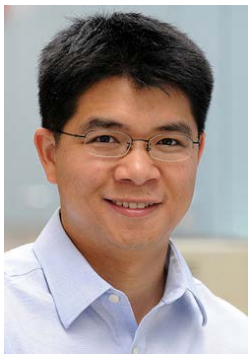
·生物元器件智能设计合成·



王宝俊 浙江大学求是讲席教授，浙江大学杭州国际科创中心生物与分子智造研究院副院长、合成生物学研究所所长，曾任爱丁堡大学生物科学学院终身教授。长期从事合成生物使能技术、基因线路设计研究及其在生物传感、智能诊疗、生物制造等领域的创新应用。主持英国自然科学基金会、美国海军研究署和国家重点研发计划重点专项、国家自然科学基金重点国际合作项目等20余项研究基金。*PLoS Biology*、*ACS Synthetic Biology* 等期刊编委。获比尔盖茨基金会全球大挑战探索基金奖、英国杰出青年科学基金奖、英国生物技术与科学基金会新研究员奖等奖励荣誉。入选英国皇家化学学会会士、教育部“长江学者奖励计划”讲席学者。

合成生物元件与线路的智能设计

毛瑞超[1]，王宝俊[1,2*]

1 浙江大学 化学工程与生物工程学院，浙江 杭州 310058
2 浙江大学 杭州国际科创中心，浙江 杭州 311215

毛瑞超, 王宝俊. 合成生物元件与线路的智能设计[J]. 生物工程学报, 2025, 41(3): 1023-1051.
MAO Ruichao, WANG Baojun. Machine learning-aided design of synthetic biological parts and circuits[J]. Chinese Journal of Biotechnology, 2025, 41(3): 1023-1051.

摘 要：合成生物学是生物学、工程学和计算机科学等多学科交叉融合的新兴前沿领域，旨在通过“自下而上”的工程化设计理念，逐级构建元件、器件和线路，以创造自然界中不存在的人工生物系统，或对已有的生物系统进行目标性改造。随着合成生物产业的飞速发展，对基因线路规模和复杂度的需求也在不断提升。然而，传统依赖经验和试错的方法在元件与线路构建中具有较低的效率和成功率，已无法满足合成生物科技创新转化的需求。这促使元件与线路的开发范式逐渐从人力型、经验型的试错模式向标准化、智能化的工程模式转变。机器学习能够揭示生物数据中隐含的结构和关联，为合成生物元件和线路的智能设计提供强大支持。本文综述了生物元件与线路设计中常用的机器学习算法，以及它们在合成启动子、RNA 调控元件、转录因子等生物元件和简单基因线路智能设计中的典型应用，探讨了当前面临的主要挑战及潜在的解决方案。最后，本文展望了机器学习与合成生物系统设计未来的融合趋势，并强调了跨学科合作的重要性。

关键词：合成生物学；生物元件；基因线路；生物设计；机器学习；深度学习

资助项目：国家重点研发计划(2023YFF1204500)；浙江省“尖兵”“领雁”研发攻关计划(2024C03011)；国家自然科学基金(32271475, 32320103001)
This work was supported by the National Key Research and Development Program of China (2023YFF1204500), the “Pioneer” and “Leading Goose” Research and Development Program of Zhejiang Province (2024C03011), and the National Natural Science Foundation of China (32271475, 32320103001).
*Corresponding author. E-mail: baojun.wang@zju.edu.cn
Received: 2024-07-25; Accepted: 2024-11-06; Published online: 2024-11-06

Machine learning-aided design of synthetic biological parts and circuits

MAO Ruichao[1], WANG Baojun[1,2*]

1 College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310058, Zhejiang, China
2 ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, Zhejiang, China

Abstract: Synthetic biology is an emerging interdisciplinary field at the convergence of biology, engineering, and computer science. It employs a bottom-up approach to progressively design biological parts, devices, and circuits, aiming to create artificial biological systems not found in nature or to redesign existing biological systems for specific purposes. With the rapid development of the synthetic biology industry, there is an increasing demand for large complex genetic circuits. However, the traditional trial-and-error methods, heavily reliant on empirical knowledge, have limited efficiency and success rates of parts/circuits construction, thereby impeding the innovation and technology translation for synthetic biology. These limitations have prompted a paradigm shift from labor-intensive, experience-driven trial-and-error models towards standardized, intelligent engineering approaches. Machine learning, capable of uncovering hidden structures and relationships within biological data, offers robust support for the intelligent design of synthetic biological parts and genetic circuits. Here, we review commonly used machine learning algorithms and analyze their typical applications in designing biological parts (e.g., synthetic promoters, RNA regulatory elements, and transcription factors) and simple genetic circuits. Additionally, we discuss the primary challenges in machine learning-aided design and propose potential solutions. Lastly, we envision the future trend of integrating machine learning with synthetic biological system design, highlighting the importance of interdisciplinary collaboration.

Keywords: synthetic biology; biological parts; genetic circuits; biodesign; machine learning; deep learning

合成生物学是一门新兴的交叉学科，旨在通过工程学原理设计和改造生物系统，以实现预设的生物学功能。合成生物元件，如启动子、增强子和转录因子等，是构建合成生物系统的基石。通过合理的设计与组合，这些元件可以整合成具有高级功能的基因线路，从而驱动宿主细胞展现出预期的生物学行为[1]。当前，研究人员已成功构建出多种功能的基因线路，如环境感知[2-3]、动态基因表达调控[4-5]和群体水平控制[6]等，这些线路在环境监测、生物制造以及代谢工程等领域得到了广泛应用，并在疾病治疗方面展现出巨大潜力[7]。

然而，传统的生物元件和基因线路设计方法主要依赖于专家经验和实验试错，设计过程复杂且耗时，极大地限制了效率与成功率[8]。尽管物理化学模型，如分子模拟技术[9]，在一定程度上提升了设计效率，但仍受到多种因素的制约。例如，为了使研究问题变得可解，这些模型往往需要做出一定的简化假设，这可能使其无法全面反映生物系统内在的复杂性和不

确定性。此外，这些方法在面对复杂系统建模时的可扩展性有限，难以满足日益增长的设计需求和系统规模的扩增。

近年来，数据驱动模型，尤其是机器学习和深度学习技术，为应对这些挑战提供了创新性的解决方案[10]。这些技术凭借其强大的数据处理及模式识别能力，能有效挖掘庞大生物数据集中的深层结构与关联，进而预测生物元件或线路的功能[11-13]。此外，深度学习还具有逆向生成的能力，能够根据预设目标功能反向设计生物元件或线路。

在实际应用中，机器学习已成功应用于预测基因表达水平[14-15]、mRNA 翻译水平[16-17]及酶活性[18-19]等领域，并取得了显著成效。与此同时，深度学习在逆向设计方面也展现出巨大潜力，成功实现了启动子[20-22]、RNA toehold 开关[23]及多肽[24-25]等生物元件的从头设计。这些创新性应用不仅大幅减少了实验迭代次数，还显著提升了研发效率与成功率，正以前所未有的速度和深度重塑着生物元件与基因线路的设计范式。

本文首先概述了合成生物元件与基因线路设计中常用的机器学习算法，接着展示了机器学习如何在合成启动子、RNA 调控元件、转录因子等生物元件及简单基因线路的智能设计中突破传统方法的局限。最后，探讨了智能化设计当前面临的主要挑战及其潜在解决方案，并展望了机器学习与合成生物系统设计未来的融合趋势，强调了跨学科合作在这一进程中的重要性。

1 智能设计原理与方法

1.1 机器学习概述

机器学习作为人工智能的核心领域，赋予计算机自主学习数据中规律和模式的能力，无需依赖预设规则。其算法种类繁多，每年均有数百种新算法被开发出来[12]。这些算法主要分为以下几类：监督学习、无监督学习、强化学习、半监督学习、主动学习和迁移学习(表 1)。

1.1.1 监督学习

监督学习是最广泛使用的机器学习方法，它利用标注数据训练模型。在训练过程中，模型通过调整参数以减少预测误差，从而提高对新数据的预测准确性。监督学习分为分类和回归这 2 种类型：在分类问题中，输出标签是离散的，算法的任务是将输入数据分配到不同的类别，如区分基因序列的编码区和非编码区[40]；回归是预测连续数值，如估计基因在特定条件下的表达量[41]。

