研究方向为合成生物智能计算、机器学习理论；廖文斌，硕士研究生，研究方向为合成生物智能计算、机器学习；陈廷柏，硕士研究生，研究方向为合成生物智能计算、机器学习；李坚强（通讯作者），教授，研究方向为人工智能、数据分析以及智能机器人，E-mail：lijq@szu.edu.cn；陈杰（通讯作者），助理教授，研究方向为合成生物智能计算、机器学习理论与算法，E-mail：chenjie@szu.edu.cn；肖敏凤（通讯作者），副研究员，研究方向为微生物基因组学与合成生物学，E-mail：xiaominfeng@genomics.cn。

Application of Artificial Intelligence in Synthetic Biology: A Review

LI Min[1,2#] LIN Zijie[3#] LIAO Wenbin[3] CHEN Tingbo[3] LI Jianqiang[3*]
CHEN Jie[3*] XIAO Minfeng[1,4*]

[1](*BGI-Shenzhen*, *Shenzhen* 518083, *China*)
[2](*School of Life Sciences*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)
[3](*College of Computer Science and Software Engineering*, *Shenzhen University*, *Shenzhen* 518060, *China*)
[4](*Shenzhen Key Laboratory of Unknown Pathogen Identification*, *Shenzhen* 518083, *China*)
[*]*Corresponding Author*: *lijq@szu.edu.cn*; *chenjie@szu.edu.cn*; *xiaominfeng@genomics.cn*
[#]*Equal Contribution*

Abstract Living systems are extremely sophisticated and difficult to accurately describe and predict, posing challenges in designing synthetic biological systems. Therefore, massively parallel trial-and-error processes are often required to optimize synthetic biological systems. In recent years, intelligent technology has experienced rapid development and has demonstrated continual learning capacity from massive data and intelligent exploring ability in unknown space, which perfectly meets the needs of the current trial-and-error platform of synthetic biology engineering and shows great potential in mining complex biological patterns and in designing biosystems. This article reviews the progresses of applying artificial intelligence (AI) in the fields of synthetic biological parts engineering, circuit engineering, metabolic engineering, and genome engineering. This article also analyzes a series of challenges in data standardization, platform intellectualization, experimental automation, and accurate prediction of cross-over studies between AI and synthetic biology. By solving these challenges, the entire workflow of "design-build-test-learn" in synthetic biology is expected to be revolutionized by AI, and creating an "AI synthetic biologist" would in turn lead to the technological advances in AI.

Keywords AI synthetic biologist; intelligent trial-and-error; synthetic biological system; biofoundry; massively parallel engineering platform

Funding This work is supported by National Key R&D Program of China (2020YFA0908700), National Natural Science Foundation of China (62072315), and Shenzhen Peacock Team Plan (KQTD2015033117210153)

1 引 言

合成生物学以人为设计和构建生命系统为目标，近年来在生物医疗技术和药物的研发[1-7]、蛋白质和其他化合物的生产[8-10]以及环境保护[5, 8-9]等领域展现出巨大的发展潜力。有别于传统生命科学，合成生物学具备多学科交叉、多技术融合的特征，遵循工程学本质，在人工设计的指导下，基于特定底盘细胞，自下而上地对生物元件、线路模块、代谢网络和基因组等进行标准化表征、通用化设计构建、可控化运行，并持续学习和优化。

随着合成生物学涉及的功能和潜在应用的不断拓展，运用合成生物学的复杂性和跨学科知识需求也在迅速增长[11-16]。然而，生命系统极其精密，包含大量不同的基因和调控元件，而元件之间又以海量不同的组合形成模块、网络，难以

精确描述和预测，因此即使设计小型的基因线路也需要反复调试[17]。工程学思维和方法是克服这一难题的利器，即大规模测试不同元件、线路模块、网络和底盘的组合，积累海量实验数据，从而指导合成生物系统的理性设计和优化[17]。合成生物自动化设施（Biofoundry）是工程学平台搭建的一大核心，依照“设计—构建—测试—学习”（Design-Build-Test-Learn，DBTL）的闭环策略组织工艺流程，通过自动化、高通量生物学实验试错获得符合预期的合成生物系统[17]。但当前工程化试错存在海量的试错空间，实验成本极其高昂，并且缺乏标准化、定量的表征手段和智能化试错、优化、学习理论与技术的系统性支撑，阻碍了工程化研究平台指导合成生物系统的设计与改造的发展。因此，需要运用一种方法将新知识和新技术流程很好地集成到合成生物学工程中，以提高试错效率、降低试错成本。

随着人工智能（Artificial Intelligence，AI）技术的快速发展，在软件、电子和机械系统等不同领域的工程设计中，使用人工智能技术来捕获人类专家知识并将其嵌入辅助工具中是很常用的思路[18]。人工智能技术基于海量数据的持续学习能力和在未知空间的智能探索能力，有效地契合了当前合成生物学工程化试错平台的需求。尽管生命体很复杂并且未被完全理解，但是人工智能技术可以找到很多突破口显著改变合成生物学工程的效能。人工智能技术的核心是机器学习模型与算法，其本质是基于一组数学规则或统计假设，对机器进行编程从而学习数据集中的模式与规律。通常说来，机器学习的目标是从给定数据集中发现特征之间的联系从而建立起预测模型，输出值可以是二元响应、多分类标签或连续值。其中，训练好的预测模型需要具有较好的泛化能力，即能较准确地预测训练集外的样本。比较经典的预测模型有逻辑回归模型、决策树模型、贝叶斯概率模型、支持向量机、卷积神经网络（Convolutional Neural Network）和循环神经网络（Recurrent Neural Network）等。在生物学和生物医学研究的大数据时代，机器学习模型与算法的一个关键优势是可自动挖掘数据中可能被忽略的模式，在发现复杂生命系统的内在规律方面起关键作用。人工智能技术在生物学领域已经具有广泛的应用，包括基因注释[19]、蛋白质功能的预测[20]、基因线路的预测[21]、代谢网络的预测[15, 22-23]和复杂微生物群落的表征[24]等。然而，合成生物学实验通常时间跨度大、成本高以及 DBTL 迭代次数有限，导致预测模型的训练数据极度不足，这也给人工智能技术带来了新的挑战。本文综述了近年来人工智能技术在合成元件工程、线路工程、代谢工程及基因组工程领域的研究进展，并在此基础上提炼归纳人工智能与合成生物学两大领域交叉融合所面临的挑战，提出开发基于人工智能完成 DBTL 闭环的“类合成生物学家”见解。

