特约评述

DOI: 10.12211/2096-8280.2023-004

基于人工智能和计算生物学的合成生物学元件设计

王晟[1]，王泽琛[1,2]，陈威华[1]，陈珂[1]，彭向达[1]，欧发芬[1]，郑良振[1,3]，孙璠原[1,4]，沈涛[1]，赵国屏[3]
（[1] 上海智峪生物科技有限公司，上海 200030；[2] 山东大学，山东 济南 250100；[3] 中国科学院深圳先进技术研究院，广东 深圳 518055；[4] 中国科学院微生物研究所，北京 100101）

摘要：合成生物学是按照一定的规律综合已有的信息设计和构建全新的生物元件、装置和系统，或者重新设计已有的天然生物系统。合成生物学的核心在于设计、改造、重建或制造生物元件、生物反应系统、代谢途径与过程，乃至创造具有生命活动能力的细胞和生物个体，为解决人类发展在环境、资源、能源等方面面临的若干重大挑战提供新技术方案。毫无疑问，从DNA重组到基因电路设计，合成生物学的发展为众多领域带来全新的解决方案，优良的催化与调控元件是设计高效、鲁棒的系统的基础。然而，合成生物学的元件通常是天然的生物大分子，其固有的复杂性限制了对其工程化改造，导致合成生物技术的潜力未能得到充分发掘。随着人工智能（artificial intelligence，AI）与计算生物学的兴起和发展，有望助力该技术更好地发挥其价值。本文主要介绍了基于AI与计算生物学的不同类型的元件设计，聚焦催化元件、调控元件、传感元件三类元件的设计和前沿进展以及生物元件改造在合成生物学研究领域中的应用方面的研究进展。

关键词：人工智能；合成生物学；计算生物学；蛋白质设计；生物序列

中图分类号：Q812　　**文献标志码：**A

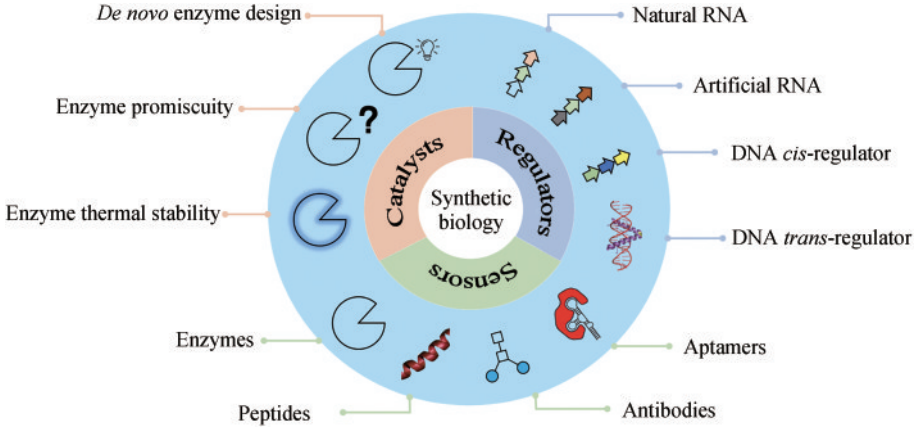
Design of synthetic biology components based on artificial intelligence and computational biology

WANG Sheng[1]，WANG Zechen[1,2]，CHEN Weihua[1]，CHEN Ke[1]，PENG Xiangda[1]，OU Fafen[1]，ZHENG Liangzhen[1,3]，SUN Jinyuan[1,4]，SHEN Tao[1]，ZHAO Guoping[3]
([1]*Shanghai Zelixir Biotech Company Ltd.*，*Shanghai 200030*，*China*；[2]*Shandong University*，*Jinan 250100*，*Shandong*，*China*；[3]*Shenzhen Institute of Advanced Technology*，*Chinese Academey of Sciences*，*Shenzhen 518055*，*Guangdong*，*China*；[4]*Institute of Microbiology*，*Chinese Academey of Sciences*，*Beijing 100101*，*China*)

Abstract: The primary objective of synthetic biology is to conceptualize, engineer, and construct novel biological components, devices, and systems based on established principles and extant information or to reconfigure existing natural biological systems. The core concept of synthetic biology encompasses the design, modification, reconstruction,

收稿日期：2023-01-11 修回日期：2023-04-03
基金项目：中国科学院国际大科学计划培育专项（153D31KYSB20170121）
引用本文：王晟，王泽琛，陈威华，陈珂，彭向达，欧发芬，郑良振，孙璠原，沈涛，赵国屏. 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
Citation：WANG Sheng，WANG Zechen，CHEN Weihua，CHEN Ke，PENG Xiangda，OU Fafen，ZHENG Liangzhen，SUN Jinyuan，SHEN Tao，ZHAO Guoping. Design of synthetic biology components based on artificial intelligence and computational biology[J]. Synthetic Biology Journal，2023，4(3)：422-443

or fabrication of biological components, reaction systems, metabolic pathways and processes, and even the creation of cells and organisms with functions or living characteristics. This burgeoning field offers innovative technologies to address challenges with sustainable development in environment, resource, energy, and so on. Undeniably, synthetic biology has yielded significant progress in numerous fields, ranging from DNA recombination to gene circuit design, yet its full potential remains insufficiently explored, but the emergence and application of artificial intelligence (AI) definitely can facilitate the development of synthetic biology for more applications. From a synthetic biology perspective, essence for life is rooted in digitalization and designability. This article reviews current advances in computational biology, particularly AI for synthetic biology to be more efficient and effective, focusing on the development of biocatalysts, regulators, and sensors. *De novo* enzyme design has been successfully implemented by using Rosetta software, as AI exhibiting significant potential for generating innovative structures and protein sequences with diverse functions. Also, the reprogramming of natural enzymes for specific purposes is crucial for synthetic biology applications. By employing various force fields and sampling techniques, promiscuity and thermal stability can be modified to accommodate specific requirements rather than those with natural hosts. AI can be integrated into the life-cycle of synthetic biology through an active learning paradigm, which enables alterations in enzyme specificity, and demonstrates potential for accurately and rapidly predicting mutation effects, surpassing force-field-based methods. The rapidly decreasing cost of sequencing has facilitated the characterization of *cis*-regulators, primarily DNA and RNA, with high-throughput. Concurrently, more trans-regulators have been identified in sequenced genomes. The expanding wealth in big data serves as a driving force for AI. AI models have successfully predicted the strength of promoters, ribosome binding sites (RBSs), and enhancers, and generated artificial protomers and RBSs. Recent progress in RNA structure prediction is expected to aid the design of RNA elements. Sensors, vital for genetic circuits and other applications such as toxin detection, typically involve interactions among various molecules, including nucleic acids, proteins, small organic molecules, and metal ions. Consequently, sensor design necessitates the integration of diverse computational biology tools to balance accuracy and computational cost. As the pool of data keeps growing, we anticipate that AI will be increasingly applied to the design of more bio-parts.



Keywords: artificial intelligence; synthetic biology; computational biology; protein design; biological sequences

合成生物学是利用工程学原理和技术来设计、构建和优化新的生物系统，从而实现具有特定功能和性能的人工生物体系。在合成生物学中，人工智能（artificial intelligence， AI）与计算生物学（computational biology）是两个重要的研究领域，它们可以为合成生物学提供强大的分析工具和模拟方法。下面是针对这些研究领域在合成生物学中的应用进行的系统分类和分析：

①元件设计和优化 在合成生物学中，元件设计和优化是非常重要的任务。元件设计可以通过人工智能或计算生物学加以实现，从而实现更好的生物表达和更高的生产效率。例如，可以使用AI算法来设计优化启动子序列，从而提高目标基因的表达水平、降低副反应等等。同时，计算生物学方法也可以用来模拟和预测酶在不同环境下的表达水平和功能。

②代谢工程和合成途径设计 代谢工程是利用基因工程技术来改造生物代谢途径，从而实现高效的生产过程。人工智能可以通过分析大量代谢数据，从而发现代谢网络中的关键酶和代谢路径。计算生物学方法也可以用来模拟和预测代谢途径在不同条件下的表现，从而帮助优化代谢工程的设计。

③模拟和预测生物系统 人工智能和计算生物学方法可以用来模拟和预测生物系统的行为和表现。例如，可以使用AI算法来预测基因调控网络的行为，并确定基因调控因子对基因表达的影响。计算生物学方法也可以用来模拟代谢途径的运行和产物生成。

④感知和控制 合成生物学中的感知和控制是指设计和构建生物感知器和生物控制器，以便实现对环境的感知和对生物行为的控制。人工智能和计算生物学方法可以用来设计和优化这些生物感知器和生物控制器。综上所述，人工智能和计算生物学在合成生物学中扮演了非常重要的角色，它们可以用来设计、优化和控制生物系统，以实现特定的功能和性能。这些技术的应用有助于推动生物技术和生物医药领域的发展。本文将重点集中讨论人工智能与计算生物学在合成生物学元件设计中的应用。

生物元件是合成生物学中基本要素之一，是合成生物学的基石[1]。生物元件是指遗传系统中最简单、最基本的生物积块（BioBrick），是具有特定功能的氨基酸或者核苷酸序列，可以在更大规模的设计中与其他元件进一步组合成具有特定生物学功能的生物学装置（device）。目前标准生物元件既包括启动子、终止子、转录单元、质粒骨架、接合转移元件、转座子、蛋白质编码区等DNA序列，也包括核糖体结合位点等RNA序列以及蛋白质结构域。目前，生物元件的挖掘鉴定和改造是合成生物学领域的一个重要研究方向（图1）。

随着计算生物学方法，尤其是相关人工智能技术的快速发展，在合成生物学领域中，使用计算方法进行合成生物学元件设计已经成为常用的工程设计思路[2]。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元件设计的需求。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

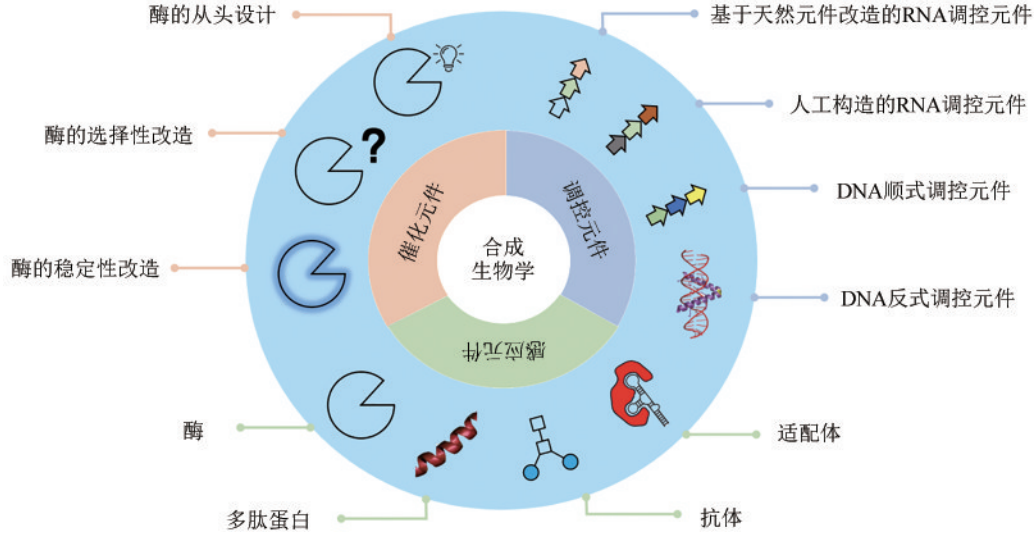


图1 基于人工智能和计算生物学的合成生物学元件设计

Fig. 1 Design for synthetic biology components based on artificial intelligence and computational biology

成生物学工程的效能[3]。计算生物学技术在合成生物学领域已经有广泛的应用：在催化元件设计领域，其应用主要包括酶的从头设计、酶的选择性改造、酶的稳定性改造；在感应元件设计领域，包括酶、多肽蛋白、抗体、适配体等的设计[4]；在调控元件领域，则包括一系列RNA调控元件（基于天然元件改造或人工构造的RNA调控元件）、DNA顺式调控元件（主要包括启动子、增强子、终止子、沉默子、绝缘子等）、DNA反式调控元件（编码转录因子的序列）等[5]。本文综述了近年来基于AI与计算生物学的不同类型的元件设计前沿进展，在此基础上提炼归纳AI与合成生物学两大领域交叉融合所面临的挑战，并对未来基于AI和计算生物学的合成生物学元件设计进行展望，有望为未来基于AI和生物计算的合成生物元件设计提供借鉴。