1.1.2 无监督学习

无监督学习不依赖标签，而是通过分析数据本身的结构来揭示隐藏的模式。它主要包括聚类和降维两大类。聚类旨在将相似数据点归为一组，而降维则旨在简化数据结构的同时保留关键信息。例如，聚类分析能帮助识别表达模式相近的基因[42]。无监督学习能从大量未标记数据中提取有用信息，但由于缺乏明确的输出标签，其性能评估具有一定挑战性。

1.1.3 强化学习

强化学习通过试错和环境反馈来优化决策过程。它不依赖于标记数据，而是通过奖励和惩罚信号来学习如何在不同情况下采取行动以最大化收益。这种学习方式模拟了生物体在自然界中通过经验积累进行学习的过程。在合成生物学领域，强化学习已被用于优化微生物共培养的代谢产物生产[43]，通过奖励成功的生产策略并惩罚失败的生产策略，它可以逐渐学会在特定条件下实现最优生产。

1.1.4 半监督学习

半监督学习结合了监督学习和无监督学习

表 1 机器学习算法的类型

Table 1 Classes of machine learning algorithms

Classes	Tasks	Common algorithms
Supervised learning	Classification	Logistic regression
		K-nearest neighbors (K-NN)
		Support vector machine (SVM)
		Neural network
		Random forests
		Naïve Bayes
	Regression	Linear regression
		Ridge regression
		Neural network
Unsupervised learning	Clustering	K-means
		Hierarchical clustering
		DBSCAN
	Dimensionality reduction	Principal component analysis (PCA)
		Singular value decomposition (SVD)
	Generative model	Variational autoencoders (VAE)[26]
		Generative adversarial networks (GAN)[27]
Reinforcement learning	Maximize long-term cumulative reward	Value function: value iteration, Q-learning[28]
		Policy gradient methods[29]: REINFORCE; actor-critic
		Deep reinforcement learning[30]
Semi-supervised learning	Pseudolabeling	Self-training[31]
		Cotraining[32]
	Consistency regularization	Virtual adversarial training (VAT)[33]
		Mean teacher[34]
Active learning	Request labels on the most informative examples	Uncertainty sampling[35]
		Representative sampling[36]
		Density weighting[37]
Transfer learning	Improve the learning of new tasks with the knowledge of source tasks	Pre-trained feature extraction[38]
		Model reuse[39]

的特点，利用少量标记数据和大量未标记数据进行训练。这种方法有助于减少对大量人工标记数据的依赖，在标记数据获取成本高昂或耗时的情况下尤为有用。尽管该方法在合成生物学领域尚未广泛应用，但近期的一些研究[44-45]已证实其巨大潜力。

1.1.5 主动学习

主动学习是半监督学习的变体，它允许算法主动从大量未标记数据中选择最具信息价值的样本进行人工标注，然后将这些新标记数据纳入训练集，以优化模型性能。主动学习的核心在于智能选择那些能够最大程度减少模型不

确定性的样本，从而在有限的标注成本下实现模型性能的快速提升。它在数据标注成本高昂或标注工作量庞大的场景中尤为适用。例如，Borkowski 等[46]使用主动学习算法探索了数百万种无细胞缓冲液组合，最大限度地提高了蛋白质产量。

1.1.6 迁移学习

迁移学习允许模型将在一个任务上学到的知识应用到另一个相关任务上。模型先在数据丰富的源任务上进行预训练，然后将学到的特征表示或参数迁移到数据稀缺的目标任务上进行微调或重用。这种方法在目标任务数据不足时尤为有效，因为它能利用源任务上的知识，加快目标任务的学习过程并提高模型性能。例如，从酵母生长率预测中学习到的特征可以用于预测酵母产生的乙醇量[47]。

1.2 常用机器学习算法

本节概述了合成生物学中常用的机器学习算法,并在表 1 中列出了它们的类别和任务类型。

1.2.1 线性回归/逻辑回归

线性回归用于分析自变量(输入特征)与因变量(输出结果)之间的关系,假设它们之间存在线性关系，并通过最小化预测值与真实值之间的残差平方和来估计模型参数(图 1A)。逻辑回归则用于分类问题，它基于线性回归，通过在输出层引入 Sigmoid 函数将输出转换为概率，以表示样本属于某一类别的可能性(图 1B)。逻辑回归因其简洁性和强解释性，在合成生物学领域得到了广泛应用，如优化 gRNA 序列[48]和识别内含活性肽断裂位点[49]等。

1.2.2 支持向量机

支持向量机(support vector machine, SVM)常用于解决分类问题，其核心思想是找到一个最优超平面来分隔不同类别，并最大化超平面与最近数据点(支持向量)之间的距离(图 1C)。

1.2.3 随机森林

随机森林(random forests, RF)是一种集成学习算法，它通过组合多个决策树来增强分类和回归任务的性能(图 1D)。每棵树在数据的随机子集上独立生长，预测时通过多数投票(分类任务)或取平均值(回归任务)的方式整合所有树的输出。这种方法可有效降低单棵树的偏差和方差，防止过拟合。因此，随机森林在处理高维复杂数据集时表现出色，如在提高目标产物产量[50]和预测启动子强度[15]等方面的应用。

1.2.4 K-最近邻

K-最近邻(K-nearest neighbors, K-NN)算法基于相似数据点聚集于邻近区域的假设(图 1E)。在预测时,它计算待分类点与训练集中各点的距离，选取最近的 K 个点，根据这些点的类别信息(通过多数投票方式)预测待分类点的类别。K-NN 算法实现简单,可以通过调整 K 值来优化预测效果，已成功应用于评估酶的功能特性[51]。然而，随着数据集维度增加，K-NN 的计算复杂度显著提升，可能导致预测速度变慢[52]。

1.2.5 人工神经网络

人工神经网络(artificial neural network, ANN)受生物神经系统启发,通过多层神经元之间的连接和权重调整来模拟复杂的非线性关系(图 1F)。在 ANN 中，输入层接收特征表示，输出层输出预测结果，而隐藏层则处理输入与输出间的复杂关系。每个神经元通过权重和激活函数传递信息。ANN 已应用于提高无细胞系统的丁醇生产[53]和优化多基因代谢途径[54]。

1.2.6 深度学习模型

深度学习是机器学习的一个分支，它使用多层神经网络来自动学习和提取数据中的高级特征。通过叠加隐藏层，深度学习模型能够构建复杂的层次结构，从而捕捉数据的深层规律和模式。常见的深度学习模型包括前馈神经网

络(feedforward neural network, FNN)、卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)[如长短期记忆网络(long short-term memory, LSTM)]、递归神经网络、图神经网络(graph neural network, GNN)以及生成模型(如生成对抗网络和变分自编码器)。这些模型在图像识别、自然语言处理和语音识别等领域取得了显著成果[55]，特别是在拥有大量标注数据的情况下，它们的性能优于传统机器学习算法[56-57]。