2　人工智能应用于合成生物学的国内外研究现状

21 世纪以来，人工智能与合成生物学交叉研究驱使元件工程、线路工程、代谢工程、基因组工程等领域取得了一些代表性的进展，并使许多具备鲜明领域交叉特色的创新研究手段和理论得以成功运用。其中，2005—2017 年为缓慢发展阶段，研究主要集中在线路工程；2018—2021 年为相对高速发展阶段，人工智能在元件工程、线路工程、代谢工程、基因组工程等领域均崭露头角。这意味着，人工智能开始有效地解决合成生物学各子领域的技术难题，开辟合成生物学发展的新道路（图 1）。

2.1　元件工程

生物元件是合成生物系统中最简单、最基本的单元，通常指一小段具有特定功能的核酸

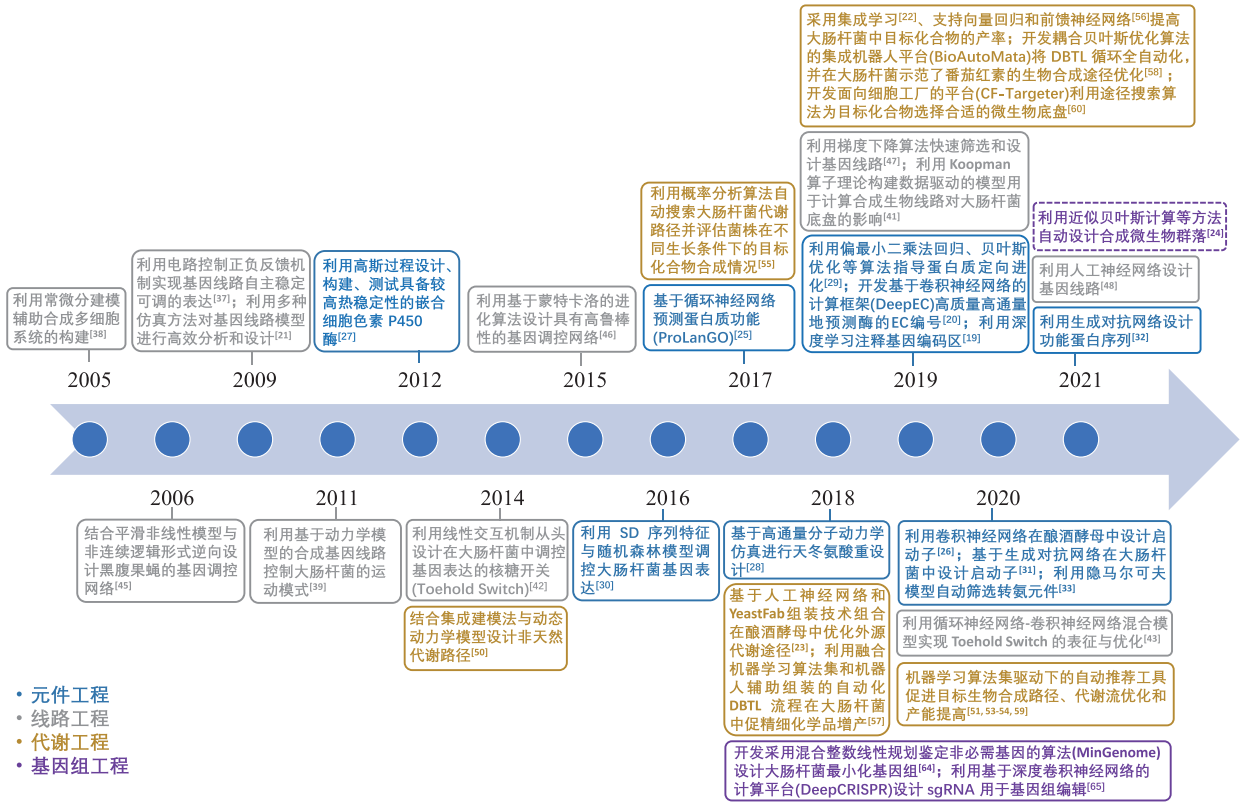

图 1　2005—2021 年人工智能应用于合成生物学的代表性进展

Fig. 1　Highlights of applying AI in synthetic biology from 2005 to 2021

和氨基酸序列。在大规模的生物智能设计中，生物元件像“搭积木”一样被用于组装具有特定生物学功能的装置和系统。在传统的生物信息学和基因组学研究中，联合多组学与序列特征分析可以得到特定的生物功能元件，如启动子、核糖体结合位点、蛋白编码基因、终止子和操纵子等。然而，从核酸和氨基酸序列到生物元件的挖掘与功能解读之间还存在巨大鸿沟。已有研究表明，人工智能技术可改善生物元件的鉴定和功能注释效率。DeepRibo[19]利用卷积神经网络和循环神经网络可有效注释基因编码区。ProLanGO[25]则是一种基于循环神经网络的神经机器翻译方法，其将蛋白质功能预测问题转化为语言翻译问题。DeepEC[20]利用 3 个相互独立的卷积神经网络联合同源分析工具 DIAMOND 预测蛋白质 EC(Enzyme Commission)编码以辅助理解酶的功能和总体细胞代谢。Kotopka 等[26]构建的卷积神经网络模型可实现对酵母启动子序列活性的高精度预测与设计。

目前，已发掘的天然生物元件结构及功能较为单一、保守，理性设计和定向进化技术是优化现有元件结构、增强其功能特性的主要策略。但这两种方法都耗时长且成本高，而机器学习通过学习序列中变异信息的特征来筛选出可能进化方向的序列，从而加速理性设计和定向进化。Romero 等[27]使用高斯过程(Gaussian Process)设计的细胞色素 P450 酶(Cytochrome P450)比先前通过嵌合染色体、理性设计或定向进化产生的酶具备更耐高温的特性。Li 等[28]利用高通量分子动力学仿真等计算机方法辅助重设计天冬氨酸酶，将其转化为不对称加氢反应的酶，由此扩大了这种酶的生产，并获得了可用于制药和其他生物活性化合物的高纯度元件。Yang 等[29]利用偏最小二乘法回归、贝叶斯优化等算法指导蛋白质