1 基于AI与计算生物学的催化元件设计

酶催化了生物体内的化学反应，对生命的关键过程如DNA复制、蛋白质合成、物质代谢和能量供给都至关重要。酶往往以催化元件的形式出现在合成生物学中。获取新酶是构建生命体中新反应、组合新途径进而合成新生命的基础。同时改造酶的选择性和稳定性是合成生物学在工业和医疗相关领域应用的迫切需求。随着基于AI与计算生物学的催化元件设计获得的突破性进展，合成生物学中催化元件的获取方式将发生根本性的变革。本节从酶的从头设计、选择性改造和稳定性改造三个方面，分别综述了传统计算生物学方法和基于人工智能（AI）方法的新进展。

1.1 基于计算生物学的催化元件设计

传统的计算生物学和生物物理方法主要包括分子对接、分子动力学（molecular dynamics simulation，MD）、以Rosetta[6]为代表的计算工具，以及基于进化信息的统计方法，被广泛用于蛋白质设计和酶工程。针对合成生物学中重要的催化元件（例如酶），利用这些方法已经发展了一些设计策略，成功进行了酶的从头设计、酶的选择性改造和稳定性提升。由于发展较早，已经有一些成熟的方法被开发成可公开使用的在线服务（表1）。

1.1.1 从头设计

从头设计可以得到催化自然界尚未观察到的生化反应的酶，这对扩展合成生物学的化学反应边界有重要的意义。天然酶可能会受到代谢网络进化产生的一些调控，从头设计这些酶也可以将其与底盘细胞解耦合。Kemp消除作为从碳上转移一个质子的模式反应［图2(a)］，其过程中被广泛研究。然而并没有发现一种能够催化该反应的酶，

表1 基于计算生物工具的在线服务器

Table 1 Online servers based on computational biology tools

名称	地址	用途与原理	参考文献
FireProt	https://loschmidt.chemi.muni.cz/fireprotweb/	通过结合进化的保守性和基于结构的能量计算，进行多点突变稳定性设计	[7]
GRAPE-WEB	https://nmdc.cn/grape-web/	结合基于物理能量和统计能量的蛋白质设计力场，进行单点突变稳定性设计，利用结构特征和实验结果使用机器学习算法组合突变分类	—
PROSS	https://pross.weizmann.ac.il/step/pross-terms/	基于Rosetta能量函数，设计可以提高稳定性的组合突变	[8]
Funclib	https://funclib.weizmann.ac.il/bin/steps	通过单点突变在进化中的概率和结构计算的能量筛选用于组合突变的候选，使用Rosetta能量函数评价组合突变的稳定性，设计突变组合突变改变底物谱、改善可溶性表达	[9]
ABACUS2	https://biocomp.ustc.edu.cn/servers/abacus-design.php	基于从蛋白质晶体结构中统计得到的能量函数，进行单点突变能量计算和序列骨架适配度打分	[10]
Swiss-Model	https://swissmodel.expasy.org/	基于序列搜索，以同源的晶体结构为模版，进行同源建模	[11]
ConSurf	https://consurf.tau.ac.il/consurf_index.php	基于序列搜索算法，分析保守性	[12]
ROSIE	https://rosie.graylab.jhu.edu/	基于Rosetta的结构relax，分子对接，突变预测等一系列任务	[13]

因此也被作为了一个从头设计酶催化剂的目标。针对Kemp消除酶的设计，David Baker课题组[14]首先设计了两套活性中心，分别使用天冬氨酸或谷氨酸作为广义碱，或使用组氨酸作为广义碱（与天冬氨酸形成二联体）作为催化位点。另外设计氢键供体来稳定中间体的酚羟基负电，额外设计了π-π堆积来形成电子离域，进一步稳定整个体系，利用量子力学（QM）计算优化了催化位点的构象，作为理论酶。在获得理论酶结构后，使用RosettaMatch计算工具，对PDB数据库中挑选出的部分适合作为骨架的蛋白尝试基于几何的骨架匹配，在完成匹配之后进行序列设计，进一步优化嫁接了活性位点后蛋白的序列，稳定催化构象，最终获得了8个具有可检测活性的Kemp消除酶[14]。这种先设计活性位点后设计骨架蛋白的方法被称为“inside-out”策略。使用这一策略，Baker课题组还针对非天然底物4-羟基-4-(6-甲氧基-2-萘基)-2-丁酮设计了催化Retro-Aldo反应的酶［图2(b)］。Retro-Aldo作为碳碳成键的关键反应，其催化过程较为复杂，但是使用“inside-out”策略仍然在72设计的序列中获得了32个具有可检测活性的酶[15]。

除了inside-out策略，Ranganathan课题组[16]还发展了基于统计模型的酶从头设计方法，设计了分支酸变位酶［图2(c)］。利用序列数据库中的天然分支酸变位酶序列的MSA进行直接耦合分析（direct coupling analysis， DCA），基于统计模型生成了一系列人工设计的序列，这些序列中有部分可以达到和天然酶相当的催化性能。

1.1.2 选择性改造

在当下的合成生物学应用中，根据需求恰当

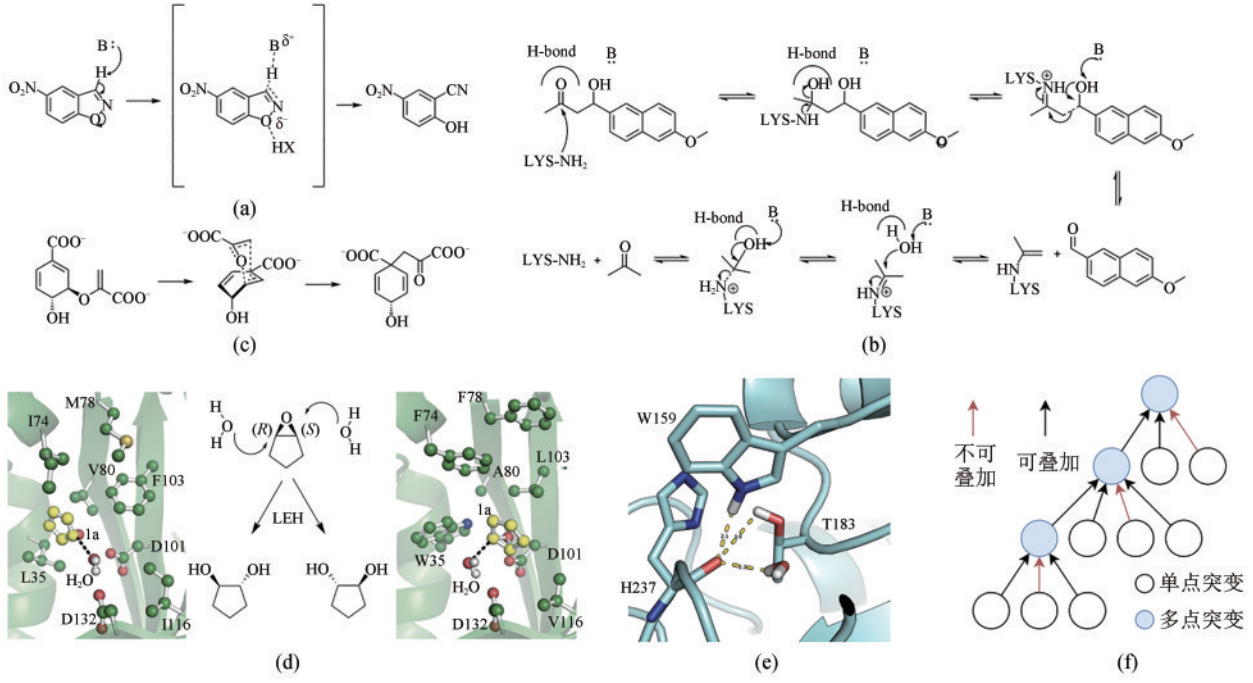


图2 基于计算生物学的催化元件设计

（a）Kemp消除反应机制；（b）Retro-Aldo反应机制；（c）分支酸变位酶催化机制；（d）柠檬烯环氧水解的两种不同的近攻击态构象示意图，左侧为pro-RR，右侧为pro-SS，结构示意图修改自文献［17］；（e）能量不利的未饱和氢键供体示意图，结构为*Is*PETase（PDB ID：5XJH），一个水分子和W159已经占据了H237的羰基可以形成的氢键（图中黄色虚线），T183的侧链羟基距离更远，难以形成氢键；（f）分组贪婪叠加策略示意图

Fig. 2 Design of the catalytic components based on computational biology

(a) Kemp elimination reaction mechanism. (b) Retro-Aldo reaction mechanism. (c) Mechanism of branching acid translocase catalysis. (d) Schematic diagram of two different near-attack state conformations of limonene epoxide hydrolysis with pro-RR on the left and pro-SS on the right, which were modified from reference [17] with permission. (e) Schematic diagram of the energetically unfavorable unsaturated hydrogen bond donor with the structure of IsPETase (PDB ID: 5XJH), where a water molecule and W159 already occupy the hydrogen bond that can be formed by the carbonyl group of H237 (yellow dashed line in the figure), and the side chain hydroxyl group of T183 is much further away and difficult to form a hydrogen bond. (f) Schematic diagram of the grouped greedy stacking strategy

改造已有天然酶成为底盘中合适的催化元件是一个重要的需求。基于计算方法的改造可以给出较小的突变体库，减少筛选的数量和时间，同时还可以实现较大的功能跃迁，也可以避免定向进化中可能出现的局部最优解。Janssen课题组[17]结合RosettaDesign、分子动力学模拟和对水分子进攻构象的限制［图2(d)］，实现了对柠檬烯环氧化物水解酶催化立体选择性的精确设计[17]。在计算的过程中，首先针对期望得到的立体选择性，放置底物在活性口袋内的静态的结构，使用RosettaDesign进行一轮序列设计来稳定活性中心的期望结构，再利用分子动力学模拟，增加对构象的采样，利用接近催化构象的几何构象出现频率定量预测立体选择性，最终获得了（*R*,*R*）和（*S*,*S*）对映体过量分别为85.5%和90.2%的突变体。天冬氨酸裂解酶（AspB）是已知选择性最严谨的酶之一，天然功能只能催化天冬氨酸的脱氨，也可催化其逆反应以富马酸和氨作为底物的氢胺化。吴边课题组[18]结合RosettaDesign针对不同底物设计序列和分子动力学模拟采样近似攻击态频率作为进一步的筛选条件，重新设计了天冬氨酸裂解酶AspB的底物谱，通过改造底物天冬氨酸的主链氨基结合区域，实现了针对多种不饱和酸的β-氢胺化反应[18]。

在此基础上，进一步改造了氨的结合口袋，获得了可以利用多种氨基供体和受体的一系列氢胺化酶[19]。Fleishman课题组[9]利用进化信息和Rosetta能量函数的打分提出了Funclib策略，实现了多样化的酶库设计。该方法利用MSA中获得的口袋位点保守性限制，作为序列组合的采样库，使用Rosetta能量函数对所有组合进行打分，选取打分能量稳定的进行实验表征，获得了能水解一系列神经毒素的酶库，且活性和可溶性表达相比于作为亲本的野生型均有提升。

1.1.3 稳定性改造

在和工业生物催化相关的应用和抗体等治疗性蛋白的应用中，往往需要蛋白质满足一定的稳定性。但是天然进化的酶往往不能够耐受高温、高盐或高浓度的有机溶剂，因此迫切需要快速提升稳定性且不损失其活性的方法。同时，改造酶的选择性等性质往往也需要一个稳定的骨架来容忍活性提升突变带来的可能的稳定性损失[20]。

计算设计通过给出少量的突变体库，易于表征，且往往能够通过少量突变大幅度提升温度稳定性。Damborsky课题组[21]提出了FireProt策略结合MSA中的保守性打分和Rosetta能量函数打分来设计更加稳定的组合突变体，成功将卤代烷脱卤酶的熔融温度（melting temperature，T_m）提升了25 ℃，将γ-六氯环己烷脱氯化氢酶的T_m提升了21 ℃。Janssan课题组[22]提出了FRESCO策略，通过利用FoldX[23]和Rosetta的ddg_monomer计算单点突变的折叠自由能变的差值（ΔΔ*G*），同时结合YASARA软件设计二硫键，随后利用MD检查突变前后突变区域的柔性，排除柔性显著增加的位点，将柠檬烯环氧化物水解酶的T_m提升了35 ℃，将肽酰胺水解酶的T_m提升了23 ℃，且大幅度提高了有机溶剂耐受性[24]。吴边课题组[25]通过结合三种基于结构的ΔΔ*G*计算工具（FoldX、Rosetta、ABACUS）和基于高温同源序列的保守性分析，辅助基于经验规则的突变结构目检，例如，不能在蛋白质非表面区域引入不参与氢键的氢键供体［图2(e)］，以排除计算结果中不合理的突变设计，设计了可能提高T_m的71个突变体，获得了21个T_m提升超过1.5 ℃的突变体，并且提出了分组贪婪叠加策略［图2(f)］，找到了可规避突变体中的负上位效应的叠加方法，成功改造了PET塑料水解酶*Is*PETase的温度稳定性，将T_m提升了31 ℃，大幅度延长了中温条件下塑料降解酶的半衰期[25]。此外，在另一个名为PROSS的算法中，研究人员通过在维持突变位点的共进化规律的同时通过Rosetta计算降低蛋白质ΔΔ*G*的情况下，可以有效提升蛋白质的热稳定性和可溶性[8]。以上这些研究指出了共进化信息和基于计算的折叠自由能变的差值对于设计提升酶的热稳定性和其他性质有非常好的指导意义。