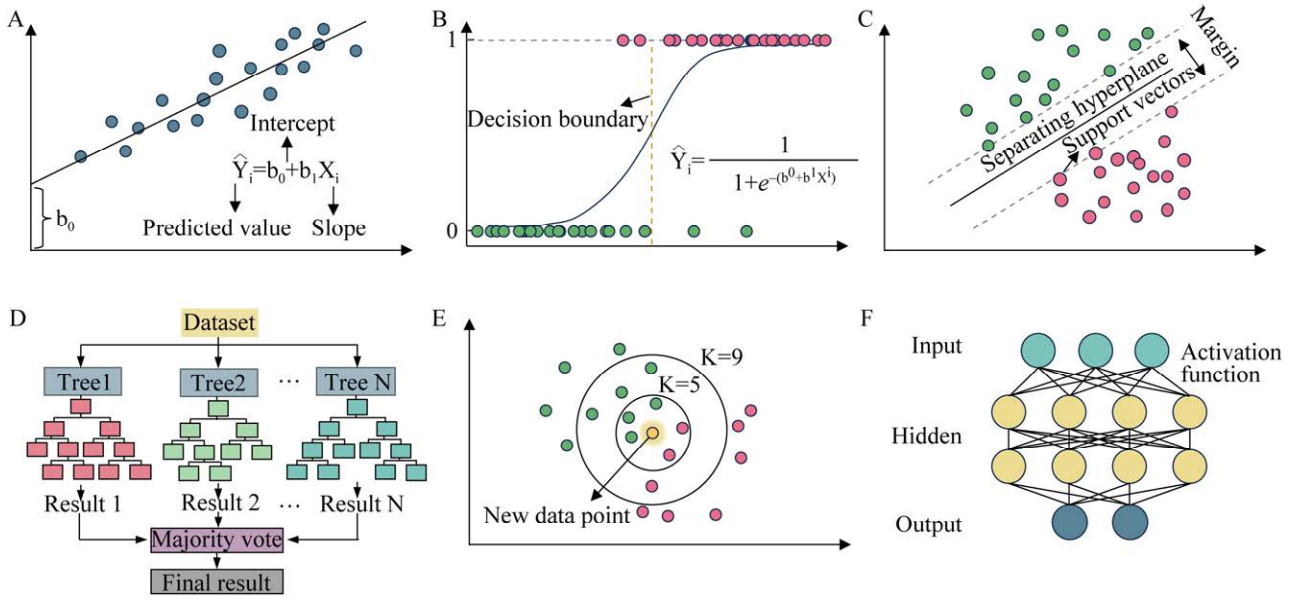


图 1　常用机器学习算法的数学框架　A：线性回归示意图。蓝色点代表数据点，黑色线为最佳拟合直线，旨在最小化所有数据点与回归线之间垂直距离的残差平方和。B：逻辑回归示意图。绿色和红色点代表两个不同类别的数据点。S 形曲线代表逻辑函数，其输出是一个介于 0 和 1 之间的概率值，表示给定数据点属于某一类的可能性。C：支持向量机示意图。其试图找到一个分隔超平面(线)，以最佳方式将数据分为不同类别。D：随机森林示意图。将数据集分为 N 棵决策树，每棵树独立生长和评估，最后使用多数投票的方式整合所有决策树的评估结果。E：K-最近邻示意图。根据数据点周围 K 个最近数据点的多数类别标签来对该数据点进行分类。F：人工神经网络示意图。通过神经元之间的连接(具有不同权重和阈值)来模拟复杂的非线性关系。

Figure 1　Mathematical framework of commonly used machine learning algorithms. A: Linear regression schematic, where blue dots represent data points and the black line denotes the best-fit line, which aims to minimize the sum of squared residuals of the vertical distances between all data points and the regression line. B: Logistic regression schematic, where green and red data dots represent data points from two different classes. The S-shaped curve represents the logistic function, which outputs a probability value between 0 and 1, indicating the likelihood of a given data point belongs to a particular class. C: Support vector machine schematic, which aims to find a separating hyperplane (line) that best divides the data into different classes. D: Random forest schematic, where the dataset is divided into N decision trees, each growing and evaluating independently, and the results of all decision trees are consolidated using majority voting. E: K-nearest neighbors schematic, which classifies a data point based on the majority class labels of the k nearest neighbors. F: Artificial neural network schematic, which simulates complex nonlinear relationships through connections between neurons with varying weights and thresholds.

2 合成生物元件的智能设计

2.1 合成生物元件概述

生物元件是合成生物系统的基本构建单元。根据分子类型，它们可分为 DNA 元件(如启动子、增强子和沉默子等)、RNA 元件(如核糖体结合位点、gRNA 和核糖开关等)、蛋白质元件(如转录因子、内含肽和信号蛋白等)及其他生物分子(如辅酶、信号分子和调控分子等)。通过模块化设计与组合，这些元件可以协同工作，实现包括信号感知、逻辑处理、代谢调控在内的多种生物功能。在此背景下，本课题组率先提出了基因线路规模化设计需要遵循的工程化设计原则(模块性和正交性)，并首次在大肠杆菌中设计出了模块化且正交化的逻辑与门、与非门控线路，并实现了具有连续可调放大倍数的模块化转录信号放大线路[58-59]。

功能明确、标准化的生物元件是合成生物系统工程化开发的基石。然而，鉴于生物系统内在的复杂性和不确定性，传统依赖实验试错的方法在设计和优化这些元件时面临巨大挑战。近年来，凭借强大的数据处理和模式识别能力，机器学习已广泛应用于合成生物元件的功能预测和从头设计(图 2)。表 2 中总结了这些应用实例，并在下文中进行了详细讨论。

2.2 合成生物元件功能预测

在合成生物元件设计中，理解元件序列如何决定其功能，即序列-功能之间的复杂映射关系，是核心问题。传统的生物物理学手段，如结构解析和分子模拟，虽然能够提供深刻的见解，但常常受限于繁琐的实验流程或漫长的计算周期。机器学习技术的引入为这一领域带来了突破，它能够直接从庞大的生物元件序列数据库(如大规模突变体文库)中发现隐含规律，通过优化模型参数，快速准确地预测未知序列的功能。这一过程摆脱了对复杂生物物理学理论的依赖，实现了序列-功能关系的定性乃至定量预测。

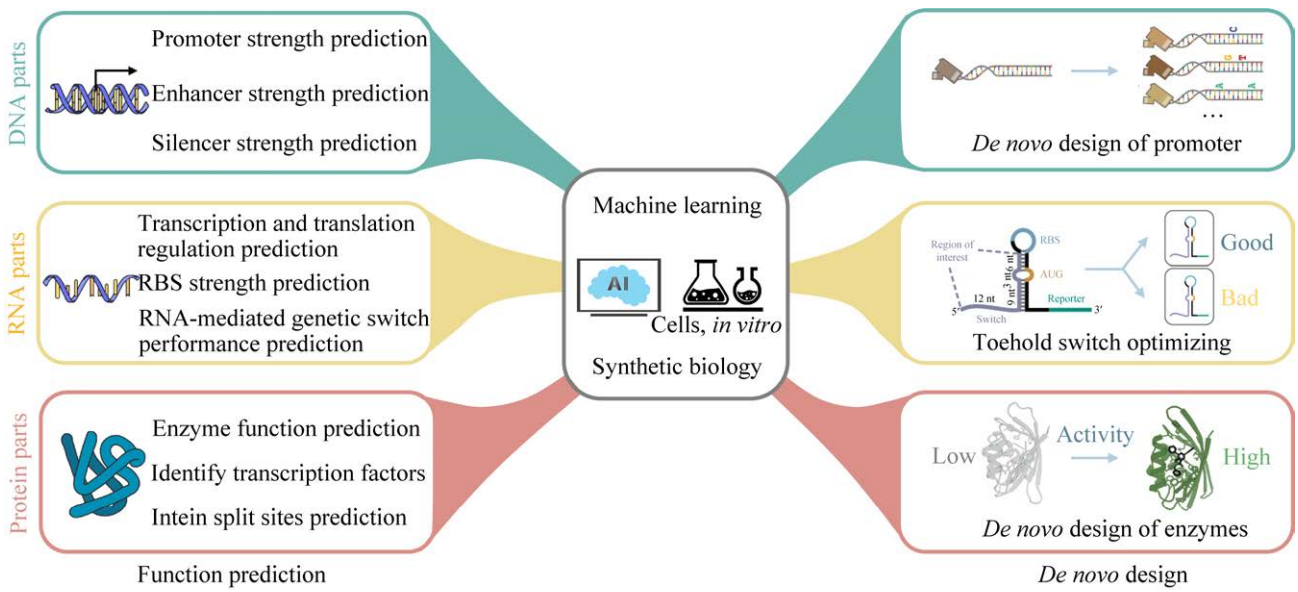


图 2 机器学习算法在合成生物元件设计中的应用

Figure 2 Applications of machine learning algorithms in the design of synthetic biological parts.