定向进化，从而提高氰化反应中蛋白质的催化效率。在蛋白质的翻译中，核糖体结合位点效率是决定蛋白质表达量的重要因素之一。Bonde 等[30]构建了一种基于随机森林的 EMOPEC（Empirical Model and Oligos for Protein Expression Changes）工具，用于全面评估核糖体结合位点上的 SD 序列（Shine-Dalgarno Sequence）对蛋白质表达的影响，并通过修改 SD 序列上的若干碱基，对大肠杆菌基因表达水平进行精准调节。

元件工程中更具挑战意义的是设计合成自然界不存在的元件，而人工智能在其中扮演着十分重要的角色。在 DNA 元件设计上，Wang 等[31]将生成对抗网络（Generative Adversarial Network）模型与支持向量机活性预测模型相结合来设计启动子，其中约 70.8% 的启动子兼具结构新颖及功能稳定的特性。该项工作为新型启动子元件的从头设计提供了端到端的方法，表明深度学习方法具有从头设计基因元件的潜力。在蛋白质元件设计上，Repecka 等[32]研究表明人工智能可辅助生成多样化的功能蛋白，其提出的 ProteinGAN 从复杂的氨基酸序列空间中学习蛋白质演化关系，并创建与天然蛋白的生物特性接近的新功能蛋白。Li 等[33]利用隐马尔可夫模型（Hidden Markov Model）对转氨酶序列和结构进行组合分析，建立高效快速的计算方法来筛选不同家族的转氨元件，最终建立了底物特异性互补的转氨元件工具箱，实现对天然 L-氨基酸的全覆盖，打通了 L-氨基酸到酮酸及相关高价值衍生物的绿色合成途径。

2.2　线路工程

人工基因线路是利用元件工程中的各类元件针对多样的需求依照电子工程中电路搭建的思维进行设计及功能优化，从而达到对生命的重编程。基于双稳态开关（Toggle Switch）[34-35]、振荡器（Oscillator）[36-37]和细胞通讯模块[38-39]等最简单的小型功能模块，研究人员根据目标重新组合或优化调整，设计出能够执行复杂逻辑功能的新颖基因线路，从而对细胞行为进行精准的时空控制，以应对复杂的生物环境[40]。

但是，合成基因线路的设计和构建远非易事。早期设计的基因线路通常需要进行多次、长时间的调试才能正常运行，且无法确定其对底盘细胞的其他影响。Hasnain 等[41]利用 Koopman 算子理论构建数据驱动的模型用于计算合成生物线路对大肠杆菌底盘的影响。Myers 等[21]开发了一种工具——iBioSim 利用多种仿真方法对基因线路模型进行高效分析和设计，可用于维护基因线路模型以及实验和仿真数据记录。尽管取得了以上进展，但在大型复杂的合成网络中，生物元件可能相互交互造成串扰，可用的生物回路元件的数量和正交性带来的限制阻碍了在活细胞中构建稳定运行的复杂回路。Green 等[42]利用线性交互机制从头设计在大肠杆菌中调控基因表达的核糖开关——Toehold Switch。Toehold Switch 不仅可以感应同源 RNA 从而激活基因表达，而且实现了较高的正交性、较低的系统串扰、可编程性以及较广的动态范围，但仍面临一定的设计瓶颈，譬如筛选有用的 Toehold Switch 通常需要开展大量实验，消耗很高的时间和经济成本。于是，Valeri 等[43]将 STORM（Sequence-based Toehold Optimization and Redesign Model）和 NuSpeak（Nucleic-Acid Speech）循环神经网络-卷积神经网络混合模型用于表征和优化 Toehold Switch。在深度学习架构中使用卷积过滤器、注意力机制和迁移学习对模型进行优化，进一步改进了面对稀疏的训练数据的性能，为调节开关的选择和设计提供了从序列到功能的深度学习框架，并增强了构建有效的生物电路和精确诊断的能力。

一个基因线路的设计被提出后，计算机仿真策略可确定该线路可以执行哪些任务，并通过修改参数以实现所需的功能。逆向工程策略利用计

算模型从基因表达数据中提取基因线路的调控结构和动力学，探索可能的基因调控线路的配置库(如基因激活或抑制强度)，以找到可以执行该功能的配置条件[44-45]。但是，由于基因线路配置的数量随基因数量的增加而迅速增加，因此这种方法的计算量巨大，需要用更高效的算法来克服这一挑战。蒙特卡洛方法提供了一种可行的替代解决方案，即反复选择最佳基因线路后对其配置进行随机更改的进化算法可成功开发出高性能的基因线路。Noman 等[46]提出一种基于蒙特卡洛的进化算法，即利用计算机对自然进化过程进行仿真，从而快速查找对噪音信息具有鲁棒性的网络拓扑(Network Topology)，这对于设计高鲁棒性的生命系统具有较高的价值。而 Hiscock 等[47]提出将机器学习中的梯度下降优化算法应用到基因线路的快速筛选和一系列不同功能的线路设计中。2021 年，Seak 等[48]尝试利用模拟人工神经网络的方法设计基因线路，进一步提升生物计算算法的潜力。

2.3 代谢工程

代谢工程最早由美国学者 Bailey[49]于 1991 年提出，是指用重组 DNA 技术有目的地改造中间代谢途径及网络，从而提高菌体生物量或代谢物产量。鉴于细胞代谢网络的复杂性，传统的设计通常整合了文献检索、代谢建模和启发式分析(Heuristic Analysis)等方法，但因为吞吐量有限，从数千个代谢反应及其调控网络等海量信息中找到合适的改造靶点非常困难。人工智能的集成建模方法有助于在代谢网络建模时兼顾动力学、调节作用、替代模型结构和参数集合等因素。例如，鲁棒性分析集成建模(Ensemble Modeling For Robustness Analysis，EMRA)将动态动力学模型与集成建模法结合以设计非天然代谢路径，可在选择代谢流改造靶点时既考虑模型性能又兼顾鲁棒性[50]。在大规模的代谢数据筛选中，机器学习平台作为高通量分析工具在促进数据驱动的目标生物合成途径优化和微生物产能提高方面得到了更广泛的应用[51-54]。EcoSynther 平台[55]使用反应数据库 Rhea 中约 10 000 条质量和电荷平衡的反应为外源反应数据源，并整合野生型大肠杆菌代谢网络模型中内源反应，利用途径搜索的概率分析算法模拟生产目标化合物的大肠杆菌菌株在不同生长条件下的整体代谢、目标化合物合成途径以及量化合成情况。将支持向量回归和前馈神经网络用于优化预测生产中核糖体结合位点和表型的关联，可将大肠杆菌中柠檬烯产量提高 60% 以上[56]。而将集成学习算法应用于 DBTL 循环数据可辅助提高大肠杆菌生产十二烷醇的效能(效价提高 21%)[22]。