1.2 基于机器学习的催化元件设计

机器学习和深度学习方法已经被广泛用于计算生物学领域，在蛋白质折叠、RNA结构预测、复合体亲和力预测以及各种生物图像分析中均有广泛应用。机器学习算法利用已有数据总结规律并对未知标签数据进行预测，例如贝叶斯模型、

逻辑回归、支持向量机、随机森林等算法也被广泛应用于生物数据建模。在已有数据（表2）的支持下，机器学习算法已经应用于合成生物学中催化元件的生成与改造。深度学习模型往往对大规模生物数据建模具有更大的优势，而且在合成生物学元件设计领域应用方兴未艾。

1.2.1 从头生成

深度学习模型在生成式任务上有一定的优势，在图像、语音、文本的合成上有广泛应用。对抗生成网络（generative adversarial network， GAN）作为一种生成式模型，通过同时训练生成样本的生成器和判别样本真假的判别器来达到生成以假乱真的样本的效果。Zelezniak课题组[35]提出了基于GAN的ProteinGAN模型，生成器和判别器均使用了卷积神经网络，利用注意力机制和膨胀卷积来捕捉序列的长距离依赖关系。将ProteinGAN在细菌来源的苹果酸脱氢酶序列数据上进行了训练，通过选取生成序列中与天然序列相似度为40%～100%的候选序列进行实验表征，鉴定了16条具有苹果酸脱氢酶活性的序列。相比于生成同家族的序列，按功能需求生成全新的酶更具有挑战性也有更广阔的应用空间。Baker课题组[36]结合在结构预测领域获得成功的RoseTTAFold，提出了RF_{joint}方法，即利用结构预测模型同时生成序列和结构。RF_{joint}提供了两种生成策略：第一种以随机序列为出发点，预测结构后计算和目标结构片段的损失函数，利用梯度更新序列或者利用蒙特卡洛采样序列，这种方式称为幻想（hallucination）；第二种通过在结构预测的训练任务上额外添加序列补全的任务，RF_{joint}能够在给定部分结构片段的情况下生成完整的蛋白质结构，这个过程称为修复（inpainting）。由于给定活性中心（几个氨基酸）生成完整的骨架（几百个氨基酸），导致了包含侧链构象的损失函数产生了过于崎岖的采样空间，使用“修复”和“幻想”方法均不能够很好地直接生成结构，因此采取了分步设计的策略，先借助不建模侧链的结构预测工具trRosetta生成主链空间[37]，再使用具有侧链生成能力的AlphaFold2生成了完整的序列和结构。作者测试了碳酸酐酶和Δ^5-3-酮类固醇异构酶，在结构预测测试中均显示了设计和预测结构的一致性。在已有的主链结构上，通过固定活性位点的氨基酸类型，以Baker课题组提出的ProteinMPNN为代表的序列生成模型也可以用于全新的催化元件设计[38]，这类方法生成的酶序列在AlphaFold的结构预测的测试上表现出较高的成功率。深度学习模型在催化元件的从头生成展现出了巨大的潜力，但是在未来的一段时间仍然需要更多的实验验证。

1.2.2 选择性改造

酶的选择性改造涉及酶蛋白和底物分子的相互作用，基于力场的方法往往能够捕捉局部的相互作用，做出正确的设计，然而为了减小计算量，力场通常会设置相互作用的截断半径，超过范围的突变产生的影响会被忽略，但是已经有大量的定向进化实验证明，远距离的突变能够显著影响选择性和活性，因此利用机器学习模型建模高阶相互作用的能力改造选择性能够克服传统方法的弊端。Shimizu课题组[39]提出了一种利用逻辑回

表2 可用于机器学习模型训练的数据库

Table 2 Databases available for machine learning model training

名称	地址	用途	参考文献
Protein Data Bank	https://www.rcsb.org/	蛋白质实验解析的结构数据库	[13]
UniProt	https://www.uniprot.org/	蛋白质序列数据库	[26]
AlphaFold DB	https://alphafold.ebi.ac.uk/	AlphaFold预测的蛋白质结构数据库	[27]
CATH	https://www.cathdb.info/	蛋白质结构域分类数据库	[28]
InterPro	https://www.ebi.ac.uk/interpro	蛋白质家族分类数据库	[29]
BRENDA	https://www.brenda-enzymes.org/	综合性酶学数据库	[30]
VariBench	http://structure.bmc.lu.se/VariBench/index.php	突变体实验测定数据库	[31]
Meltome	http://meltomeatlas.proteomics.wzw.tum.de:5003/	蛋白质熔融温度数据库	[32]
FireProt-DB	https://loschmidt.chemi.muni.cz/fireprotdb/	蛋白质点突变稳定性数据库	[33]
SABIO-RK	http://sabio.h-its.org/	酶动力学性质数据库	[34]

归和系统发育分析来估计每个氨基酸残基对底物特异性贡献的方法，以大肠杆菌的苹果酸脱氢酶作为模式酶，证明这种方法能够改变苹果酸脱氢酶对辅因子NAD^+和$NADP^+$的选择性。司同课题组[40]结合机器学习模型和贝叶斯优化，开发了BO-EVO，通过迭代的自动化实验与机器学习，高效搜索蛋白质组合突变空间，理论上将实验量大大降低。通过应用BO-EVO方法，将鼠李糖酯合酶RhlA对含有C_8底物的特异性提升了4.8倍。随着蛋白质-小分子相互作用模型的逐渐完善，可以期待机器学习模型未来在选择性改造上可以更好地利用目前大量的高精度蛋白质结构数据。

1.2.3 稳定性改造

目前有大量基于机器学习的蛋白质突变稳定性预测方法被开发出来，有部分算法已经接受了实验的检验并且获得了显著的成果。Ellington实验室[41]开发了基于3D卷积神经网络的MutCompute算法用于单点突变预测。MutCompute将三维空间划分为20×20×20的体素（voxel），用蛋白质的全原子模型填充体素对应的通道并将中心氨基酸排除在输入特征之外，训练模型预测输入中缺少的中心氨基酸的类型。Alper课题组[42]利用MutCompute改造了塑料降解酶PETase，改造后的Fast-PETase可以在50 ℃下48 h内将未经处理的PET包装盒几乎完全降解，并且实现了废弃塑料的降解后重新聚合。由于依赖结构的稳定性改造方法优于基于序列的方法，Fleishmank课题组[43]利用基于深度学习的trRosetta预测了过氧化物酶的结构，基于预测的结构应用前期开发的PROSS策略稳定且功能多样的多功能过氧化物酶。目前已经有比较多的基于深度学习和机器学习的稳定性预测模型，其中很多在一些数据集上接近或者超过了Rosetta和FoldX的$\Delta\Delta G$预测精度，但是这些算法在未知蛋白质热稳定性预测上的稳健性仍需要验证。

1.3 催化元件设计的挑战与展望

基于机器学习与计算生物学的合成生物学催化元件设计仍面临许多挑战。机器学习主要面临数据数量不足的问题：由于生化表征的昂贵性，目前可用于催化元件设计的数据量有限，需要利用高通量的实验方法获取更多的数据。此外，现有数据可能存在噪声和偏差，需要进行适当的预处理和清洗。计算生物学主要面临结构的多尺度计算问题：酶对催化的影响是多尺度的，分子力场通常只能描述到原子水平，对电子水平的影响是无法表现的。量子化学计算方法的计算代价过大，难以用于大量突变体的预先计算。

随着数据集和计算资源的不断增加，机器学习和计算生物学方法将成为合成生物学催化元件设计的有力工具。针对数据数量的问题，利用高通量实验增加数据量或利用预训练模型降低对数据量的需求。针对结构多尺度计算的问题，目前基于机器学习的分子动力学方法有望解决速度与精度的取舍问题[44]。这将促进催化元件的设计和合成，以实现更高效、更可持续和更具针对性的生物工程应用。

2 基于AI与计算生物学的调控元件设计

基因调控是基因表达时控制表达的基因类别及表达的时间、位置和表达量的过程。通过基因调控机制，功能上相关但表达方式各异的一组基因得以协调一致、共同表达。在生命体生长发育全过程中，这种协调表达能使生物更好地适应环境，维持生长发育。基因调控可以通过对DNA分子本身的修饰，以及对转录和翻译过程的控制来实现。天然基因调控机制复杂，涉及不同分子类型的多种元件，如启动子、增强子等DNA元件，以各种转录因子为代表的蛋白质元件，在某些机制中还涉及RNA元件，例如抑制翻译的反义RNA[45]和激活翻译的小RNA[46]。

2.1 DNA相关调控元件

DNA相关的调控元件可以分为顺式调控元件与反式调控元件。前者涉及启动子、增强子、终止子、绝缘子、沉默子等DNA元件；后者主要涉及转录因子。

2.1.1 DNA顺式调控元件

人类基因组包括2万个蛋白质编码基因，基因

突变会导致疾病的产生。然而，基因在人类基因组中的比重不到2%，基因组的大多数区域不编码蛋白质。曾经很长一段时间内，人们认为基因组的非编码区域是无用的，现在我们知道基因组的大部分非编码区域包含着不同的非编码调控元件（non-coding regulatory element，NCRE）。据统计，超过90%的疾病相关序列变异位于基因组的非编码部分，这表明了NCRE对细胞正常生理活动的重要性[47]。NCRE控制着基因的转录，目前一般认为，转录的起始阶段是基因表达全过程调控中最重要的环节。

DNA顺式调节模块（*cis*-regulatory module，CRM）是一段长度为100～1000个碱基对、可影响自身基因表达活性的DNA序列。CRM本身不参与任何蛋白质的编码，而是通过与转录因子结合来调节基因转录。它们包括启动子、增强子、终止子、沉默子、绝缘子以及其他参与调控基因表达的片段（图3）。由于它们通常与控制的基因位于同一DNA链上，因此被称为顺式调控元件。

随着基因组学技术的发展，研究者对顺式调控元件的探索逐渐深入，并逐步丰富调控元件的功能注释，表3中展示了一些收集顺式调控元件的数据库。

图3 顺式调控元件

Fig. 3 *Cis*-regulatory elements

然而，人类对基因组的探索仍然处于起步阶段。对于顺式调控元件在合成生物学中的应用而言，相比于从头设计，在天然细胞中对天然顺式作用元件的识别仍然是探索新调控元件的主要途径。通过实验识别调控元件序列、监测其活性强度及作用机制需要较高的时间和经济成本。由于DNA序列特征的高度多样性和转录调控的组织特异性，通过计算手段精确识别调控元件也是一项具有挑战性的任务[59]。2020年，Zrimec等[60]构建了一个深度学习网络，直接从天然DNA序列中预

表3 顺式调控元件相关的数据库

Table 3 Databases for *cis*-regulatory elements

名称	描述	链接	年份	参考文献
EPDNew	真核启动子数据库	http://epd.vital-it.ch/	2015	[48]
dbSUPER	包含小鼠和人类超级增强子信息	http://asntech.org/dbsuper/	2016	[49]
SEA	包含人类、小鼠等多种生物的超级增强子信息	http://sea.edbc.org/	2016	[50]
DiseaseEnhancer	包含143种人类疾病中的847种疾病相关的增强子信息	http://biocc.hrbmu.edu.cn/DiseaseEnhancer/	2018	[51]
HEDD	人类增强子疾病数据库，包含约280万人类增强子的基因组信息	https://zdzlab.einsteinmed.edu/1/hedd.php	2018	[52]
SEdb	人类超级增强子数据库，注释了超级增强子在基因调控中的功能	http://www.licpathway.net/sedb/	2019	[53]
PlantPAN3.0	从78种植物中收集了17 230个转录因子，部分包含结合位点信息	http://plantpan.itps.ncku.edu.tw/	2019	[54]
REDfly	包含实验验证的果蝇的CRM信息	http://redfly.ccr.buffalo.edu/	2019	[55]
EnhancerAtlas2.0	包含586种组织/细胞中的13 494 603个增强子	http://www.enhanceratlas.org/indexv2.php	2016	[56]
UCSC Genome Browser database	提供了人类、小鼠和SARS-Cov-2的基因组数据	http://genome.ucsc.edu	2021	[57]
SilencerDB	包含33 060个试验确定的沉默子和5 045 547个机器学习算法预测的沉默子	http://health.tsinghua.edu.cn/silencerdb/	2021	[58]