表 2 机器学习算法在合成生物元件功能预测和从头设计中的应用

Table 2 Applications of machine learning in the functional prediction and *de novo* design of synthetic biological parts

Types	Parts	Applications	Algorithms	References
DNA parts	Promoter	Promoter strength prediction	XgBoost	[15]
		Promoter strength prediction	ANN	[60]
		Transcription initiation frequency prediction	CNN	[61]
		Transcriptional activity prediction	Logistic regression/CNN	[62]
		Promoter activity prediction	CNN	[41]
		Promoter activity prediction	CNN	[63]
		De novo design of promoter sequences	VAE/CNN	[20]
		De novo design of promoter sequences	GAN/VAE/Diffusion/LSTM/CNN	[64]
	Enhancer	Identify enhancer sequences	SVM	[65]
		Predict enhancer strength	SVM	[66]
		Identify enhancer sequences/predict strength	RF	[67]
	Silencer	Identify silencer sequences	SVM	[68]
		Identify silencer sequences	CNN/ANN	[69]
		Identify silencer sequences	CNN	[70]
		Identify silencer sequences	SVM	[71]
		Identify silencer sequences/predict strength	Autoencoder/CNN	[72]
RNA parts	mRNA	Mean ribosome load prediction	CNN	[16]
		Half-life prediction	Hybrid model (CNN/RNN)	[73]
		Mean ribosome load/half-life prediction	Hybrid model (CNN/RNN)	[17]
	RNA toehold switch	Toehold switch function prediction	CNN/LSTM	[74]
		Optimizing performance	CNN/RNN/language model	[23]
	gRNA	gRNA activity prediction	SVM/Gradient Boosting Decision Tree	[48]
		Cas9 cleavage efficiency prediction	RF	[75]
		Targeted knockout efficiency prediction	CNN	[76]
		Specificity prediction	CNN/ANN	[77]
Protein parts	Transcription factors	Identification of transcription factors	CNN	[78]
		Binding sites prediction	CNN/RNN	[79]
	Intein	Split sites prediction	Logistic regression	[49]
		Insertion sites prediction	SVM	[80]

ANN: Artificial neural network; CNN: Convolutional neural network; VAE: Variational autoencoder; GAN: Generative adversarial network; LSTM: Long short-term memory; SVM: Support vector machine; RF: Random forests; RNN: Recurrent neural network.

目前，基于机器学习的生物元件功能预测已成为跨学科研究的热点，并催生了一系列创新性的科研成果。接下来，将针对几种典型的 DNA、RNA 和蛋白质调控元件，通过具体实例探讨机器学习在其功能预测中的成功应用，以展示该领域的发展动态和潜力。

2.2.1 DNA 元件的功能预测

(1) 启动子

启动子是基因表达的关键调控元件，负责启动转录过程并决定 RNA 聚合酶在 DNA 上的

起始位置。在基因线路设计中，为了精确调控关键基因的表达速率，发掘具有特定强度或活性的启动子至关重要。对于传统的启动子工程，构建、筛选和表征流程繁琐且耗时，这促使人们寻求更快捷、准确的启动子强度预测方法。

为了实现这一目标，研究人员致力于建立启动子序列与其转录活性之间的联系。早期，Jensen 等[81]展示了基于位置权重矩阵的统计方法，通过分析大肠杆菌中 69 个 $P_{L\text{-}\lambda}$ 启动子变体，定位了影响启动子强度的关键核苷酸位点。随后，De Mey 等[82]和 Liu 等[83]进一步采用偏最小二乘(partial least squares, PLS)模型分析了大肠杆菌和枯草芽孢杆菌中数百个合成启动子，揭示了启动子序列特征与活性之间的关联。然而，这些研究因数据集规模小、建模方法单一，预测准确率不高。为解决数据量不足的问题，Wang 等[84]和 Kiryu 等[85]利用大肠杆菌的转录组数据库和微阵列数据训练启动子强度预测模型，虽在一定程度上提高了预测精度，但因数据集并非专门针对启动子强度收集，模型性能仍显不足，仅能获得较低的皮尔逊相关系数(0.51–0.57)。

如图 3A 所示，为了构建一个准确且庞大的启动子活性预测专用数据集，Zhao 等[15]通过执行 83 轮突变-构建-筛选-表征(mutation-construction-screening-characterization, MCSC)流程，鉴定了 66 026 个不同的启动子突变体，并筛选出 3 665 个独特的 Trc 启动子突变体构建合成启动子库；该数据集被随机划分为训练集(90%)和测试集(10%)，基于训练集，作者运用了多种机器学习模型，包括 PLS、AdaBoost、XgBoost、RNN、RF 和梯度提升决策树等，来建立启动子强度预测模型，通过十折交叉验证来评估模型的预测性能，并在训练过程中消除多重共线性问题，使用相对强度的平均绝对误差(mean absolute error, MAE)作为衡量模型性能的指标，最终，XgBoost 模型表现最佳，具有最低的平均交叉验证 MAE (0.204)和最高的 R^2 (0.77)，能有效预测人工设计启动子序列的强度，模型的性能显著超过了上述基于位置权重矩阵、最小二乘回归以及非专业数据集的方法。

值得注意的是，上述研究多采用传统机器学习模型，这主要是因为可用于训练的突变体文库规模较为有限(小于 10^4)。然而，近年来，随着高通量实验技术的迅猛发展，特别是大规模并行报告基因检测(massively parallel reporter assays, MPRA)技术[86]的出现，构建大规模启动子突变库已成为可能，同时也推动了以 CNN 为代表的深度学习算法在启动子活性预测领域的广泛应用[14,41,61-63]。

如图 3B 所示，Van Brempt 等[61]利用 FACS-seq 技术[结合流式细胞分选(fluorescence-activated cell sorting, FACS)与二代测序(next-generation sequencing, NGS)的 MPRA 技术]，创建了一个包含大肠杆菌 σ70 依赖性启动子及枯草芽孢杆菌 σB、σF 和 σW 依赖性启动子的大型转录起始频率(transcription initiation frequency, TIF)数据库；在数据预处理阶段，研究人员剔除了非间隔区含有突变的读取样本；随后，采用深度学习方法，通过多个卷积层处理稀疏的独热(one hot)编码序列，再经丢弃层和全连接层分析以获得与启动子 TIF 相关的潜在变量；最终训练出的 CNN 模型不仅能准确预测启动子的 TIF，还能识别不同 σ 因子特异性启动子的正交性；基于这一模型，研究人员开发了一款在线启动子设计工具，使用户能够根据具体的研究需求定制启动子，显著提高了模型的实用价值。此外，MPRA 技术和 CNN 算法的结合在酿酒酵母[41]、植物[63]乃至人类细胞[62]的启动子活性预测中也取得了显著成果，这些研究充分证明了深度学习与高通量实验技术融合的巨大潜力。