合成生物学 DBTL 循环通常需要大规模采集和分析数据，且循环中往往受到实验成本高昂、可变性高、采样偏差以及传统数据分析方法局限性的限制。而自动化 DBTL 流程在微生物底盘生化途径的快速原型设计和优化应用中，集成了一系列独特的新技术组合，能大大降低实验成本和噪声，并且不依赖于研究人员对生物学机制的理解。Pablo 等[57]开发的 DBTL 平台使用计算机仿真选择候选酶，通过自动化元件设计，融合机器学习算法集优化技术指导和机器人辅助组装生化途径，随后进行快速测试和理性重设计，仅用两个 DBTL 循环就能大规模压缩可能的参数和变数组态(Configuration)数目，将大肠杆菌的类黄酮产量较以往报道的水平提高了 500 倍。Hamedirad 等[58]开发了一个耦合贝叶斯优化等机器学习算法的集成机器人平台——BioAutoMata，并用于 DBTL 循环优化番茄红素的生物合成途径。实验证明，仅测试不到 1% 的可能变异体就能发掘高产菌株，其产量超出随机筛选法选出的最优菌株产量的 77%。

由于不同微生物之间的差异，目标化合物的产量和合成途径也可能因底盘的不同而异。除了上述以大肠杆菌作为底盘，Zhou 等[23]基于人工

神经网络和 YeastFab 组装技术组合在酿酒酵母中优化外源代谢途径来提高目标代谢物的产量。此外，一种基于贝叶斯优化的自动推荐工具——ART（Automated Recommendation Tool）使得酵母中色氨酸的效价和生产率提升比例分别高达 74% 和 43%。该工具利用机器学习和概率建模技术以系统的方式指导合成生物学，而无需对生命系统有完整的理解[59]。Ding 等[60]开发的生物学推理系统 CF-Targeter 基于已有代谢反应库，利用途径搜索算法（Pathway-Searching Algorithm）对每个目标化合物执行 1 400 000 次搜索，可为指定的目标化合物选择合适的底盘。

2.4　基因组工程

随着基因测序、DNA 合成和基因编辑等技术的发展，合成生物学能对生物体的整个基因组甚至细胞进行工程改造，从而为直接探测基因型和表型之间的关系提供新工具，并为了解生物体基因组复杂功能体系提供一种全新的方式。在基因组工程领域，合成生物学与计算机技术的最早交互是通过一系列 Perl 脚本设计需改造的染色体序列及实现分层组装策略[61-63]。2018 年，Wang 等[64]提出使用计算机仿真自上而下地合成最小化基因组，利用混合整数线性规划（Mixed-Integer Linear Programming）标记已知的必需基因或导致显著适应性损失的基因，避免合成致死缺失，并在大肠杆菌中成功验证。

除了基因组合成外，基因组编辑、微生物组或群落的设计也涉及合成生物学与人工智能技术的交互。2018 年，DeepCRISPR 通过深度学习实现对 sgRNA 的靶点和靶点外预测，超越了其他软件工具的准确性，这将有助于实现高灵敏度和高特异性的 sgRNA 优化设计并应用于精准编辑基因组[65]。人工智能辅助合成生物学技术在调节肠道益生菌的治疗和营养方面也展现出一定价值。例如，将来自健康人群和肠道疾病患者的肠道微生物组的元基因组数据与机器学习算法（如逻辑回归、随机森林和支持向量机等）协同建模，可以更好地促进健康、免疫、消化、大脑功能等方面的研究[66]。2021 年，Karkaria 等[24]以合成生物学中的计算环路设计为基础，借助近似贝叶斯计算（Approximate Bayesian Computation）和蒙特卡洛采样法的模型选择和参数优化算法，提出了自动化合成微生物共生系统设计器，并构建稳定的双菌和三菌共生系统。该方法不但能给出构建稳定共生系统的基本设计原则，而且能揭示控制共生系统组成的关键参数。

3　人工智能与合成生物学交叉研究的关键瓶颈及未来方向

人工智能作为一门快速发展的新兴学科，其数学模型的训练主要基于数据驱动。然而，当前合成生物学研究存在数据来源广、数据形式异构、高质量训练数据不足等问题，这导致小数据稀疏监督下人工智能模型难以得到有效训练。鉴于生命系统极其复杂，很难用传统数学模型精确描述，当前技术仍无法有效预测复杂的基因线路。构建工程化平台是合成生物系统的重要研究手段，但当前工程化试错存在标准化的数据缺乏、海量的试错空间、定量的表征手段较少等问题，且智能化试错、优化、学习的理论支撑不足，工程化平台仍无法有效指导合成生物系统的设计与改造（图 2）。本小节将介绍人工智能技术与合成生物学的交叉研究在数据标准化、试错智能化、实验自动化、预测精准化方面存在的挑战。

3.1　数据标准化

合成生物工程自动化水平低，很大程度上受限于复杂的生命系统下用于人工智能模型训练的标准化数据[67]。例如，在生物信息系统中，转录调控和免疫信号转导网络数据通常存在类型不统一、有效数据缺乏和数据层次多等问题，且现有的 KEGG、GO 等公共数据库、公开文献数据