测基因表达水平，在7种生物体中实现了较高的准确性，表明了在原核生物和真核生物中，基因表达水平不是由单独的编码区域和顺式调控元件决定，而是由整个基因调控结构共同决定的。2021年，Umarov等提出了ReFeaFi方法，该方法包含两个深度学习模型，第一个模型用于扫描基因组并识别假定的调控区域，另一个模型确定转录起始位点，展现了深度学习在全基因组调控元件预测的潜力。2022年，Zrimec等[61]又设计了一个生成对抗网络（GAN）来生成预先指定的mRNA水平的DNA调控元件。值得肯定的是，机器学习尤其是深度学习在基因组学和合成生物学领域的应用，为调控元件的识别、活性强度预测、元件从头设计等方面提供了新思路。

启动子（promoter）序列是RNA聚合酶执行基因转录起始位置的DNA序列，启动子的强度或活性在调节基因转录中起着关键作用。当前，活细胞已经进化出许多具有一系列强度的启动子来微调关键基因的表达，从而实现特定的生理功能[62]。随着合成生物学的发展，人们开始依靠更强大的工具来控制转录过程，其中启动子是最基本的组成部分。基因表达的精确控制是发展从合成生物学到治疗学等多种应用的必要技术。尽管基因表达受到许多因素的控制，但启动子仍是基因转录的基本驱动因素。细胞工程通常会改造或设计启动子元件来控制基因表达[63]。在此过程中，对启动子活性的准确预测是进行成功设计的重要因素。设计准确的计算方法来预测启动子活性将为实验生物学家的研究提供指导。Meng等[62]基于人工神经网络和支持向量机设计了预测大肠杆菌DNA序列中启动子强度的机器学习模型；2019年，Oubounyt等[64]基于卷积神经网络（CNN）和长短期记忆网络（LSTM）构建了一个鲁棒的深度学习模型DeepPromoter，用于分析短真核启动子序列的特征，并准确识别人体和小鼠启动子序列；2020年，Wang等[65]提出了一个基于人工智能的大肠杆菌从头启动子设计框架，该模型以从天然启动子中学习到的序列特征为指导，可以捕获不同位置的核苷酸之间的相互作用，进而在计算机中设计新的启动子；2021年，Sudheer等[66]提出了一种新的计算模型，通过深度学习与伪二核苷酸（pseudo-dinucleotide）组成来识别原核启动子并预测其强度。

增强子（enhancer）是指导特定种类细胞转录的重要基因组调控元件[67]。当被转录因子结合时，其会增强相关基因的转录。增强子序列作用于同一DNA分子上的基因，但其具体位置通常不确定。比如，增强子序列可以位于距被调节基因的转录起始位点数千个碱基对的位置，但由于DNA在细胞核中折叠和盘绕，增强子实际上可能位于折叠状态下的转录起始位点附近。此外，增强子发挥作用与其序列的正反方向无关。在基因组学技术的推动下，人们意识到增强子和启动子之间有一些共同的特征和功能[68]。比如，它们的染色质和序列结构非常相似[69]，有的启动子还会有增强子活性[70]，有的增强子能够通过自身在其边界上驱动局部转录起始[71]。

增强子在发育和疾病过程中起着关键作用。由于人类基因组中的增强子控制着基因在特定种类细胞中的表达，因此其导致的变异存在引起疾病的风险[72-73]。由于对DNA序列和调控活性之间的关系知之甚少，增强子的从头设计一直具有挑战性。长期以来，识别增强子位置及其活性一直是科学研究的焦点。

2020年，Khanal等[74]基于word2vec和CNN从原始DNA序列中提取增强子的特征，开发了一个准确预测增强子活性的计算工具iEnhancer-CNN；2021年，Min等[75]设计了一个包含CNN和双向门控递归单元（Bi-GRU）的混合神经网络，仅使用DNA序列作为输入，实现了增强子-启动子相互作用预测；2022年，Almeida等[76]从黑腹果蝇S2细胞的DNA序列出发，结合深度学习算法构建了一个预测增强子活性的模型DeepSTAR。DeepSTARR在40 000个野生型和突变型的果蝇和人类增强子上进行了测试，证明其可推广到人类增强子的预测，最后，作者从头设计了特定活性的增强子并得到了验证[76]。

终止子（terminator）是位于基因编码区下游，能够给予RNA聚合酶转录终止信号的DNA序列。转录终止是基因表达的一个重要调控步骤。如果没有终止子，转录就不会停止，从而导致基因表达异常。根据作用机制，终止子可分为Rho因子依

赖性和非Rho因子依赖性。非Rho因子依赖性终止子富含GC碱基的反向重复序列，其转录出的mRNA可形成茎环结构，可以阻止RNA聚合酶的前进。Rho因子依赖的终止子没有形成强的茎环结构，因而不能自发终止转录，需要借助Rho因子来实现转录终止。Rho因子通过其解链酶的活性，强行解开转录泡上的RNA/DNA形成的杂交双螺旋，使RNA转录物得到释放，从而终止转录。

准确识别转录终止子在转录调控的研究和合成生物学应用中非常重要。2019年，Feng等[77]基于支持向量机（SVM）开发了一种识别转录终止子的预测模型iTerm-PseKNC，并在大肠杆菌和枯草芽孢杆菌的终止子上进行测试，证明了该模型可以成为细菌终止子识别的有力工具。

在真核生物基因组中，绝缘子（insulator）既是一种边界元件，又是一种控制基因表达的调控元件。绝缘子本身并不直接作用于基因的表达，其作用仅仅是不让其他调控元件对基因表达产生影响。随着生物的进化，DNA序列变得越来越复杂，绝缘子的出现能够将DNA序列划分为不同的区域，只有区域内的顺式作用元件可以调控该区域基因的表达，区域外的顺式作用元件则不能控制该区域内的基因表达。因此，绝缘子的出现能够保障基因免受无关调控元件的干扰。比如，当绝缘子位于增强子和启动子之间时，其可以保护启动子免受上游增强子的影响。此外，绝缘子具有阻断增强子与细胞癌基因调控元件相互作用的潜力，将有效抑制肿瘤的发生。这将为基因治疗方法在癌症、遗传病的应用提供重要启示。相比于对其他调控元件，绝缘子的识别和设计尚未成为研究的热点。

沉默子（silencer）是一段能够与转录调控因子结合的DNA序列，当抑制剂蛋白与DNA的沉默子区域结合时，将会阻止RNA聚合酶转录，从而阻止基因被表达为蛋白质。为保证细胞进行正常的生理活动，所有基因表达的激活和沉默必须保持微妙的平衡。尽管沉默子是非编码调节元件，但失去它们会导致发育缺陷，比如可能导致胚胎死亡。这也强调了基因组非编码区域的调节多功能性，并更好地解释了沉默子在基因调节框架中的重要性。此外，发现基因沉默的机制也可能为抑制肿瘤基因表达提供新视角，从而促进抗肿瘤领域的研究。但是，目前人们对沉默子的识别表征和作用规律的研究远远不及其他调控元件[47]。

2.1.2 DNA反式调控元件

不同于顺式调控元件，反式调控元件（*trans*-regulatory element，TRE）可位于基因组的任何地方。反式调控元件通过反式作用因子（*trans*-acting factor）介导实现对基因表达的干预，而反式作用因子通过与顺式调节元件相互作用以实现对结构基因表达的调节作用。尽管TRE突变影响基因表达，但它也是进化的主要驱动力。

反式作用因子又称转录因子（transcription factor，TF），是指能直接或间接地识别或结合在各类顺式作用元件核心序列上参与调控结构基因转录效率的蛋白质。反式作用因子通过与顺式作用元件相互作用，来参与基因表达的调控。真核生物中，按照反式作用因子的功能特性，可将其分为基本转录因子（general transcription factor）和特异转录因子（special transcription factor）。基本转录因子是RNA聚合酶结合启动子所必需的一组蛋白质因子，决定着三种RNA的转录类别。特异转录因子为个别基因转录所必需，决定该基因的时间、空间特异性表达。在不同类型的组织细胞中，会出现不同的特异转录因子。能够使某些基因表达增强的称为转录激活因子，比如增强子结合蛋白与增强子结合，使基因转录增强；使某些基因表达减弱的称为转录抑制因子，比如沉默子结合蛋白与沉默子结合，使某些基因的表达减弱。

随着基因组技术的发展，转录因子的实验数据库在不断扩展，并成为准确的计算方法开发的基础。表4展示了转录因子相关的数据库。

当前合成生物学相关方向对转录因子的研究尚集中于转录因子的识别和结合位点预测，这些工作是未来进行转录因子设计的基础和重要组成部分。近年来，越来越多的基于深度学习的方法被提出用于预测转录因子的结合位点，并获得了惊人的预测性能。比如2020年，Fu等[90]开发了一个深度学习预测框架scFAN，该框架不仅能预测转录因子的结合基序，还可用于分析单细胞表观基因组学和预测细胞类型；2021年，Kim等[91]提出了一个深度学习方法DeepTFactor，来预测蛋白质

表4 转录因子相关的数据库

Table 4 Databases for transcription factor

名称	描述	链接	年份	参考文献
GTRD	包含试验确定的人类和小鼠转录因子结合位点	http://gtrd.biouml.org/	2017	[78]
HOCOMOCO	包含人类和小鼠转录因子结合位点	https://hocomoco11.autosome.org/	2016	[79-80]
MeDReaders	包含人类和小鼠中731个转录因子与甲基化DNA序列结合的信息	http://medreader. org/	2018	[81]
TRRUST	人类TF-靶标相互作用数据库	https://www.grnpedia.org/trrust/	2015	[82-83]
AnimalTFDB 3.0	包含97个动物基因组中125 135个TF基因和80 060个转录辅助因子基因	http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/	2019	[84]
SalMotifDB	包含5个鲑鱼基因组中的转录因子及其顺式调控结合位点	https://salmobase.org/SalMotifDB/	2019	[85]
hTFtarget	包含人类转录因子及其靶体	http://bioinfo.life.hust.edu.cn/hTFtarget#!/	2020	[86]
KnockTF	包含人类组织/细胞中转录因子及其靶基因	http://www.licpathway.net/KnockTF/	2020	[87]
JASPAR 2022	包含真核生物转录因子结合位点	https://jaspar.genereg.net/	2022	[88]
PCRMS	提供了人类和小鼠基因组中CRM和转录因子结合位点的预测数据	https://cci-bioinfo.uncc.edu/	2022	[89]

是否为转录因子；同年，Zhang等[92]应用CNN设计了一个预测转录因子结合位点的方法FCNA；Zheng等[93]提出了一个机器学习框架AgentBind，用于预测基因组中某个转录因子基序在特定的细胞类型中是否结合，并识别关键的上下游碱基。这些方法的提出，也表明了机器学习和深度学习在当今大数据时代中，在解决生物问题上具有巨大潜力。

2.2 RNA调控元件

人们基于天然RNA调控元件设计了一些人工RNA调控元件，比较典型的有基于反义RNA，增强了结构稳定性的PTRNA元件[94]，以及源自大肠杆菌系统，利用结合蛋白Hfq稳定的基因抑制RNA元件[95]。这些调控元件使用模块化的结构，基于可变序列与目标mRNA的碱基互补配对实现对目标mRNA的特异性抑制。基于此原理的调控元件设计时序列自由度有限，但其序列特性（如靶向mRNA的位置及互补配对数的多少）可以和抑制效率形成一定的相关性。计算模型可用于评估这种相关性，以实现精细的抑制调控[96]。此外，还有另一种不依赖目标序列的RNA调控元件设计思路：在目标序列的适当位置引入人工的开/关模块，此等模块通常包含一段顺式感应序列和一段反式调控序列，基于感应序列和调控序列的特异性结合进行调控，典型例子有翻译激活调控的Toehold开关[97]和转录激活调控的STAR系统[98-100]。由于不依赖于目标序列，这一思路下调控元件的序列设计自由度很高，利用NUPACK[101]等基于最小自由能算法的RNA二级结构预测模型可以设计出高度正交的顺/反序列对。在大型的基因调控网络和复杂的合成生物电路中，元件的高正交、低串扰特性具有尤其重要的意义。