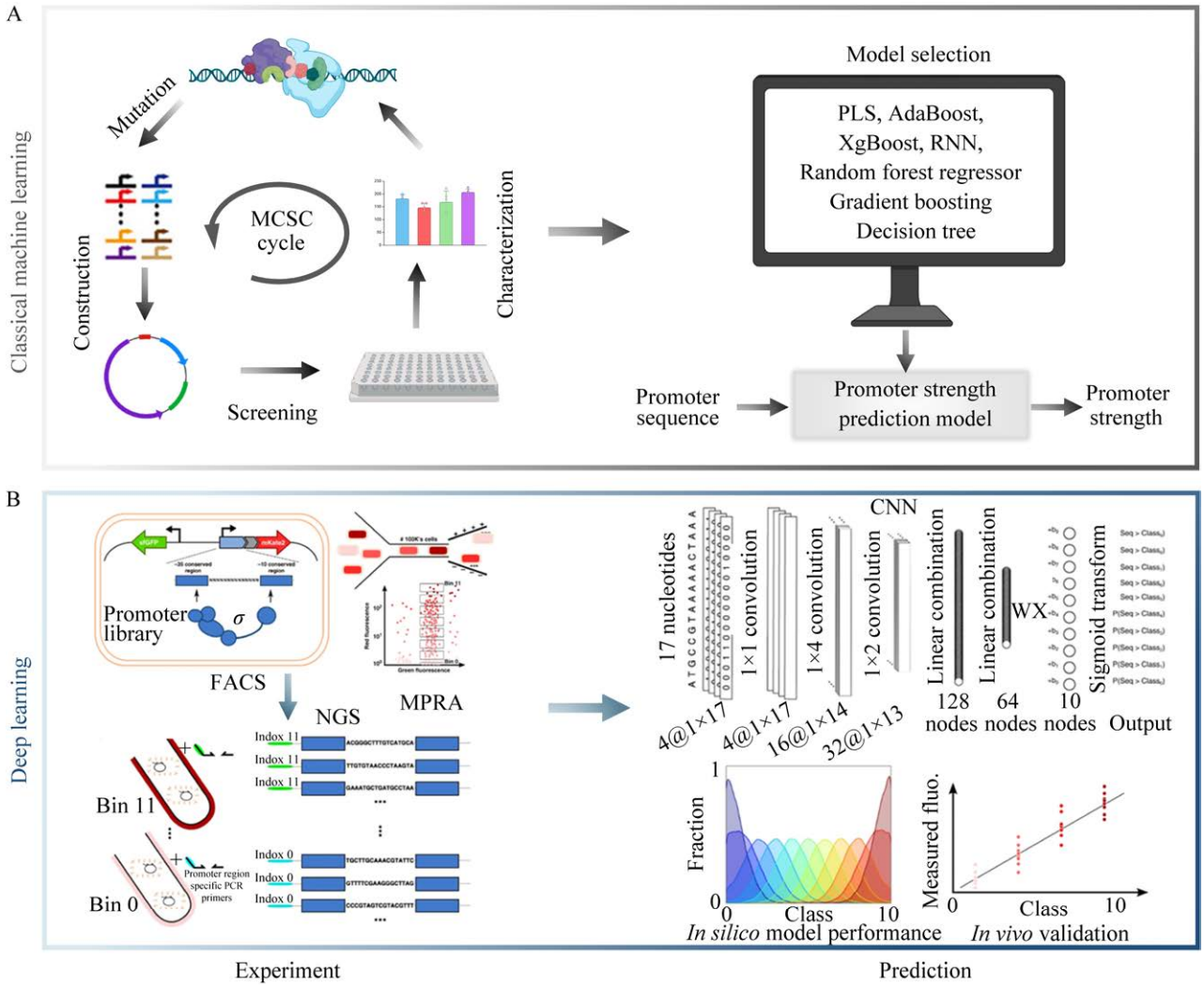


图 3　机器学习算法在启动子强度预测中的应用　A：低通量实验技术和经典机器学习算法的结合[15]；B：高通量实验技术和深度学习算法的结合[61]。

Figure 3　Applications of machine learning algorithms in promoter strength prediction. A: Integration of low-throughput experiments with classical machine learning algorithms[15]; B: Integration of high-throughput experiments with deep learning algorithms[61].

(2) 增强子

增强子是提升基因转录效率的 DNA 元件，能通过促进转录因子与启动子的结合来增强基因表达。增强子的识别及其强度的评估是合成生物元件设计的重要研究领域。

对于增强子的识别，传统方法主要有转录因子结合检测[87]和 DNA 酶 I 超敏感性分析[88]。然而，前者可能漏检未结合转录因子的增强子，导致假阴性结果；后者则可能误将非增强子区域识别为增强子，产生假阳性。

近年来，机器学习技术的发展为增强子的精准、高效识别提供了新途径。研究人员开发了多种高精度机器学习算法，如 EnhancerFinder[89]、BiRen[90]和 gkm-SVM[65]等，显著加速了识别过程并降低了成本。特别是 Liu 等[66]于 2016 年推出的 iEnhancer-2L 预测器，实现了增强子识别

及其强度评估的同步预测。随后，可以同时识别增强子并预测其强度的新机器学习框架开始层出不穷，如 iEnhancer-EBLSTM[91]、iEnhancer-ECNN[92]以及 Enhancer-IF[93]等，这些方法均表现出卓越的预测性能。最近，Butt 等[67]利用上述 iEnhancer-2L 数据集[66]，通过剔除相似序列，采用随机森林算法和十折交叉验证，实现了增强子识别(91.68%)和强度预测(84.53%)的高准确率，成为该领域目前最准确的预测方法，为基因表达调控研究提供了强大工具。

(3) 沉默子

与增强子相反，沉默子是抑制基因表达的 DNA 元件。它们通过与转录抑制因子结合，阻止转录复合物的形成或降低其稳定性，从而避免因基因过度表达导致蛋白质过剩，进而维持生物系统环境的稳定。然而，长期以来，基因表达调控研究重心多集中于增强元件，相比之下，沉默子的识别与鉴定方法较为匮乏[94]，这主要源于已发现的沉默子数量稀少及其功能机制尚不明晰。

近年来，随着研究的深入，人类、小鼠及果蝇等生物体中的沉默子逐渐被揭示，其调控机制也得以解析，使得沉默子成为计算生物学领域的研究热点[95-96]。当前，主流的沉默子识别模型包括基于经典机器学习算法的 gkm-SVM[68]和基于深度学习算法的 DeepSilencer[69]、SEPredict[70]，这些模型高度依赖于沉默子的序列特征作为输入，但沉默子的活性具有细胞特异性，仅依赖序列信息可能无法全面捕捉其在不同细胞类型中的功能差异。

随着测序技术的发展，多种生物组学数据的出现为构建基于多维度生物信号的沉默子预测模型提供了可能。Huang 等[71]基于组蛋白修饰及转录因子信息构建了 SVM 模型进行沉默子识别，尽管这一研究初步揭示了特定生物信号与沉默子之间的关联性，但它尚未充分挖掘生物信号组合在细胞特异性沉默子识别中的潜在价值，而这些组合在沉默子中更为普遍，并已在细胞特异性增强子识别中显示出重要作用[97-98]。在此背景下，Zhang 等[72]最近提出了一种名为 DeepICSH 的深度学习框架：通过整合多级序列特征、转录特征和表观遗传特征，并采用 CNN 算法捕捉与沉默子密切相关的生物信号组合，DeepICSH 可以有效识别人类基因组中的细胞特异性沉默子。研究表明，DeepICSH[72]在 HepG2 和 K562 细胞系的测试集上超越了 gkm-SVM[68]、DeepSilencer[69]和 SEPredict[70]等现有基于序列的沉默子识别方法，并能对特定细胞内的沉默子进行分类[接收者操作特征曲线下面积(area under the receiver operating characteristic curve, AUROC)的值为 0.703]，成为首个能区分强弱沉默子的计算模型。

2.2.2 RNA 元件的功能预测

(1) mRNA

mRNA 是一种单链核糖核酸分子，负责将 DNA 中的遗传信息转录并传递到核糖体，指导蛋白质的合成。其结构包括 5′帽子、5′非翻译区(5′ untranslated region, 5′UTR)、开放阅读框、3′非翻译区(3′UTR)以及多聚腺苷酸尾，这些部分共同调控 mRNA 的表达水平和稳定性。

2019 年，Sample 等[16]通过改进的 MPRA 技术，结合多聚核糖体分析与 RNA 高通量测序，评估了 280 000 种具有随机 5′UTR 序列的 mRNA 分子在 HEK293T 细胞中的平均核糖体负载(mean ribosome load, MRL)；研究人员使用其中 260 000 个带有标签的序列训练了一款基于 CNN 的深度学习模型——Optimus 5-Prime，剩余的 20 000 个序列用于性能测试。通过穷尽网格搜索优化超参数，Optimus 5-Prime 模型[16]能够接受 5'UTRs 序列的独热编码作为输入，并

准确预测mRNA分子的MRL值,准确率高达93%。

2022年，Agarwal与Kelley[73]构建了一个跨物种mRNA半衰期数据集，包含39个人类和27个小鼠样本，涉及13 000个人类和18 000个小鼠mRNA序列(由5′UTR、3′UTR和编码区构成)；利用这一数据集并结合混合卷积与循环神经网络技术，研究人员开发了Saluki模型，实现了对mRNA半衰期的高精度预测(r=0.77)。