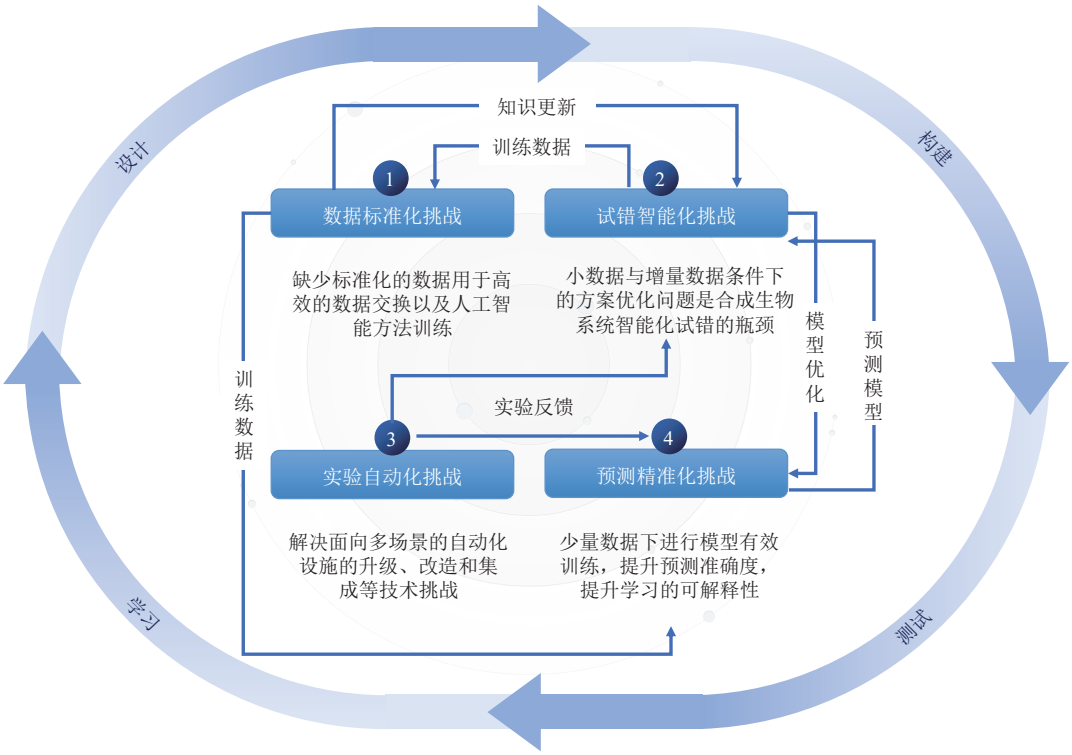

图 2 人工智能应用于合成生物学的挑战

Fig. 2 Challenges of applying AI in synthetic biology

及实验结果反馈的数据标准不统一，这需要研发构建多源融合的标准合成生物元件信息库的方法和技术，提供智能化查询、检索和推荐等功能[68-69]。高效利用公开数据库也是为机器学习算法提供训练数据的有效手段。在标准化数据的支持下，机器学习算法具有挖掘更多生物元件的潜力——采用生物信息学以及基因数据挖掘技术，从已有的元件库和未知微生物中挖掘更多的生物元件：结合生物学实验，将已有的生物元件作为输入，设计并训练机器学习模型，挖掘已有元件的模式，用于指导相应元件进行修饰、重组和改造，从而生成新的生物元件信息资源。然而，现实中存在着海量的还未发现的自然元件数据，这需要我们研发用于未知元件数据的自动化注释与标注的机器学习方法。

3.2 试错智能化

智能试错利用 DBTL 闭环中产生的数据，选择下一个迭代的实验设计，可以提高实验数据质量，减少估计误差[11]。上述过程适合利用强化学习[70]等优化决策理论框架进行建模，目标是输出累积奖励最高的实验设计序列。然而，由于合成生物实验通常时间跨度大、成本高，DBTL 迭代次数有限，可用于训练强化学习决策模型的数据极度不足。因此，解决小数据与增量数据条件下的方案优化问题[71]是合成生物系统设计、试错智能化的瓶颈问题。机器学习领域中一些小数据集下模型训练的理论框架具有应对上述挑战的潜力：分级强化的理念可减轻由于合成生物系统状态和可用改造手段的数量巨大，导致实验轨迹数据相对稀疏问题；生成对抗学习[72]框架产生高质量的实验轨迹可解决稀疏实验轨迹数据带来训练不足的问题；迁移学习框架也可复用已有相近源域的实验数据/模型，解决目标域由于稀疏实验轨迹数据无法有效训练设计策略模型的问题。将上述通用理论框架与合成生物领域场景相结合，可发展出一系列服务于试错智能化的新型机器学习

算法。

3.3　实验自动化

实验自动化旨在设计专用的人工智能技术以提高 DBTL 闭环中构建和测试两个环节的构建效率和测试质量。构建环节主要依赖于高灵活度的协议，优化构建规划与资源调度和提高自动化执行的能力。研究机器人[73-74]、不确定性环境下的优化规划[75]等人工智能技术可减少人工干预、提高构建的效率[18]。测试环节主要检验基因改造后细胞的行为是否符合预期。其中，最大的挑战是如何准确建立起基因型与表型之间的联系[18]。例如，定量地建立代表性真核细胞、原生生物、病毒基因型和表型(基因转录水平、蛋白表达量、小分子生成量、个体生存和功能水平)之间的关系。面向多场景的合成生物自动化设施的升级、改造和集成等给实验自动化带来了巨大的技术挑战。实现实验自动化可确保高通量的实验数据源源不断地进入 DBTL 闭环中，驱动循环，从而促使各个环节中机器学习方法提高性能。

3.4　预测精准化

由于合成生物系统复杂度高(可获取的数据极其复杂，通常具有数以万计的变量)，数据总量却严重不足，所以难以训练出一个高精度的机器学习模型[76]。迁移学习是在少量数据条件下通过迁移相关的两个或多个领域之间的知识结构进行模型有效训练的一种思路。例如，描述不同合成生物系统生物元件的基因水平上的调控信息、蛋白质水平上的相互作用和翻译后修饰信息[77-78]等，可在稀疏数据条件下提高预测准确性。此外，许多预测能力强的机器学习模型(图卷积神经网络等)存在“黑盒问题”，难以从生物学角度对模型输出进行解释，这阻碍了机器学习模型发现生物学内在机制的能力。合成生物应用存在大量的领域知识，通过融合机器学习模型与领域内知识可以更好地理解内部机制，提高预测的精准度[78]。而通过对生物内部机制的理解也可为建立全新的人工智能算法带来启发，如对进化生物学、脑科学和行为科学的研究启发了进化计算、人工神经网络以及强化学习等机器学习理论。合成生物系统中通过基因间的精密相互交互，动态形成调控网络，从而产出目标因子的工作方式，揭示了粗放型的传统机器学习模型——依赖大量数据、学习内在模式的方式已无法满足需求，亟需研究可精确融合领域知识的新型通用机器学习算法框架。