2.2.1 Toehold开关

Toehold开关由一对顺/反互补RNA序列组成，其中顺式的mRNA序列用于在目标mRNA的5′-UTR区引入人工发卡结构，并将目标mRNA的核糖体结合位点（RBS）包含在发卡结构的环区。未激活状态下，该人工结构阻止核糖体结合，抑制目标mRNA的翻译，因此顺式序列称为顺式抑制mRNA（*cis*-repressed mRNA， crRNA）。相应的反式互补序列则称为反式激活RNA（*trans*-activating RNA， taRNA），它特异性地与顺式抑制mRNA形成双链，使RBS暴露从而激活目标mRNA的翻译。利用RNA二级结构和自由能预测算法，Toehold开关的设计团队已经设计出了多组互相高度正交的低串扰组合，其中有18个组合的串扰低于2%，26个组合的串扰低于12%[97]。

2.2.2 STAR系统

STAR系统也包括一对顺反序列，不同于Toehold开关，STAR系统的顺式感应序列位于目标

mRNA的RBS上游，在转录时，这一段序列形成转录终止子，直接阻止目标mRNA的转录。而反式序列（称为STAR，small transcription activating RNA）与顺式终止子发卡的5′端及其上游序列互补。反式序列的配对会打开终止子发卡并允许转录延伸至下游的mRNA，从而实现目标mRNA的转录激活[99]。STAR系统的表现与反式STAR序列对顺式终止子发卡5′端上游线性单链的识别呈密切相关，在终止子发卡结构不变的情况下，改变该线性单链的序列即可改变STAR系统的泄露率和激活倍数[98]。STAR系统的激活倍数还和线性单链的二级结构程度呈反相关[98]。因此，RNA二级结构和自由能预测算法同样可以用来设计STAR系统[98]。

2.3 挑战和展望

调控元件一直是生物学研究的重要方向，获诺贝尔奖表彰的乳糖操纵子模型即其中开创性和代表性的研究成果，2000年发表的合成生物学基因电路开关的设计和表征工作也得益于DNA调控元件的成功运用[102]。由于生物调控元件固有的复杂性，传统的生物信息学方案表征的成功率有限。不过随着大量与调控元件相关的实验数据的积累，特别是算法的飞速进展，在处理复杂性问题上表现优越的深度学习算法在DNA顺式调控元件的识别、设计以及功能活性预测等方面都取得了长足的进步。计算方法能以极高的效率筛选极其庞大的序列空间，可以针对无法跨越细胞膜或具有细胞毒性的配体设计核糖开关，也可以利用生成式人工智能算法发展新的设计思路，这些算法的实现将弥补实验成本高、耗时长的局限，跨越式提升合成生物学工程化改造生物系统的能力。

值得关注的是，近年来RNA调控元件设计方面亦有较快进展，利用计算方法进行RNA调控元件设计也可以大大加快RNA调控元件的开发。基于RNA二级结构和自由能预测的计算模型已经在Toehold开关和STAR系统等人工调控系统的元件设计中有了成功的应用。由于当前计算领域对RNA结构和生物学属性认识的局限，目前的计算方法还只能基于RNA的序列和有限的二级结构开展设计，且可设计部分集中于模块化结构组装体的可变序列部分。未来随着计算领域对RNA三维结构特性的进一步理解，利用RNA的更高级结构可以设计出更加多样化的元件库，结合利用机器学习等方法建立调控效果的直接预测模型，RNA元件的计算设计将能发挥更加强大的威力。

3 基于计算生物学的感应元件设计

生物传感器（又称为感应元件）是一种用于检测被分析物的分析设备。顾名思义，生物传感器就是把生物成分和物理化学检测器结合在一起的一种设备，是由固定化的各种生物敏感材料作识别元件、适当的理化换能器（如氧电极、光敏管、场效应管、压电晶体等等）及信号放大装置构成的分析工具，其目的就是为了把待分析物质的种类、浓度等性质通过一系列的信号转化为能够容易被人们检测的量化数据，便于分析[103]。生物传感器与物理/化学传感器的主要区别在于生物传感器的识别元件是生物物质或者是仿生物物质。

生物传感器被广泛运用在医疗健康、食物质量检测、环境监测等方面（图4），特别适合用在需要经济有效的检测工具的场景。设计具有新功能特性或者新配体的生物传感器已快速发展成为一个新的生物医学和生物技术的分支领域。根据生物敏感元件制备来源的不同，生物传感器的特点各异。

3.1 基于不同生物元件的生物传感器设计

基于酶的生物传感器被认为是最适用的生物传感器分析工具之一，酶能够催化广泛不同的目标化合物的变化，目标化合物也可以通过抑制或者改变酶的催化活性而被检测到。这些基于酶的生物传感器具有独特的优点，如高特异性、选择性、可重复使用、低成本、易于制备和易于小型化。然而，它们也存在一些缺陷，如在苛刻的实验条件下不可恢复的酶的变性、储存困难、对样品基质的敏感性以及对温度和pH变化的敏感性[104]。

多肽和蛋白也是非常好的设计生物传感器的起始材料，因为它们在与目标分子相互作用的时候有着巨大的结构多样性，基于肽的生物传感器

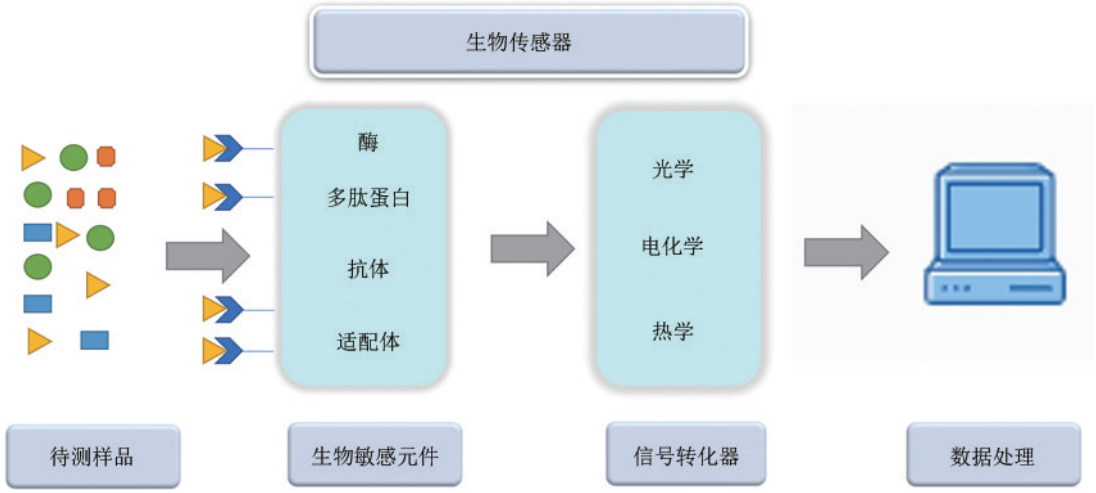


图 4　传感器设计

Fig. 4　Design of biosensor

具有优异的水溶性和生物相容性，使它们能够检测各种目标。不过，也有一些缺陷限制了它们的应用，例如耐久性低、灵敏度有限、在恶劣条件下会降解[105]。

基于抗体的生物传感器也常常用于检测多种目标，而且具有很高的灵敏度。但是基于抗体的生物传感器不耐温度变化，温度变化的时候容易降解变性失活。另外，抗体分子量较高，合成和修饰抗体是一个相对耗时的过程。需要用实验动物来产生筛选抗体，这增加了基于抗体的生物传感器的成本[106]。

核酸适配体是另外一种广泛应用的生物传感器。核酸适配体由单链核酸构成，对广泛的目标化合物（小到金属离子大到蛋白分子）都有很好的识别效果。与基于抗体的生物传感器相比，核酸适配体的优势在这些方面表现突出。核酸适配体可以通过体外的方式产生和进化，这样就不像抗体那样需要实验动物。核酸适配体分子量较小，这样生产和修饰相对容易。核酸适配体在广泛的温度和pH范围内都可以发挥功能，而且还可以在变性后容易重新折叠复性成有功能的状态。甚至有的时候它们可以在变性条件下发挥其功能，比如在高尿素的环境中[107]。核酸适配体一般经过SELEX的实验过程产生，随机核酸文库通过重复的结合、选择、扩增的过程得到亲和力越来越高的适配体。SELEX技术开发核酸适配体已经获得了很多成功案例，不过一般需要几周甚至几月的时间来进行富集筛选，多轮次的筛选工作也使得这项技术复杂和花费不少，而且有时候筛选到的适配体还要进行修饰来提高适配体的效率。另外为了提高SELEX技术的筛选效率，有时候需要准备一系列的不同类型的配体。这项技术的成功也依赖于初始文库的表示比例，有时候最高亲和力的适配体会因为不完整的文库而在筛选的过程中丢失[108]。

有各种实验技术方案可以将生物传感器的信号转化成可仪器测量的数据。这些技术包括表面等离子体共振（SPR）、荧光共振能量转移（FRET）、小角度X射线散射（SAXS）、电化学发光（ECL）、表面增强拉曼光谱（SERS）和光电化学（PEC）[103]。在过去的十年中，液晶（LC）在材料、化学和生物科学领域被广泛应用，因为它们有独特的物理和光学优势，对外部刺激的高反应性能、敏感的定向反应和光学各向异性[109-110]。生物传感器的液晶显示有如下显著的优势：容易制作，灵敏度高，反应快，成本低而且无需标记。它的缺陷是不能在高温环境中使用，也不适合在高光亮的环境中使用[111]。

3.2　基于计算生物学的生物传感器设计

除了实验方法外，包括量子力学（quantum mechanics，QM）、分子动力学（MD）和分子对接等在内的一系列计算方法被用于设计和研究生物传感器。生物传感器设计的计算机方法一般需要

准确的生物传感器片段的结构预测、热稳定性和分子间相互作用的精确描述。

量子力学方法可以在电子结构尺度上研究化学反应的热力学或动力学机理的方法。该方法可分为从头计算、密度泛函理论（DFT）和半经验算法等。由于研究对象是分子内的电子结构和运动，QM计算可以描述生物或化学分子的化学键的形成断裂、电荷转移等[112]。此能力是当前MD或分子对接等其他计算方法所不具备的。在生物传感器设计方面，使用QM的一个重要切入点是计算量子电导，其大小与生物传感器的选择性有着重要关联[113-114]。另外，由于QM方法具有高度的准确性，基于QM的相互作用能计算和溶剂化计算也被用于开发各种生物或化学传感器[115]。

限制QM方法在生物传感器领域大范围应用的主要因素是其高昂的计算成本。为了提高计算效率，人们开发了基于分子力场类的方法。分子力场假设分子的能量是分子内或分子间各个原子相对位置的函数。这些函数是经验性的，通常来自对QM结合和实验结果的拟合，其计算量要比QM类方法小数个数量级，但在适当的范围内，比如不涉及化学反应的大分子复杂体系，计算精度与QM是可比较的。以分子力场为基础的上层算法包括分子动力学模拟、蒙特卡洛方法、分子对接等等。

分子动力学（MD）是使用最为广泛的分子力场类方法，它通过迭代牛顿运动方程来模拟体系的构象在分子力场下随着时间的运动变化，以提供原子水平的简介。在足够的模拟时间和采样下，模拟产生的构象集合可用于机理分析[116]，几何形状、能量和许多物理化学特性的统计计算。配体和受体的结合能计算是生物传感器重要基础，基于MD技术开发的多种结合能计算方法已经在该领域取得了令人瞩目的进展，包括但不限于直接自由能计算[117]、平均力势（the potential mean force，PMF）方法[118]、MM-GBSA或MM-PBSA方法[119-122]。可靠的模拟结果往往需要高质量的采样，比如有人使用steered MD（SMD）研究生物受体与其靶标亲和力[123]。Thyparambil等[124-125]使用metadynamics类MD模拟设计了基于肽的霉菌毒素生物传感器。