近期，借助上述MRL[16]以及半衰期[73]数据集，清华大学汪小我团队在mRNA序列功能解析领域取得了突破性进展[17]，研究人员借助自主研发的NeuronMotif工具，构建了一套全面的RNA序列到功能模型定制管道，包括基序发现、基序贡献评估及基序间交互分析这3大核心环节，不仅能精准识别mRNA中的关键顺式作用RNA基序，还可以揭示它们之间的复杂交互关系；基于这一流程和上述两个大型数据集，研究团队开发了基于混合卷积与递归神经网络的预测系统，包括平均核糖体负载预测器(R^2=0.935)和半衰期预测器(r=0.73)，两者均展现出卓越的预测精度(图4)，其中，由于排除了小鼠数据，半衰期预测器的表现稍弱于Saluki模型[73] (r=0.77)，而平均核糖体负载预测器与上述Optimus 5-Prime平台[16] (R^2=0.934)的预测性能相当。

(2) RNA toehold开关

RNA toehold开关是一种由特定RNA序列构成的生物元件,能通过与互补的RNA序列(触发序列)结合来调整自身构象，从而控制基因表达[99]。这一机制使得RNA toehold开关能在检测到特定触发序列信号时精准调控蛋白质生产，为合成生物学提供了一种强大的基因表达调控手段。

然而在实际应用中，相当数量的toehold开关存在功能欠佳甚至完全失效的问题[100-101]。尽管有基于序列的热力学预测工具，但其预测结果与实际性能的相关系数较低，仅为0.22左

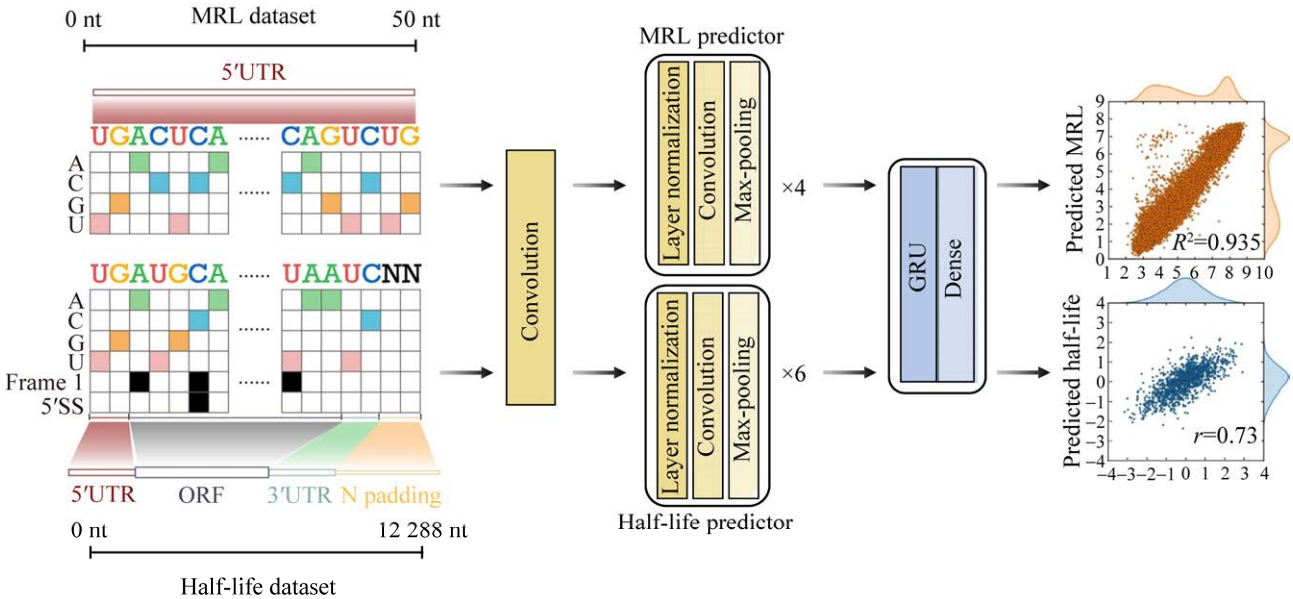


图4　基于深度学习算法预测mRNA的平均核糖体负载和半衰期[17]

Figure 4　Deep learning-based prediction of the mean ribosome load and half-life of mRNA[17].

右[102-103]，无法满足高精度设计的需求。Green等[104]通过生物物理分析确定了一系列与toehold开关性能紧密相关的热力学参数，为理性设计提供了指导。但单一参数难以全面反映toehold开关复杂的性能特征，因此需要开发一种专门针对toehold开关、仅依赖RNA序列的改进预测框架。

为解决这一问题，Angenent-Mari等[74]利用高通量DNA合成与测序技术，构建了一个包含91 534个toehold开关的大型数据库，涵盖了23个病毒基因组和906个人类转录因子，并基于此训练了一个深度神经网络模型，该模型在Toehold开关性能预测上的表现(R^2=0.43–0.70)显著优于传统的热力学模型(R^2=0.04–0.15)[100,102,105]；此外，通过将核苷酸互补矩阵作为输入，研究人员提高了模型的透明度，使其能可视化学习到的重要二级结构，从而直观地解释RNA toehold开关的性能，这种方法在准确性与可解释性方面均超越了传统的热力学建模方法。

(3) gRNA

为了更高效地改造细胞功能，除了开发用于调控基因表达的生物元件之外，还需引入基因组编辑系统(如CRISPR-Cas)以剔除冗余基因或整合外源分子。CRISPR-Cas系统通过特异性核酸酶(Cas蛋白)和短链RNA(gRNA)实现对活细胞内DNA序列的精准编辑，从而在目标位点引发双链断裂。在这个过程中，核酸酶负责DNA切割，而gRNA则负责精确地引导到特定序列。这种模块化的设计意味着仅需更新gRNA序列，就可以将其定向到新基因上，从而实现高效且经济的突变菌株构建。

在设计和优化gRNA时，研究人员重点关注其切割效率和靶向特异性。理想的gRNA应兼具高切割效率(高靶向编辑率)和低脱靶风险(高特异性)。机器学习算法能够从已表征的gRNA数据集中挖掘潜在模式，从而指导gRNA在切割效率和靶向特异性两方面的优化设计。

在切割效率的预测上，Doench等[48]和Abadi等[75]分别采用SVM和随机森林回归等传统机器学习模型来评估gRNA引导的Cas9复合体在不同基因组位点的切割效率，研究发现gRNA序列及基因座特性等因素对切割活性具有决定性影响。在深度学习领域，Kim等[76]开发的DeepCpf1工具采用了CNN算法，在大规模gRNA数据集上进行训练以预测靶向切割活性；在外部测试集中，该工具的预测性能(Spearman相关系数为0.61–0.79)明显优于传统机器学习算法(Spearman相关系数为0.34–0.44)。同样地，在脱靶预测方面，深度学习也显示出了显著优势[77]，通过在脱靶数据集CRISPOR上训练和测试CNN及ANN算法，发现这两种算法的AUROC值均超过0.97，显著高于随机森林、梯度提升树和逻辑回归等传统机器学习模型(AUROC值为0.83–0.93)。

2.2.3 蛋白质元件的功能预测

(1) 转录因子

转录因子(transcription factors, TFs)是一种具有序列特异性的DNA结合蛋白，它们通过影响RNA聚合酶的结合来调控基因转录。识别TFs对于理解生物体的转录调控至关重要。传统方法依赖于已知TFs的DNA结合域的序列同源性[106-107]。因此，那些与已报道的TFs没有同源性的TFs很难被预测出来。为了解决这一问题，Kim等[78]基于CNN算法开发了DeepTFactor预测工具，可以直接根据蛋白质序列预测其是否为转录因子，对真核和原核TFs识别的F1值分别达到0.82和0.80，在大肠杆菌K-12 MG1655基因组数据集中，DeepTFactor成功识别了332个潜在的TFs，其中3个得到了实验验证。