3.5　四大挑战间的联系

解决数据标准化、试错智能化、实验自动化、预测精准化四大挑战是相辅相成的。解决数据标准化挑战，建立起动态融合的知识库，可以作为其他三个方面开展的基础。其中，高通量实验数据的采集及智能试错技术进行优化，可为预测模型提供数据标准。而解决试错智能化的挑战则可在小数据稀疏监督下利用人工智能有效指导实验设计，提高元件库中新元件的挖掘效率以及标准化建库的质量；海量设计方案空间的优化探索，也可提高构建合成生物系统预测模型的效率。解决实验自动化挑战，实现高通量实验来增加训练数据总量，从源头上为智能试错算法和预测模型缓解小数据与稀疏监督的问题。解决预测精准化挑战，可根据基因型对合成生物系统表现型进行精准预测，以此显著提升强化学习模型策略效率，从而减少对真实实验数据的依赖。解决上述挑战可助力构建基于人工智能完成 DBTL 闭环的“类合成生物学家”智能体(图 3)，不断在循环过程中进行学习与试错优化，从而在数据标准化、实验自动化、预测精准化方面大大降低真实生物学实验的试错空间和成本。

4　总结与展望

人工智能与合成生物学交叉融合的研究工作仍处于发轫之始阶段：(1)常用于实现智能化元

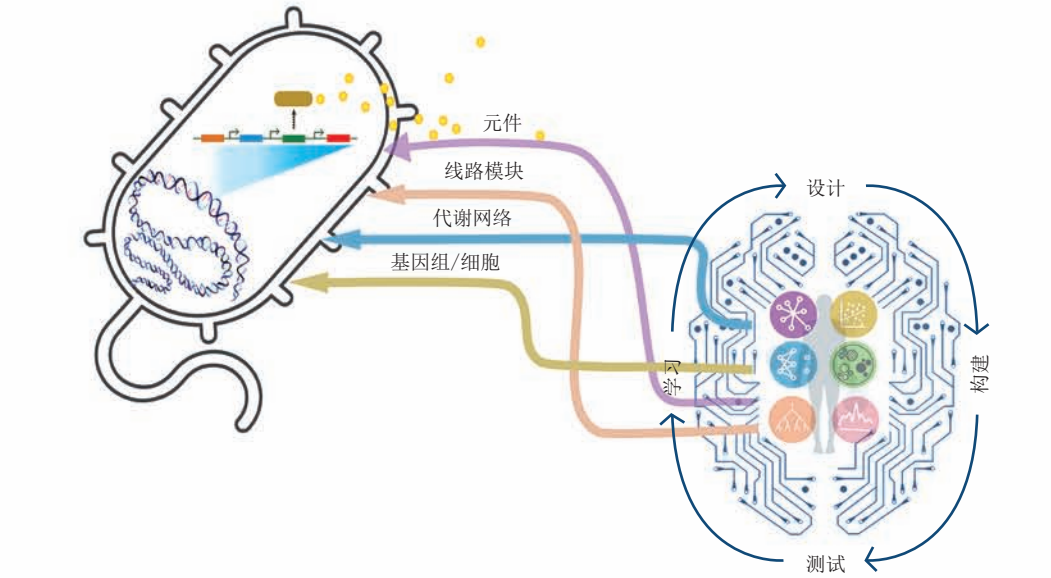

图 3 基于人工智能的“类合成生物学家”概念

Fig. 3 Concept of “AI synthetic biologist” based on artificial intelligence

件工程、线路工程、代谢工程和基因组工程的底盘生物仍局限于大肠杆菌和酿酒酵母；(2) 全基因组、微生物组或群落水平的智能化设计和合成仍寥寥无几；(3) 人工智能与合成生物学的融合多发生于 DBTL 循环的个别步骤，而 DBTL 全循环实现智能化的研究仍屈指可数。可喜的是，2020 年国家重点研发计划“合成生物学”专项立项名单中涌现了一批合成生物学与智能算法融合的项目，包括“基于合成生物学的新型活疫苗设计与开发”、“面向合成生物系统海量工程试错优化的人工智能算法研究与应用”、“数字细胞建模与人工模拟”、“新蛋白质元件人工设计合成及应用”、“正交化蛋白质复合物元件的人工设计构建与应用”等。值得关注的是，“面向合成生物系统海量工程试错优化的人工智能算法研究与应用”项目通过开发具有持续学习能力的自动化海量试错优化平台实现 DBTL 全循环智能化，利用人工智能的优势给工业合成生物学和医学合成生物学领域研究带来新的思路，并结合合成生物学的特色在微藻油脂细胞工厂、固有免疫细胞、人造噬菌体三大生物学应用场景下开展人工智能的算法研究。

受限于生命系统内部机理复杂以及合成生物实验周期长、成本高，以及适合训练人工智能方法的数据量极度不足，现有的机器学习方法均不足以支持高精度预测和实验设计优化。因此，研究小数据/零数据下的服务于海量工程试错的强化学习模型、具有生物可解释性的机器学习预测模型，可同时促进人工智能和合成生物学两大领域的发展。通过数据驱动及持续学习，“类合成生物学家”依照 DBTL 循环策略，部署多种基于人工智能的工具进行工程化的海量试错，可在快速合成具备目标功能的生命系统的同时孵化智能技术的革新。

参 考 文 献

[1] Bober JR, Beisel CL, Nair NU. Synthetic biology approaches to engineer probiotics and members of the human microbiota for biomedical applications [J]. Annual Review of Biomedical Engineering, 2018, 20(1): 277-300.

[2] Chien T, Doshi A, Danino T. Advances in bacterial

cancer therapies using synthetic biology [J]. Current Opinion in Systems Biology, 2017, 5: 1-8.

[3] Charlton Hume HK, Vidigal J, Carrondo MJ, et al. Synthetic biology for bioengineering virus-like particle vaccines [J]. Biotechnology and Bioengineering, 2019, 116(4): 919-935.

[4] Kitney RI, Bell J, Philp J. Build a sustainable vaccines industry with synthetic biology [J]. Trends in Biotechnology, 2021, 8: S0167-7799(0120)30331-30330.

[5] McCarty NS, Ledesma-Amaro R. Synthetic biology tools to engineer microbial communities for biotechnology [J]. Trends in Biotechnology, 2019, 37(2): 181-197.

[6] Weber W, Fussenegger M. Emerging biomedical applications of synthetic biology [J]. Nature Reviews Genetics, 2012, 13(1): 21-35.

[7] Tan X, Letendre JH, Collins JJ, et al. Synthetic biology in the clinic: engineering vaccines, diagnostics, and therapeutics [J]. Cell, 2021, 184(1): 881-898.

[8] Gao XY, Sun T, Pei GS, et al. Cyanobacterial chassis engineering for enhancing production of biofuels and chemicals [J]. Applied Microbiology and Biotechnology, 2016, 100(8): 1-13.