分子对接（molecular docking）是一种通过搜索分子之间的空间匹配和能量匹配来预测配体与生物受体最可能匹配模式的方法。分子对接对配体-受体之间潜在相互作用的预测对于生物传感器准确性和特异性的改进是非常有帮助的[126]。相对于传统的MD方法，简化的势能函数和采样策略让分子对接具有高效的计算效率。分子对接可以用于筛选不同靶标或者不同的构象异构体[127]，也可以通过突变分析或位点饱和诱变（SSM）寻找甚至构建关键相互作用[128]，以实现高效的生物传感器。虚拟筛选（virtual screening）技术建立在分子对接技术的基础上，它将包含大量化合物的库与目标的三维模型进行对接。虚拟筛选既可以基于配体（ligand-based virtual screening）又可以基于受体（receptor-based virtual screening），已经被广泛应用于不同的生物系统[129]，包括筛选具有高结合亲和力的化合物用于生物传感器设计[130]，或者设计用作气态传感器的肽分子[131-132]。

上述各个方法均有着自己的优缺点和适用范围，表5中对各种方法进行了优缺点的总结，根据不同的问题，将不同方法组合起来可以更为有效地设计生物传感器。比如：分子对接可以为MD提供结合位点和结合模拟，MD则给出更精确的结合能计算[133-137]；QM和分子力场（MM）的结合（QM/MM）则可以用来计算核心区域的化学反应或电荷转移，MM的引入大大降低了QM的计算成本[138]，而QM/MM可提供更为准确的配合和受体之间的相互作用并评估[139-142]。该方法还可以应用于基于FRET的生物传感器的计算设计[143]。此外，上述计算方法与SELEX等实验技术的联用也取得了成功[144-146]，为设计生物传感器提供了有效策略。

3.3 生物传感器设计展望

随着合成生物学的快速发展，生物传感器受到了越来越多的关注，应用场景也在不断拓展。合成生物学里面的DBTL（Design—Build—Test—Learn）循环中的测试环节需要更高通量的检测方案，生物传感器实时快速检测的特点使其在合成基因回路、酶工程、代谢工程等领域中得到广泛关注[147-148]。别构转录因子来源的生物传感器在生

表5 计算方法比较

Table 5 Comparison of the computation methods

计算方法	优点	缺点	在生物传感器领域的应用
QM方法	精确、可以计算化学反应或电荷转移	计算成本高、难以直接应用于生物大分子的计算	高精度的评估结合能；计算化学反应（结合QM/MM）；计算量子电导
MD方法	在一定范围内有着可靠的精度，计算效率比QM高数个数量级	当电子运动不可忽略时，精度不够可靠；面临优化序列空间的问题是计算效率依然不够高	对生物传感器的作用机制或分子的构象变化做机理分析；评估配体和受体之间的亲和力；在一定范围内做序列优化
分子对接和虚拟筛选	高的计算效率	精度低于MD方法	高效评估配体和受体的结合状态和结合模式；从打数据库或序列空间寻找潜在的配体或受体待优化对象

物体内原本是发挥感受分子信号的功能作用，被开发研究得比较多[149-151]。同时，别构转录因子来源的生物传感器也可以像传统生物传感器那样在体外进行样品检测，CRISPR分子诊断技术与别构转录因子的结合使用是一个很有用的技术平台[152]。计算方法辅助生物传感器的设计方面，蛋白设计技术的进步，使得基于经典的四螺旋束蛋白框架的生物传感器的设计也取得了一些有应用价值的进展[153]，电荷基团导致的大规模结构相变原理上与别构转录因子的信号转化功能作用相似。可以预计的是，合成生物学的快速发展以及DBTL循环中对生物传感器的应用需求会导致越来越多有应用价值的技术方案和成功案例的出现。在经历了前期的各种有意义的生物学的技术方案的积累和发展后，计算机辅助的生物传感器的设计方案将会扮演越来越重要的角色。

4 总结和展望

目前，AI与计算生物学在合成生物学中的应用十分广泛，本文着重介绍了催化元件、调控元件和生物传感器的挖掘和设计中与AI和计算生物学结合的内容。受益于计算化学领域的前期发展，针对蛋白质、有机小分子的力场和采样工具都比较成熟，因此在设计和改造催化元件时，取得了较为显著的应用成果。机器学习方法的发展时间较短，一些方法还没有经受实验的广泛检验，但是也已经取得了一些成果，例如成功改造FAST-PETase用于塑料降解。针对特定的功能或性质，直接生成全新的酶或设计高阶突变体将是未来AI用于催化元件设计的发展方向之一。同时，受益于自动化实验设施的发展，AI将可以嵌入合成生物学的DBTL循环，进一步促进合成生物的整体发展。

随着基因组学技术的发展，对DNA序列非编码区域的功能研究不断深入，逐渐阐明了一些调控元件的功能和作用机制。这些调控元件虽然不直接参与基因的表达，但是对基因表达起调控作用。由于调控元件的突变会导致基因表达的变化，进而引起疾病的产生。因此，对于人类体内调控元件的识别、分类以及转录因子结合位点的分析，可以极大地促进人们对调控元件在基因表达、疾病产生中的作用的认识，为实验提供参考依据。随着调控元件实验数据的积累以及深度学习算法的开发，深度学习已经被应用于调控元件的识别、设计以及功能活性预测，而开发更有效的机器学习算法必将使现有基因组数据更丰富，促进基因组学的发展。

尽管取得了很大的成就，但在合成生物学中，人工智能和计算生物学的应用面临一些挑战和难点。下面是其中一些关键问题：

①复杂性和可扩展性。合成生物学中的生物系统往往非常复杂，包括多个基因、酶和代谢途径。这使得优化这些系统的设计和控制变得非常困难。同时，需要考虑生物系统的可扩展性，因为在实际应用中需要生产大量的产品，需要保证系统能够扩展到大规模生产。

②数据处理和模型构建。人工智能和计算生物学需要大量的数据支持，但生物学数据通常非常复杂，需要处理大量的噪声和变异。同时，需要建立准确的模型来描述生物系统的行为和功能，这需要对生物系统的深入理解和对建模技术的

掌握。

③生物实验和数据采集：合成生物学需要进行大量的生物实验和数据采集，以测试和验证生物系统的性能。这需要大量的时间和资源，同时还需要考虑如何最大限度地减少实验成本和提高实验效率。

④安全和规范：合成生物学中的新生物系统可能对环境和人类健康造成潜在威胁，因此需要制定相关的安全规范和评估方法，以确保生物系统的安全性和可控性。

因此，人工智能和计算生物学需要解决上述难点和关键问题，以实现合成生物学的进一步发展和应用。这需要多个学科领域之间的紧密合作和交叉创新，同时需要不断发展和完善相关技术和方法。

参 考 文 献

[1] LV X Q, HUESO-GIL A, BI X Y, et al. New synthetic biology tools for metabolic control[J]. Current Opinion in Biotechnology, 2022, 76: 102724.

[2] FAULON J L, FAURE L. *In silico*, *in vitro*, and *in vivo* machine learning in synthetic biology and metabolic engineering[J]. Current Opinion in Chemical Biology, 2021, 65: 85-92.

[3] LAWSON C E, MARTÍ J M, RADIVOJEVIC T, et al. Machine learning for metabolic engineering: a review[J]. Metabolic Engineering, 2021, 63: 34-60.

[4] MADHAVAN A, ARUN K B, BINOD P, et al. Design of novel enzyme biocatalysts for industrial bioprocess: harnessing the power of protein engineering, high throughput screening and synthetic biology[J]. Bioresource Technology, 2021, 325: 124617.

[5] DE JONGH R P H, VAN DIJK A D J, JULSING M K, et al. Designing eukaryotic gene expression regulation using machine learning[J]. Trends in Biotechnology, 2020, 38(2): 191-201.

[6] ALFORD R F, LEAVER-FAY A, JELIAZKOV J R, et al. The Rosetta all-atom energy function for macromolecular modeling and design[J]. Journal of Chemical Theory and Computation, 2017, 13(6): 3031-3048.

[7] MUSIL M, STOURAC J, BENDL J, et al. FireProt: web server for automated design of thermostable proteins[J]. Nucleic Acids Research, 2017, 45(W1): W393-W399.

[8] GOLDENZWEIG A, GOLDSMITH M, HILL S E, et al. Automated structure- and sequence-based design of proteins for high bacterial expression and stability[J]. Molecular Cell, 2016, 63(2): 337-346.

[9] KHERSONSKY O, LIPSH R, AVIZEMER Z, et al. Automated design of efficient and functionally diverse enzyme repertoires[J]. Molecular Cell, 2018, 72(1): 178-186.e5.

[10] XIONG P, HU X H, HUANG B, et al. Increasing the efficiency and accuracy of the ABACUS protein sequence design method[J]. Bioinformatics, 2020, 36(1): 136-144.

[11] SCHWEDE T, KOPP J, GUEX N, et al. SWISS-MODEL: an automated protein homology-modeling server[J]. Nucleic Acids Research, 2003, 31(13): 3381-3385.

[12] ASHKENAZY H, EREZ E, MARTZ E, et al. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids[J]. Nucleic Acids Research, 2010, 38(suppl_2): W529-W533.

[13] MORETTI R, LYSKOV S, DAS R, et al. Web-accessible molecular modeling with Rosetta: the Rosetta Online Server that Includes Everyone (ROSIE)[J]. Protein Science, 2018, 27(1): 259-268.

[14] RÖTHLISBERGER D, KHERSONSKY O, WOLLACOTT A M, et al. Kemp elimination catalysts by computational enzyme design[J]. Nature, 2008, 453(7192): 190-195.

[15] JIANG L, ALTHOFF E A, CLEMENTE F R, et al. *De novo* computational design of Retro-Aldol enzymes[J]. Science, 2008, 319(5868): 1387-1391.

[16] RUSS W P, FIGLIUZZI M, STOCKER C, et al. An evolution-based model for designing chorismate mutase enzymes[J]. Science, 2020, 369(6502): 440-445.

[17] WIJMA H J, FLOOR R J, BJELIC S, et al. Enantioselective enzymes by computational design and *in silico* screening[J]. Angewandte Chemie International Edtion, 2015, 54(12): 3726-3730.

[18] LI R F, WIJMA H J, SONG L, et al. Computational redesign of enzymes for regio- and enantioselective hydroamination[J]. Nature Chemical Biology, 2018, 14(7): 664-670.

[19] CUI Y L, WANG Y H, TIAN W Y, et al. Development of a versatile and efficient C-N lyase platform for asymmetric hydroamination *via* computational enzyme redesign[J]. Nature Catalysis, 2021, 4(5): 364-373.

[20] MENG Q L, CAPRA N, PALACIO C M, et al. Robust ω-*trans*-aminases by computational stabilization of the subunit interface[J]. ACS Catalysis, 2020, 10(5): 2915-2928.

[21] BEDNAR D, BEERENS K, SEBESTOVA E, et al. FireProt: energy- and evolution-based computational design of thermostable multiple-point mutants[J]. PLoS Computational Biology, 2015, 11(11): e1004556.

[22] WIJMA H J, FLOOR R J, JEKEL P A, et al. Computationally designed libraries for rapid enzyme stabilization[J]. Protein Engineering, Design and Selection, 2014, 27(2): 49-58.

[23] DELGADO J, RADUSKY L G, CIANFERONI D, et al. FoldX 5.0: working with RNA, small molecules and a new graphical interface[J]. Bioinformatics, 2019, 35(20): 4168-4169.

[24] WU B, WIJMA H J, SONG L, et al. Versatile peptide C-terminal functionalization *via* a computationally engineered peptide amidase[J]. ACS Catalysis, 2016, 6(8): 5405-5414.

[25] CUI Y L, CHEN Y C, LIU X Y, et al. Computational redesign of a PETase for plastic biodegradation under ambient condition by the GRAPE strategy[J]. ACS Catalysis, 2021, 11(3): 1340-1350.

[26] THE UNIPROT CONSORTIUM. UniProt: a worldwide hub of protein knowledge[J]. Nucleic Acids Research 2019, 47(D1), D506-D515.

[27] VARADI M, ANYANGO S, DESHPANDE M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models[J]. Nucleic Acids Research, 2022, 50(D1): D439-D444.

[28] ORENGO C A, MICHIE A D, JONES S, et al. CATH—a hierarchic classification of protein domain structures[J]. Structure, 1997, 5(8): 1093-1108.

[29] BLUM M, CHANG H Y, CHUGURANSKY S, et al. The InterPro protein families and domains database: 20 years on[J]. Nucleic Acids Research, 2021, 49(D1): D344-D354.

[30] CHANG A, JESKE L, ULBRICH S, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates[J]. Nucleic Acids Research, 2021, 49(D1): D498-D508.

[31] SARKAR A, YANG Y, VIHINEN M. Variation benchmark datasets: update, criteria, quality and applications[J]. Database, 2020, 2020: baz117.