TFs 通过识别转录因子结合位点(transcription factor binding site, TFBS)来调控基因转录。传统预测 TFBS 的方法如位置权重矩阵(position weight matrix, PWM)和序列特征分析[108-109]，假设 TFBS 中的每个核苷酸对 TFs-DNA 的亲和力贡献是相互独立的。然而，TFs 与 DNA 之间的相互作用实际上更为复杂，包括单个氨基酸与多个碱基的互作[110]，以及通过诱导 DNA 扭曲或通过水分子间接互作[111-112]等。这些复杂相互作用使得传统预测模型难以达到较高的精度。与传统模型相比，基于机器学习的预测方法能够整合 DNA 序列信息、DNA 形状特征、染色质结构和动力学信息等多维描述符，从而显著提高 TFBS 的预测精度[113-116]。例如，Barissi 等[79]提出了基于物理描述符的机器学习方法，通过利用分子动力学模拟获得的 DNA 物理特性(碱基的静电模式、氢键和疏水性等)作为输入，准确预测了不同实验技术测定的 TFs-DNA 结合亲和力，显著超越了传统基于 PWM 和序列特征分析的方法。

(2) 内含肽

内含肽是一类可以执行蛋白质剪接反应的插入蛋白，它们在宿主蛋白质内部翻译，并巧妙地切除自身，将两侧的外显肽以近乎无痕的方式连接起来，从而形成完整的目标蛋白[117]。与单一基因编码的连续内含肽相比，由 2 个独立基因编码的断裂内含肽在基因线路设计中展现了更高的灵活性和调控能力，因此备受研究人员青睐。近期，本课题组通过合成生物学技术成功构建了一个正交断裂内含肽元件库，大幅扩展了在一个反应系统中同时使用多个断裂内含肽的可能性[118]。目前，断裂内含肽已在蛋白质连接、环化、片段同位素标记和翻译后修饰等多个领域发挥重要作用[119-122]。

随着微生物基因组测序技术的迅猛发展，目前已有数千种连续内含肽序列被揭示[123]。为了进一步挖掘这些天然内含肽的潜力，研究者们致力于通过人工断裂的方式，创造具有高效剪接活性的非天然断裂内含肽，从而丰富合成生物学工具库[124-125]。

最初，断裂位点的选择主要依赖于启发式方法[126]和简单的结构分析[127]，这通常伴随着大量的试错实验以验证断裂后的剪接活性。为简化这一过程，研究者们开发了多种计算预测工具。2012 年，Lee 等[128]提出了一种环状排列方法，它通过连接蛋白质末端残基的氨基和羧基并产生新的骨架开口，然后根据开口对蛋白质折叠的影响来评估其作为断裂位点的潜力。随后，2018 年，Dagliyan 等[129]推出了 SPELL 算法，该算法通过计算蛋白质结构的分裂能量来预测断裂位点，然而，该方法的设计目标主要用于构建化学遗传和光遗传分裂蛋白质，在其他连续内含肽断裂位点预测上的适用性仍有待进一步证实。

为了寻求更加简单、可靠且通用的解决方案，最近，Schmitz 等[49]推出了一款名为 Int&in 的机器学习模型，实现了对连续内含肽中活性与非活性分裂位点的高精度预测；研究人员首先通过蛋白质印记实验对源自 gp41-1、Npu DnaE 和 CL 内含肽的 126 个随机断裂位点进行了活性表征；接着，基于断裂内含肽的片段亲和力、溶剂可及表面积、断裂位点残基的保守性、二级结构及捕获亲和力等 5 项生物物理特征，使用逻辑回归算法训练预测模型；最终，Int&in 不仅在内部验证中达到了 0.79 的高准确度，而且在外部测试集(包含 97 个已发表文献中报道的分裂位点)上也保持了 0.78 的准确度，证实了其良好的泛化能力；这一模型的开发为新型分裂内含肽的工程化设计提供了有效工具。

2.3 智能设计非天然元件

著名理论物理学家理查德·费曼曾言："What I cannot create, I do not understand"，即通过创造，人们可以更深刻地理解事物的本质。这一理念在合成生物学中得到了生动的体现。合成生物学旨在开发能够双向工作的模型：一方面，根据给定的序列正向预测其功能；另一方面，根据所需功能反向生成全新的序列(图2)。长期以来，优质的生物元件主要源于自然界，尤其是细菌和真菌等微生物。它们的发掘过程往往依赖于大量的实验试错，耗时冗长且效率低下。随着合成生物学的发展和基因线路设计复杂性的增加，传统的元件发掘策略已难以满足研究人员对于新颖、高效元件的迫切需求。因此，合成生物学家们正积极转变思路，从被动地寻找自然界中的元件转向主动设计与创造。

过去几十年里，非深度学习方法在元件的设计和改造中发挥了重要作用。例如，Salis 等[130]通过构建基于生物物理学模型的热力学框架，成功设计出能够精确调控蛋白质表达水平的RBS。同时，基于物理学原理的蛋白质结构解析和分子动力学模拟可以精准识别特定 RNA 序列的最低自由能构象，从而指导 RNA 设计[131]。此外，定向进化策略也广泛应用于发掘具备增强功能的蛋白质变体[132]。尽管这些策略在合成生物元件设计中取得了一定成功，但同时具有设计效率低下和容易陷入局部最优解等明显缺陷。

近年来，深度生成模型，如变分自编码器、生成对抗网络和自回归模型等，在合成生物元件设计中展现了巨大的潜力，有效弥补了传统方法的局限。在 DNA 元件的生成设计中，Seo 等[20]构建了一个基于深度学习算法的蓝藻启动子生成模型框架，包括生成、预测和验证这 3 个步骤(图 5)：生成步骤中，使用蓝藻的天然启动子序列训练 VAE 模型，并生成大量合成启动子；随后，利用 CNN 模型预测启动子强度，并通过无细胞转录试验测试生成的启动子活性，结果表明，生成的启动子能够有效驱动转录。值得一提的是，清华大学汪小我教授团队最近推出了一款名为 GPro 的启动子设计工具包[64]，它提供了一套标准流程，覆盖从模型训练到优化再到性能评估的整个启动子设计周期，其用户友好的开发理念使得上述复杂的生成框架得以简便操作，预期将大幅加速定制化合成启动子的设计流程。在 RNA 元件的设计与优化方面，Valeri 等[23]将梯度提升策略与深度生成模型相结合，通过深入探索 RNA 序列的突变空间，不仅成功重塑了性能不佳的 toehold 开关，还揭示了影响其性能的关键因素，为 RNA 元件的精准调控提供了强有力的研究工具。在蛋白质元件设计领域，深度生成模型同样取得了巨大成功。例如，最近，Ni 等[133]通过对来自 PDB 数据库的 7 026 种蛋白质进行大规模的拉伸分子动力学模拟，从中提取应力信息作为输入特征，训练了一款名为 ForceGen 的蛋白质语言扩散模型，可以根据研究人员的具体需求，设计出具有指定拉伸应力特性的合成蛋白质序列。尤为值得一提的是，美国华盛顿大学蛋白设计研究所的 David Baker 团队在蛋白质生成式设计领域取得了开创性成果，这些成就使 Baker 荣获 2024 年"诺贝尔化学奖"。该团队利用深度学习模型进行蛋白质结构生成的研究已在近期的综述文献中得到详尽描述[134]，本文不再赘述。