[9] Zhang W, Nielsen DR. Synthetic biology applications in industrial microbiology [J]. Frontiers in Microbiology, 2014, 5: 451.

[10] 饶聪, 云轩, 虞沂, 等. 微生物药物的合成生物学研究进展 [J]. 合成生物学, 2020, 1(1): 92-102.
Rao C, Yun X, Yu Y, et al. Recent progress of synthetic biology applications in microbial pharmaceuticals research [J]. Synthetic Biology Journal, 2020, 1(1): 92-102.

[11] Carbonell P, Radivojevic T, Martín H. Opportunities at the intersection of synthetic biology, machine learning, and automation [J]. ACS Synthetic Biology, 2019, 8(7): 1474-1477.

[12] Chen Y, Banerjee D, Mukhopadhyay A, et al. Systems and synthetic biology tools for advanced bioproduction hosts [J]. Current Opinion in Biotechnology, 2020, 64: 101-109.

[13] Dixon TA, Curach NC, Pretorius IS. Bio-informational futures: the convergence of artificial intelligence and synthetic biology [J]. EMBO Reports, 2020, 21(3): e50036.

[14] Freemont PS. Synthetic biology industry: data-driven design is creating new opportunities in biotechnology [J]. Emerging Topics in Life Sciences, 2019, 3(5): 651-657.

[15] Radivojevi T, Costello Z, Workman K, et al. A machine learning automated recommendation tool for synthetic biology [J]. Nature Communications, 2020, 11(1): 4879.

[16] Zhang JZ, Chen YC, Fu LH, et al. Accelerating strain engineering in biofuel research via build and test automation of synthetic biology [J]. Current Opinion in Biotechnology, 2021, 67: 88-98.

[17] 赵国屏. 合成生物学: 开启生命科学 “会聚” 研究新时代 [J]. 中国科学院院刊, 2018, 33(11): 1135-1149.
Zhao GP. Synthetic biology: unsealing the convergence era of life science research [J]. Bulletin of Chinese Academy of Sciences, 2018, 33(11): 1135-1149.

[18] Yaman F, Adler A, Beal J. AI challenges in synthetic biology engineering [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 7884-7886.

[19] Clauwaert J, Menschaert G, Waegeman W. DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns [J]. Nucleic Acids Research, 2019, 47(6): e36-e36.

[20] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers [J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 13996-14001.

[21] Myers CJ, Barker N, Jones K, et al. iBioSim: a tool for the analysis and design of genetic circuits [J]. Bioinformatics, 2009, 25(21): 2848-2849.

[22] Opgenorth P, Costello Z, Okada T, et al. Lessons from two design-build-test-learn cycles of dodecanol production in *Escherichia coli* aided by machine learning [J]. ACS Synthetic Biology,

2019, 8(6): 1337-1351.

[23] Zhou Y, Li G, Dong JK, et al. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae* [J]. Metabolic Engineering, 2018, 47: 294-302.

[24] Karkaria BD, Fedorec AJ, Barnes CP. Automated design of synthetic microbial communities [J]. Nature Communications, 2021, 12(1): 1-12.

[25] Cao RZ, Freitas C, Chan L, et al. ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network [J]. Molecules, 2017, 22(10): 1732.

[26] Kotopka BJ, Smolke CD. Model-driven generation of artificial yeast promoters [J]. Nature Communications, 2020, 11(1): 1-13.

[27] Romero PA, Krause A, Arnold FH. Navigating the protein fitness landscape with Gaussian processes [J]. Nature Methods, 2012, 110(3): E193-E201.

[28] Li RF, Wijma HJ, Lu S, et al. Computational redesign of enzymes for regio- and enantioselective hydroamination [J]. Nature Chemical Biology, 2018, 14(W1): 664-670.

[29] Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering [J]. Nature Methods, 2019, 16(8): 687-694.

[30] Bonde MT, Pedersen M, Klausen MS, et al. Predictable tuning of protein expression in bacteria [J]. Nature Methods, 2016, 13(3): 233.

[31] Wang Y, Wang H, Wei L, et al. Synthetic promoter design in *Escherichia coli* based on a deep generative network [J]. Nucleic Acids Research, 2020, 48(12): 6403-6412.

[32] Repecka D, Jauniskis V, Karpus L, et al. Expanding functional protein sequence spaces using generative adversarial networks [J]. Nature Machine Intelligence, 2021, 3(4): 324-333.

[33] Li T, Cui XX, Cui YL, et al. Exploration of transaminase diversity for the oxidative conversion of natural amino acids into 2-ketoacids and high-value chemicals [J]. ACS Catalysis, 2020, 10(14): 7950-7957.

[34] Atkinson MR, Savageau MA, Myers JT, et al. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli* [J]. Cell, 2003, 113(5): 597-607.

[35] Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli* [J]. Nature, 2000, 403(6767): 339-342.

[36] Elowitz MB, Liebler S. A synthetic oscillatory network of transcriptional regulators [J]. Nature, 2000, 403(1): 335-338.

[37] Tigges M, Marquez-Lago TT, Stelling J, et al. A tunable synthetic mammalian oscillator [J]. Nature, 2009, 457(7227): 309-312.

[38] Basu S, Gerchman Y, Collins C, et al. A synthetic multicellular system for programmed pattern formation [J]. Nature, 2005, 434(7037): 1130-1134.

[39] Liu C, Fu X, Liu L, et al. Sequential establishment of stripe patterns in an expanding cell population [J]. Science, 2011, 334(6053): 238-241.

[40] 娄春波, 杜沛, 孟凡康, 等. 人工基因线路的研究进展和未来挑战 [J]. 中国科学院院刊, 2018, 33(11): 1158-1165.
Lou CB, Du P, Meng FK, et al. Development and challenges of synthetic genetic circuits [J]. Bulletin of Chinese Academy of Sciences, 2018, 33(11): 1158-1165.

[41] Hasnain A, Sinha S, Dorfan Y, et al. A data-driven method for quantifying the impact of a genetic circuit on its host [C] // 2019 IEEE Biomedical Circuits and Systems Conference, 2019: 1-4.

[42] Green A, Silver P, Collins J, et al. Toehold switches: de-novo-designed regulators of gene expression [J]. Cell, 2014, 159(4): 925-939.

[43] Valeri JA, Collins KM, Ramesh P, et al. Sequence-to-function deep learning frameworks for engineered riboregulators [J]. Nature Communications, 2020, 11(1): 5058.