[32] JARZAB A, KURZAWA N, HOPF T, et al. Meltome atlas—thermal proteome stability across the tree of life[J]. Nature Methods, 2020, 17(5): 495-503.

[33] STOURAC J, DUBRAVA J, MUSIL M, et al. FireProt^{DB}: dataBase of manually curated protein stability data[J]. Nucleic Acids Research, 2021, 49(D1): D319-D324.

[34] WITTIG U, REY M, WEIDEMANN A, et al. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics[J]. Nucleic Acids Research, 2018, 46(D1): D656-D660.

[35] 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.

[36] WANG J, LISANZA S, JUERGENS D, et al. Scaffolding protein functional sites using deep learning[J]. Science, 2022, 377 (6604): 387-394.

[37] ANISHCHENKO I, PELLOCK S J, CHIDYAUSIKU T M, et al. *De novo* protein design by deep network hallucination[J]. Nature, 2021, 600(7889): 547-552.

[38] DAUPARAS J, ANISHCHENKO I, BENNETT N, et al. Robust deep learning-based protein sequence design using ProteinMPNN[J]. Science, 2022, 378(6615): 49-56.

[39] SUGIKI S, NIIDE T, TOYA Y, et al. Logistic regression-guided identification of cofactor specificity-contributing residues in enzyme with sequence datasets partitioned by catalytic properties[J]. ACS Synthetic Biology, 2022, 11(12): 3973-3985.

[40] HU R Y, FU L H, CHEN Y C, et al. Protein engineering *via* Bayesian optimization-guided evolutionary algorithm and robotic experiments[J]. Briefings in Bioinformatics, 2023, 24(1): bbac570.

[41] SHROFF R, COLE A W, DIAZ D J, et al. Discovery of novel gain-of-function mutations guided by structure-based deep learning[J]. ACS Synthetic Biology, 2020, 9(11): 2927-2935.

[42] LU H Y, DIAZ D J, CZARNECKI N J, et al. Machine learning-aided engineering of hydrolases for PET depolymerization[J]. Nature, 2022, 604(7907): 662-667.

[43] BARBER-ZUCKER S, MINDEL V, GARCIA-RUIZ E, et al. Stable and functionally diverse versatile peroxidases designed directly from sequences[J]. Journal of the American Chemical Society, 2022, 144(8): 3564-3571.

[44] DOERR S, MAJEWSKI M, PÉREZ A, et al. TorchMD: a deep learning framework for molecular simulations[J]. Journal of Chemical Theory and Computation, 2021, 17(4): 2355-2363.

[45] GREEN P J, PINES O, INOUYE M. The role of antisense RNA in gene regulation[J]. Annual Review of Biochemistry, 1986, 55: 569-597.

[46] PAPENFORT K, VANDERPOOL C K. Target activation by regulatory RNAs in bacteria[J]. FEMS Microbiology Reviews, 2015, 39(3): 362-378.

[47] PANG B X, VAN WEERD J H, HAMOEN F L, et al. Identification of non-coding silencer elements and their regulation of gene expression[J]. Nature Reviews Molecular Cell Biology, 2022: 1-13.

[48] DREOS R, AMBROSINI G, PÉRIER R C, et al. The eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools[J]. Nucleic Acids Research, 2015, 43(Database issue): D92-D96.

[49] KHAN A, ZHANG X G. dbSUPER: a database of super-enhancers in mouse and human genome[J]. Nucleic Acids Research, 2016, 44(D1): D164-D171.

[50] WEI Y J, ZHANG S M, SHANG S P, et al. SEA: a super-enhancer archive[J]. Nucleic Acids Research, 2016, 44(D1): D172-D179.

[51] ZHANG G X, SHI J, ZHU S W, et al. DiseaseEnhancer: a resource of human disease-associated enhancer catalog[J]. Nucleic Acids Research, 2018, 46(D1): D78-D84.

[52] WANG Z, ZHANG Q W, ZHANG W, et al. HEDD: human enhancer disease database[J]. Nucleic Acids Research, 2018, 46 (D1): D113-D120.

[53] JIANG Y, QIAN F C, BAI X F, et al. SEdb: a comprehensive human super-enhancer database[J]. Nucleic Acids Research, 2019, 47(D1): D235-D243.

[54] CHOW C N, LEE T Y, HUNG Y C, et al. PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants[J]. Nucleic Acids Research, 2019, 47(D1): D1155-D1163.

[55] RIVERA J, KERÄNEN S V E, GALLO S M, et al. REDfly: the transcriptional regulatory element database for Drosophila[J]. Nucleic Acids Research, 2019, 47(D1): D828-D834.

[56] GAO T S, HE B, LIU S, et al. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types[J]. Bioinformatics, 2016, 32(23): 3543-3551.

[57] GONZALEZ J N, ZWEIG A S, SPEIR M L, et al. The UCSC genome browser database: 2021 update[J]. Nucleic Acids Research, 2021, 49(D1): D1046-D1057.

[58] ZENG W W, CHEN S Q, CUI X J, et al. SilencerDB: a comprehensive database of silencers[J]. Nucleic Acids Research, 2021, 49(D1): D221-D228.

[59] UMAROV R, LI Y, ARAKAWA T, et al. ReFeaFi: genome-wide prediction of regulatory elements driving transcription initiation[J]. PLoS Computational Biology, 2021, 17(9): e1009376.

[60] ZRIMEC J, BÖRLIN C S, BURIC F, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure[J]. Nature Communications, 2020, 11: 6141.

[61] ZRIMEC J, FU X Z, MUHAMMAD A S, et al. Controlling gene expression with deep generative design of regulatory DNA[J]. Nature Communications, 2022, 13: 5099.

[62] MENG H L, MA Y F, MAI G Q, et al. Construction of precise support vector machine based models for predicting promoter strength[J]. Quantitative Biology, 2017, 5(1): 90-98.

[63] CAZIER A P, BLAZECK J. Advances in promoter engineering: novel applications and predefined transcriptional control[J]. Biotechnology Journal, 2021, 16(10): e2100239.

[64] OUBOUNYT M, LOUADI Z, TAYARA H, et al. DeePromoter: robust promoter predictor using deep learning[J]. Frontiers in Genetics, 2019, 10: 286.

[65] WANG Y, WANG H C, 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.

[66] MENON S, PIRAMANAYAKAM S, AGARWAL G. Computational identification of promoter regions in prokaryotes and eukaryotes[J]. EPRA International Journal of Agriculture and Rural Economic Research, 2021, 9(7): 21-28.

[67] CATARINO R R, STARK A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation[J]. Genes & Development, 2018, 32(3/4): 202-223.

[68] ANDERSSON R, SANDELIN A. Determinants of enhancer and promoter activities of regulatory elements[J]. Nature Reviews Genetics, 2020, 21(2): 71-87.

[69] ANDERSSON R, REFSING ANDERSEN P, VALEN E, et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species[J]. Nature Communications, 2014, 5: 5336.

[70] DIAO Y R, FANG R X, LI B, et al. A tiling-deletion-based genetic screen for *cis*-regulatory element identification in mammalian cells[J]. Nature Methods, 2017, 14(6): 629-635.

[71] KIM T K, HEMBERG M, GRAY J M, et al. Widespread transcription at neuronal activity-regulated enhancers[J]. Nature, 2010, 465(7295): 182-187.

[72] FULCO C P, NASSER J, JONES T R, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations[J]. Nature Genetics, 2019, 51(12): 1664-1669.

[73] MAURANO M T, HUMBERT R, RYNES E, et al. Systematic localization of common disease-associated variation in regulatory DNA[J]. Science, 2012, 337(6099): 1190-1195.

[74] KHANAL J, TAYARA H, CHONG K T. Identifying enhancers and their strength by the integration of word embedding and convolution neural network[J]. IEEE Access, 2020, 8: 58369-58376.

[75] MIN X P, YE C M, LIU X R, et al. Predicting enhancer-promoter interactions by deep learning and matching heuristic[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa254.

[76] DE ALMEIDA B P, REITER F, PAGANI M, et al. DeepSTARR predicts enhancer activity from DNA sequence and enables the *de novo* design of synthetic enhancers[J]. Nature Genetics, 2022, 54(5): 613-624.

[77] FENG C Q, ZHANG Z Y, ZHU X J, et al. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators[J]. Bioinformatics, 2019, 35(9): 1469-1477.

[78] YEVSHIN I, SHARIPOV R, VALEEV T, et al. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments[J]. Nucleic Acids Research, 2017, 45(D1): D61-D67.

[79] KULAKOVSKIY I V, VORONTSOV I E, YEVSHIN I S, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse *via* large-scale ChIP-Seq analysis[J]. Nucleic Acids Research, 2018, 46(D1): D252-D259.

[80] KULAKOVSKIY I V, VORONTSOV I E, YEVSHIN I S, et al. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models[J]. Nucleic Acids Research, 2016, 44(D1): D116-D125.

[81] WANG G H, LUO X M, WANG J N, et al. MeDReaders: a database for transcription factors that bind to methylated DNA[J]. Nucleic Acids Research, 2018, 46(D1): D146-D151.

[82] HAN H, CHO J W, LEE S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory

interactions[J]. Nucleic Acids Research, 2018, 46(D1): D380-D386.

[83] HAN H, SHIM H, SHIN D, et al. TRRUST: a reference database of human transcriptional regulatory interactions[J]. Scientific Reports, 2015, 5: 11432.

[84] HU H, MIAO Y R, JIA L H, et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors[J]. Nucleic Acids Research, 2019, 47(D1): D33-D38.

[85] MULUGETA T D, NOME T, TO T H, et al. SalMotifDB: a tool for analyzing putative transcription factor binding sites in salmonid genomes[J]. BMC Genomics, 2019, 20(1): 694.

[86] ZHANG Q, LIU W, ZHANG H M, et al. hTFtarget: a comprehensive database for regulations of human transcription factors and their targets[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 120-128.

[87] FENG C C, SONG C, LIU Y J, et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors[J]. Nucleic Acids Research, 2020, 48(D1): D93-D100.

[88] CASTRO-MONDRAGON J A, RIUDAVETS-PUIG R, RAULUSEVICIUTE I, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles[J]. Nucleic Acids Research, 2022, 50(D1): D165-D173.

[89] NI P Y, SU Z C. PCRMS: a database of predicted *cis*-regulatory modules and constituent transcription factor binding sites in genomes[J]. Database, 2022, 2022: baac024.

[90] FU L Y, ZHANG L H, DOLLINGER E, et al. Predicting transcription factor binding in single cells through deep learning[J]. Science Advances, 2020, 6(51): eaba9031.

[91] KIM G B, GAO Y, PALSSON B O, et al. DeepTFactor: a deep learning-based tool for the prediction of transcription factors[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(2): e2021171118.

[92] ZHANG Q H, WANG S G, CHEN Z H, et al. Locating transcription factor binding sites by fully convolutional neural network[J]. Briefings in Bioinformatics, 2021, 22(5): bbaa435.

[93] ZHENG A, LAMKIN M, ZHAO H Q, et al. Deep neural networks identify sequence context features predictive of transcription factor binding[J]. Nature Machine Intelligence, 2021, 3(2): 172-180.

[94] NAKASHIMA N, TAMURA T, GOOD L. Paired termini stabilize antisense RNAs and enhance conditional gene silencing in *Escherichia coli*[J]. Nucleic Acids Research, 2006, 34(20): e138.

[95] NA D, YOO S M, CHUNG H, et al. Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs[J]. Nature Biotechnology, 2013, 31(2): 170-174.

[96] ZHANG R H, ZHANG Y, WANG J, et al. Development of antisense RNA-mediated quantifiable inhibition for metabolic regulation[J]. Metabolic Engineering Communications, 2021, 12: e00168.

[97] GREEN A A, SILVER P A, COLLINS J J, et al. Toehold switches: *de-novo*-designed regulators of gene expression[J]. Cell, 2014, 159(4): 925-939.

[98] CHAPPELL J, WESTBROOK A, VEROSLOFF M, et al. Computational design of small transcription activating RNAs for versatile and dynamic gene regulation[J]. Nature Communications, 2017, 8: 1051.

[99] CHAPPELL J, TAKAHASHI M K, LUCKS J B. Creating small transcription activating RNAs[J]. Nature Chemical Biology, 2015, 11(3): 214-220.

[100] JANG S H, JANG S Y, YANG J N, et al. RNA-based dynamic genetic controllers: development strategies and applications[J]. Current Opinion in Biotechnology, 2018, 53: 1-11.

[101] ZADEH J N, STEENBERG C D, BOIS J S, et al. NUPACK: analysis and design of nucleic acid systems[J]. Journal of Computational Chemistry, 2011, 32(1): 170-173.

[102] GARDNER T S, CANTOR C R, COLLINS J J. Construction of a genetic toggle switch in *Escherichia coli*[J]. Nature, 2000, 403(6767): 339-342.

[103] CHEN T, CHENG G Y, AHMED S, et al. New methodologies in screening of antibiotic residues in animal-derived foods: Biosensors[J]. Talanta, 2017, 175: 435-442.

[104] SONGA E A, OKONKWO J O. Recent approaches to improving selectivity and sensitivity of enzyme-based biosensors for organophosphorus pesticides: a review[J]. Talanta, 2016, 155: 289-304.

[105] PUIU M, BALA C. Peptide-based biosensors: from self-assembled interfaces to molecular probes in electrochemical assays[J]. Bioelectrochemistry, 2018, 120: 66-75.

[106] SHARMA S, BYRNE H, O'KENNEDY R J. Antibodies and antibody-derived analytical biosensors[J]. Essays in Biochemistry, 2016, 60(1): 9-18.

[107] WANG R E, ZHANG Y, CAI J, et al. Aptamer-based fluorescent biosensors[J]. Current Medicinal Chemistry, 2011, 18(27): 4175-4184.

[108] BLIND M, BLANK M. Aptamer selection technology and recent advances[J]. Molecular Therapy Nucleic Acids, 2015, 4 (1): e223.

[109] HONG P T K, JANG C H. Sensitive and label-free liquid crystal-based optical sensor for the detection of malathion[J]. Analytical Biochemistry, 2020, 593: 113589.

[110] KIM H S, AN Z F, JANG C H. Label-free optical detection of thrombin using a liquid crystal-based aptasensor[J]. Microchemical Journal, 2018, 141: 71-79.

[111] O'NEILL M, KELLY S M. Liquid crystals for charge transport, luminescence, and photonics[J]. Advanced Materials, 2003, 15

(14): 1135-1146.

[112] BYKHOVSKI A, ZHANG W D, JENSEN J, et al. Analysis of electronic structure, binding, and vibrations in biotin-streptavidin complexes based on density functional theory and molecular mechanics[J]. The Journal of Physical Chemistry B, 2013, 117(1): 25-37.

[113] PAULLA K K, FARAJIAN A A. Concentration effects of carbon oxides on sensing by graphene nanoribbons: *ab initio* modeling[J]. The Journal of Physical Chemistry C, 2013, 117(24): 12815-12825.

[114] KUMAR N, SEMINARIO J M. Design of nanosensors for fissile materials in nuclear waste water[J]. The Journal of Physical Chemistry C, 2013, 117(45): 24033-24041.

[115] NEZHADALI A, MOJARRA M. Computational study and multivariate optimization of hydrochlorothiazide analysis using molecularly imprinted polymer electrochemical sensor based on carbon nanotube/polypyrrole film[J]. Sensors and Actuators B: Chemical, 2014, 190: 829-837.

[116] LIU Q Y, ZUO F, ZHAO Z G, et al. Molecular dynamics investigations of an indicator displacement assay mechanism in a liquid crystal sensor[J]. Physical Chemistry Chemical Physics: PCCP, 2017, 19(35): 23924-23933.

[117] CAUSIN P, SACCO R, VERRI M. A multiscale approach in the computational modeling of the biophysical environment in artificial cartilage tissue regeneration[J]. Biomechanics and Modeling in Mechanobiology, 2013, 12(4): 763-780.

[118] ZHANG W J, DU Y Q, CRANFORD S W, et al. Biosensor design through molecular dynamics simulation[J]. World Academy of Science, Engineering and Technology, International Journal of Biomedical and Biological Engineering 2016, 10(1): 10-14.

[119] KHOSHBIN Z, HOUSAINDOKHT M R, IZADYAR M, et al. Theoretical design and experimental study of new aptamers with the improved target-affinity: new insights into the Pb^{2+}-specific aptamers as a case study[J]. Journal of Molecular Liquids, 2019, 289: 111159.

[120] ZHUANG S L, WANG H F, DING K K, et al.. Interactions of benzotriazole UV stabilizers with human serum albumin: atomic insights revealed by biosensors, spectroscopies and molecular dynamics simulations[J]. Chemosphere, 2016, 144: 1050-1059.

[121] KOLLMAN P A, MASSOVA I, REYES C, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models[J]. Accounts of Chemical Research, 2000, 33(12): 889-897.

[122] KHOSHBIN Z, HOUSAINDOKHT M R. Computer-aided aptamer design for sulfadimethoxine antibiotic: step by step mutation based on MD simulation approach[J]. Journal of Biomolecular Structure & Dynamics, 2021, 39(9): 3071-3079.

[123] DO P C, LEE E H, LE L. Steered molecular dynamics simulation in rational drug design[J]. Journal of Chemical Information and Modeling, 2018, 58(8): 1473-1482.

[124] THYPARAMBIL A A, ABRAMYAN T M, BAZIN I, et al. Site of tagging influences the ochratoxin recognition by peptide NFO4: a molecular dynamics study[J]. Journal of Chemical Information and Modeling, 2017, 57(8): 2035-2044.

[125] THYPARAMBIL A A, BAZIN I, GUISEPPI-ELIE A. Evaluation of ochratoxin recognition by peptides using explicit solvent molecular dynamics[J]. Toxins, 2017, 9(5): 164.

[126] SALMASO V, STURLESE M, CUZZOLIN A, et al. Combining self- and cross-docking as benchmark tools: the performance of DockBench in the D3R grand challenge 2[J]. Journal of Computer-Aided Molecular Design, 2018, 32(1): 251-264.

[127] LIU Q J, WANG H, LI H L, et al. Impedance sensing and molecular modeling of an olfactory biosensor based on chemosensory proteins of honeybee[J]. Biosensors & Bioelectronics, 2013, 40(1): 174-179.

[128] ABOLHASAN R, MEHDIZADEH A, RASHIDI M R, et al. Application of hairpin DNA-based biosensors with various signal amplification strategies in clinical diagnosis[J]. Biosensors & Bioelectronics, 2019, 129: 164-174.

[129] CROSS S, BARONI M, CAROSATI E, et al. FLAP: GRID molecular interaction fields in virtual screening. validation using the DUD data set[J]. Journal of Chemical Information and Modeling, 2010, 50(8): 1442-1450.

[130] MA D L, CHAN D S H, LEE P, et al. Molecular modeling of drug-DNA interactions: virtual screening to structure-based design[J]. Biochimie, 2011, 93(8): 1252-1266.

[131] PIZZONI D, MASCINI M, COMPAGNONE D, et al. Virtual screening peptide selection for a peptide based gas sensors array[C/OL]//Proceedings of the Second National Conference on Sensors, Rome, Italy, 19-21 February, 2014, 2015: 89-93[2023-01-01]. https://link. springer. com/chapter/10.1007/978-3-319-09617-9_16.

[132] MASCINI M, PIZZONI D, PEREZ G, et al. Tailoring gas sensor arrays *via* the design of short peptides sequences as binding elements[J]. Biosensors & Bioelectronics, 2017, 93: 161-169.

[133] FRANCA E F, LEITE F L, CUNHA R A, et al. Designing an enzyme-based nanobiosensor using molecular modeling techniques[J]. Physical Chemistry Chemical Physics: PCCP, 2011, 13(19): 8894-8899.

[134] HONG ENRIQUEZ R P, PAVAN S, BENEDETTI F, et al. Designing short peptides with high affinity for organic molecules: A combined docking, molecular dynamics, and Monte Carlo approach[J]. Journal of Chemical Theory and Computation, 2012, 8(3): 1121-1128.

[135] SHCHERBININ D S, GNEDENKO O V, KHMELEVA S A, et al. Computer-aided design of aptamers for cytochrome P450[J].

Journal of Structural Biology, 2015, 191(2): 112-119.

[136] PRANDI I G, RAMALHO T C, FRANÇA T C C. Esterase 2 as a fluorescent biosensor for the detection of organophosphorus compounds: docking and electronic insights from molecular dynamics[J]. Molecular Simulation, 2019, 45(17): 1432-1436.

[137] SHAHBAAZ M, KANCHI S, SABELA M, et al. Structural basis of pesticide detection by enzymatic biosensing: a molecular docking and MD simulation study[J]. Journal of Biomolecular Structure & Dynamics, 2018, 36(6): 1402-1416.

[138] CHAKRAVORTY D K, PARKER T M, GUERRA A J, et al. Energetics of zinc-mediated interactions in the allosteric pathways of metal sensor proteins[J]. Journal of the American Chemical Society, 2013, 135(1): 30-33.

[139] GROENHOF G. Introduction to QM/MM simulations[J]. Methods in Molecular Biology, 2013, 924: 43-66.

[140] PAPAMICHAEL E M, STAMATIS H, STERGIOU P Y, et al. Enzyme kinetics and modeling of enzymatic systems[M/OL]. Advances in Enzyme Technology, Amsterdam: Elsevier, 2019: 71-104[2023-01-01]. https://www.sciencedirect.com/science/article/pii/B9780444641144000030?via%3Dihub.

[141] RYDE U. QM/MM calculations on proteins[J]. Methods in Enzymology, 2016, 577: 119-158.

[142] WONG M W, XIE H F, KWA S T. Anion recognition by azophenol thiourea-based chromogenic sensors: a combined DFT and molecular dynamics investigation[J]. Journal of Molecular Modeling, 2013, 19(1): 205-213.

[143] CHARCHAR P, CHRISTOFFERSON A J, TODOROVA N, et al. Understanding and designing the gold-bio interface: insights from simulations[J]. Small, 2016, 12(18): 2395-2418.

[144] ZHU C, LI L S, YANG G, et al. High-efficiency selection of aptamers for bovine lactoferrin by capillary electrophoresis and its aptasensor application in milk powder[J]. Talanta, 2019, 205: 120088.

[145] YARIZADEH K, BEHBAHANI M, MOHABATKAR H, et al. Computational analysis and optimization of carcinoembryonic antigen aptamers and experimental evaluation[J]. Journal of Biotechnology, 2019, 306: 1-8.

[146] KHAVANI M, IZADYAR M, HOUSAINDOKHT M R. Theoretical design and experimental study on the gold nanoparticles based colorimetric aptasensors for detection of neomycin B[J]. Sensors and Actuators B: Chemical, 2019, 300: 126947.

[147] MITCHLER M M, GARCIA J M, MONTERO N E, et al. Transcription factor-based biosensors: a molecular-guided approach for natural product engineering[J]. Current Opinion in Biotechnology, 2021, 69: 172-181.

[148] HOSSAIN G S, SAINI M, MIYAKE R, et al. Genetic biosensor design for natural product biosynthesis in microorganisms[J]. Trends in Biotechnology, 2020, 38(7): 797-810.

[149] LIANG W F, CUI L Y, CUI J Y, et al. Biosensor-assisted transcriptional regulator engineering for *Methylobacterium extorquens* AM1 to improve mevalonate synthesis by increasing the acetyl-CoA supply[J]. Metabolic Engineering, 2017, 39: 159-168.

[150] KASEY C M, ZERRAD M, LI Y W, et al. Development of transcription factor-based designer macrolide biosensors for metabolic engineering and synthetic biology[J]. ACS Synthetic Biology, 2018, 7(1): 227-239.

[151] DE PAEPE B, MAERTENS J, VANHOLME B, et al. Chimeric LysR-type transcriptional biosensors for customizing ligand specificity profiles toward flavonoids[J]. ACS Synthetic Biology, 2019, 8(2): 318-331.

[152] LIANG M D, LI Z L, WANG W S, et al. A CRISPR-Cas12a-derived biosensing platform for the highly sensitive detection of diverse small molecules[J]. Nature Communications, 2019, 10: 3672.

[153] MCCANN J J, PIKE D H, BROWN M C, et al. Computational design of a sensitive, selective phase-changing sensor protein for the VX nerve agent[J]. Science Advances, 2022, 8(27): eabh3421.

通讯作者：王晟（1983—），男，博士，上海智峪生物科技有限公司 CEO，中国科学院深圳先进技术研究院客座研究员。研究方向为基于深度学习的蛋白质结构预测、基于人工智能的合成生物学。
E-mail：wangsheng@zelixir.com

第一作者：王泽琛（1997—），男，博士研究生。研究方向为基于深度学习的蛋白质-配体相互作用预测和虚拟筛选。
E-mail：wangzechen@mail.sdu.edu.cn

第一作者：陈威华（1982—），男，博士研究生。研究方向为合成生物学方向，基因合成与组装。
E-mail：chenweihua@zelixir.com