[44] Crombach A, Wotton KR, Cicin-Sain D, et al. Efficient reverse-engineering of a developmental gene regulatory network [J]. PLoS Computational Biology, 2012, 8(7): e1002589.

[45] Perkins TJ, Jaeger J, Reinitz J, et al. Reverse

engineering the *Gap* gene network of *Drosophila* melanogaster [J]. PLoS Computational Biology, 2006, 2(5): e51.

[46] Noman N, Monjo T, Moscato P, et al. Evolving robust gene regulatory networks [J]. PLoS One, 2015, 10(1): e0116258.

[47] Hiscock TW. Adapting machine-learning algorithms to design gene circuits [J]. BMC Bioinformatics, 2019, 20(1): 1-13.

[48] Seak L, Lo O, Suen CW, et al. Next-generation biocomputing: mimicking artificial neural network with genetic circuits [Z]. bioRxiv, 2021, DOI: 10.1101/2021.03.12.435120.

[49] Bailey JE. Toward a science of metabolic engineering [J]. Science, 1991, 252(5013): 1668-1675.

[50] Lee Y, Lafontaine Rivera JG, Liao JC. Ensemble modeling for robustness analysis in engineering non-native metabolic pathways [J]. Metabolic Engineering, 2014, 25: 63-71.

[51] Culley C, Vijayakumar S, Zampieri G, et al. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth [J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(31): 18869-18879.

[52] Presnell KV, Alper HS. Systems metabolic engineering meets machine learning: a new era for data-driven metabolic engineering [J]. Biotechnology Journal, 2019, 14(9): 1800416.

[53] Ramzi AB, Baharum SN, Bunawan H, et al. Streamlining natural products biomanufacturing with omics and machine learning driven microbial engineering [J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 608918.

[54] Zhou K, Ng W, Cortés-Pea Y, et al. Increasing metabolic pathway flux by using machine learning models [J]. Current Opinion in Biotechnology, 2020, 66: 179-185.

[55] Ding SZ, Liao XP, Tu WZ, et al. EcoSynther: a customized platform to explore biosynthetic potential in *E. coli* [J]. ACS Chemical Biology, 2017, DOI: 10.1021/acschembio.7b00605.

[56] Jervis AJ, Carbonell P, Vinaixa M, et al. Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli* [J]. ACS Synthetic Biology, 2019, 8(1): 127-136.

[57] Pablo C, Jervis AJ, Robinson CJ, et al. An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals [J]. Communications Biology, 2018, 1(1): 568-576.

[58] Hamedirad M, Chao R, Weisberg S, et al. Towards a fully automated algorithm driven platform for biosystems design [J]. Nature Communications, 2019, 10(1): 5150.

[59] Zhang J, Petersen SD, Radivojevic T, et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism [J]. Nature Communications, 2020, 11(1): 4880.

[60] Ding SZ, Cai PL, Yuan L, et al. CF-Targeter: a rational biological cell factory targeting platform for biosynthetic target chemicals [J]. ACS Synthetic Biology, 2019, 8(10): 2280-2286.

[61] Dai J, Boeke JD, Luo Z, et al. Sc3.0: revamping and minimizing the yeast genome [J]. Genome Biology, 2020, 21(1): 1215-1220.

[62] Hutchison III CA, Chuang RY, Noskov VN, et al. Design and synthesis of a minimal bacterial genome [J]. Science, 2016, 351(6280): aad6253.

[63] Richardson SM, Mitchell LA, Stracquadanio G, et al. Design of a synthetic yeast genome [J]. Science, 2017, 355(6329): 1040-1044.

[64] Wang L, Maranas CD. MinGenome: an *in silico* top-down approach for the synthesis of minimized genomes [J]. ACS Synthetic Biology, 2018, 7(2): 462-473.

[65] Chuai GH, Ma HH, Yan JF, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning [J]. Genome Biology, 2018, 19(1): 1-18.

[66] Kumar P, Sinha R, Shukla P. Artificial intelligence and synthetic biology approaches for human gut microbiome [J]. Critical Reviews in Food Science and Nutrition, 2020: 1-19.

[67] Carbonell P, Le Feuvre R, Takano E, et al. *In silico* design and automated learning to boost next-

generation smart biomanufacturing [J]. Synthetic Biology, 2020, 5(1): ysaa020.

[68] Li JQ, Deng GQ, Wei W, et al. Design of a real-time ECG filter for portable mobile medical systems [J]. IEEE Access, 2017, 5: 696-704.

[69] Li JQ, Huang LX, Zhou YM, et al. Computation partitioning for mobile cloud computing in a big data environment [J]. IEEE Transactions on Industrial Informatics, 2017, 13(4): 2009-2018.

[70] He Y, Wang YH, Qiu C, et al. Blockchain-based edge computing resource allocation in IoT: a deep reinforcement learning approach [J]. IEEE Internet of Things Journal, 2021, 8(4): 2226-2237.

[71] Chen J, Low KH, Yao YJ, et al. Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems [J]. IEEE Transactions on Automation Science and Engineering, 2015, 12(3): 901-921.

[72] Li JQ, Chen ZZ, Chen J, et al. Diversity-sensitive generative adversarial network for terrain mapping under limited human intervention [J]. IEEE Transactions on Cybernetics, 2020: 1-12.

[73] Li J, Deng G, Luo C, et al. A hybrid path planning method in unmanned air/ground vehicle (UAV/UGV) cooperative systems [J]. IEEE Transactions on Vehicular Technology, 2016, 65(12): 9585-9596.

[74] Li JQ, Hu SQ, Li QJ, et al. Global visual and semantic observations for outdoor robot localization [J]. IEEE Transactions on Network Science and Engineering, 2020, DOI: 10.1109/TNSE.2020.3045263.

[75] Chen J, Zhang YF, Li JQ, et al. Integrated air-ground vehicles for UAV emergency landing based on graph convolution network [J]. IEEE Internet of Things Journal, 2021, DOI: 10.1109/JIOT.2021.3058192.

[76] Camacho DM, Collins KM, Powers RK, et al. Next-generation machine learning for biological networks [J]. Cell, 2018, 173(7): 1581-1592.

[77] Ma LJ, Li JQ, Lin QZ, et al. Reliable link-inference for network data with community structures [J]. IEEE Transactions on Cybernetics, 2019, 49(9): 3347-3361.

[78] Teng T, Chen J, Zhang YH, et al. Scalable variational Bayesian kernel selection for sparse Gaussian process regression [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 5997-6004.