3 基因线路的智能设计

3.1 基因线路概述

基因线路是由多个基因及其调控元件组成的网络结构，这些元件通过相互作用精确调控

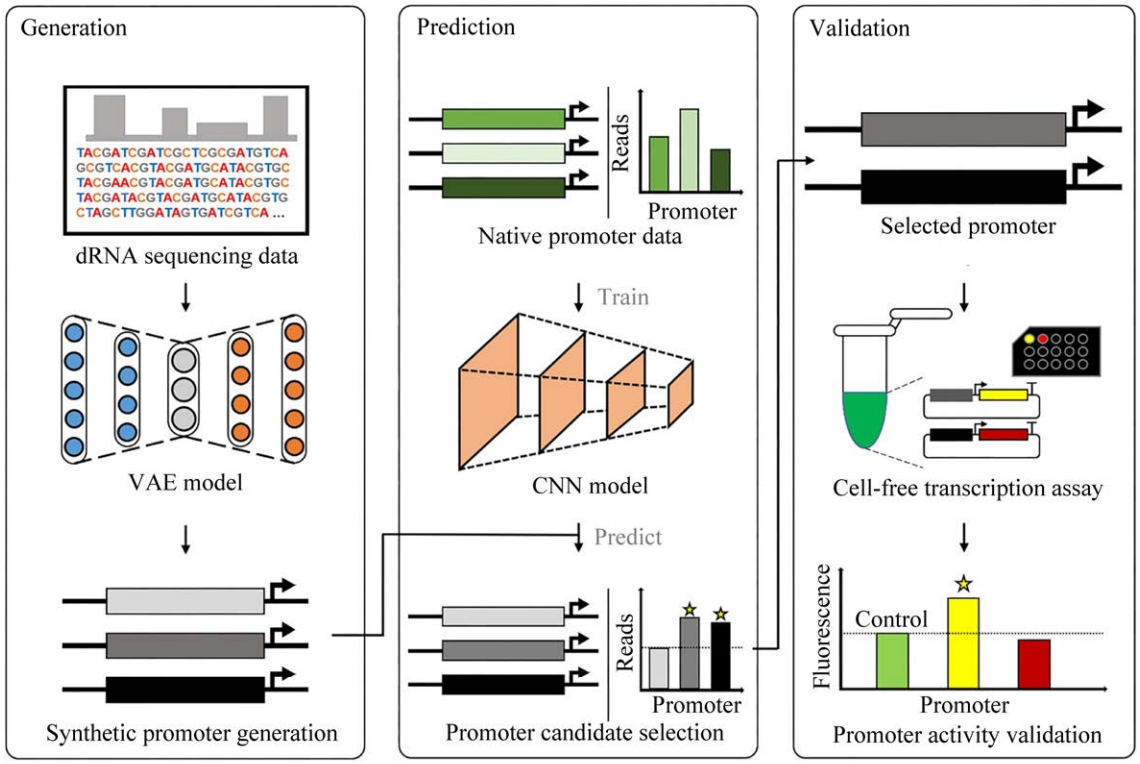


图 5　智能设计蓝藻合成启动子的深度学习框架[20]　该框架包括 3 个步骤：序列生成(左)、活性预测(中)和活性验证(右)。

Figure 5　Deep learning framework for intelligent design of cyanobacterial synthetic promoters[20]. The framework consists of three steps: sequence generation (left), activity prediction (middle), and activity validation (right).

基因表达和细胞行为。类似于电子电路中的工作原理，基因线路能接收输入信号，经由一系列生化反应产生特定的输出响应。它们在自然界中普遍存在，负责调控细胞分裂、代谢调控和信号传导等各种生物过程[135]。

在合成生物学中，研究人员通过设计和构建人工基因线路以实现特定的生物功能[136-137]。例如，本课题组开发了一种可全面优化合成基因线路设计的核酸海绵多功能基因调控工具[138]，并通过级联多层正交模块化的转录信号放大线路，设计出了超敏感的基于大肠杆菌底盘的砷和汞等重金属污染合成细胞传感器[139]。基因线路在生物工程和生物医学等领域具有广泛的应用潜力。然而，传统设计与构建方法通常依赖于对特定生物通路的深刻理解，并涉及大量元件及其参数的组合与筛选，以及反复实验以验证其性能[140]。这一过程耗时且低效，往往需要多次迭代才能确定最佳的元件及参数组合。

近年来，随着计算生物学和机器学习技术的发展，研究人员开始采用更先进的方法来设计基因线路，主要包括 2 个方面：一是利用机器学习算法预测基因线路的行为(功能)；二是通过计算模拟自动优化元件及其参数组合。接下来将分别探讨这 2 个方面，展示机器学习算法在基因线路设计中的前沿应用及其潜在优势。

3.2　基因线路功能预测

在合成生物学中，由于细胞的复杂动态

特性，精确预测基因线路的行为一直颇具挑战[141]。然而，近年来机器学习的引入为此提供了创新解决方案。2021 年，Zhu 等[142]率先将机器学习算法应用于多顺反子基因线路的基因表达水平预测，使用 RF 与 ANN 算法构建的回归模型成功预测了基因表达强度，R^2 分数高达 0.87 和 0.95；此外，他们还构建了分类模型，实现了对合成基因线路模式的完美识别，准确率达 100%。这些模型有望帮助研究人员在合成基因线路设计中减少试错成本，并快速筛选出最优元件组合。

2022 年，Daniels 等[143]将机器学习应用于 T 细胞中嵌合抗原受体(chimeric antigen receptor, CAR)的功能预测，研究人员构建了一个包含约 2 300 个合成共刺激结构域的 CARs 文库，通过病毒转导技术将这些 CAR 导入原代 T 细胞中，并系统评估了细胞的自我更新能力和细胞毒性；基于这些丰富的实验数据，他们开发了一个融合 CNN 与 LSTM 的深度学习模型，能准确预测新合成的共刺激域组合对细胞表型的影响，为 CAR-T 细胞疗法在癌症治疗领域的定制化设计提供了实用工具。

最近，Qin 等[144]研究了 T7 RNA 聚合酶激活模块与阻遏蛋白模块在非模式微生物中的跨物种应用，发现这些模块在不同生物体中的活性具有相关性，因此推测其在多种生物背景下的表现可能具有一定的可预测性；基于这一假设，研究人员构建了一个包含非特异性宿主参数的组合模型，并成功实现了对线路行为在不同生物背景下的定量预测，这一成果为合成生物学的跨物种研究提供了新的视角。此外，Rai 等[145]推出了一种新型表征平台——CLASSIC，该平台结合了长读和短读测序技术，能够在人类细胞中进行大规模并行分析合成基因线路，一次性实验即可解析超过十万种药物诱导的基因线路设计表达；利用 CLASSIC 平台积累的大量数据，研究人员训练了随机森林模型，该模型不仅能根据基因线路的元件组成精确预测表达输出，还能揭示基因线路设计的隐含规则，为合成基因线路的优化设计提供了强大的研究工具。

3.3 基因线路的智能设计

随着对基因线路及其调控规则的深入了解，研究人员日益重视对基因线路的人工设计和调整，以实现对细胞行为在时间和空间上的精准调控。为此，合成生物学家借鉴电子设计自动化的思想，构建了基因设计自动化框架。自 2009 年以来，多个强大工具相继问世，包括 Peccoud 团队的 GenoCAD (用于从基因元件构建基因线路的网络应用程序)[146]、Nguyen 团队的 iBioSim (支持合成基因线路学习、分析与设计的计算平台)[147]以及 Voigt 团队的 Cello (通过硬件描述语言 Verilog 定义基因线路功能，并从布尔逻辑门库中自动生成特定功能的 DNA 序列)[148]，这些自动化平台极大程度地简化了基因线路设计流程，提高了设计效率和准确性。

然而，生物元件并非孤立运作，基因线路设计的复杂性不仅体现在线路架构(布线图)的构建上，更在于精确调控元件之间错综复杂的相互作用强度。随着基因线路规模的扩张，元件间的互作关系急剧复杂化：一个包含 N 个基因的线路，其潜在的互作模式可达 N^2种，即需要面对至少 N^2个参数的调控挑战。在如此高维的参数空间中精准定位使线路性能最优的参数组合，无疑是一项艰巨的任务。为应对这一挑战，近期研究尝试将机器学习技术融入基因线路设计流程中，旨在高效探索其参数设计空间并获得最佳参数组合[149-152]。

如图 6A 所示，Merzbacher 等[151]将混合整数贝叶斯优化技术应用于基因线路的参数优化

中，通过学习基因线路性能景观并迭代搜索参数组合空间，该技术能够智能地确定哪些参数(代谢物和酶的浓度、代谢物与基因表达调控元件之间的相互作用强度等)最有可能产生最优的线路设计；相比于传统的穷举法或随机搜索，这种方法具有更高的效率和准确度，特别适合处理包含大量变量和复杂相互作用的大型基因线路设计问题。此外，Hiscock[152]基于梯度下降法开发了 GeneNet 工具(图 6B)，该工具通过定义与基因线路输出目标紧密相关的成本函数，并运用 Adam 优化算法迭代调整基因参数(包括转录因子浓度、基因互作强度及降解速率等)，以实现成本函数的最小化，从而锁定最佳参数组合；实践证明，GeneNet 可以成功设计出执行振荡、脉冲检测和生物计数等具有不同复杂度的基因线路。

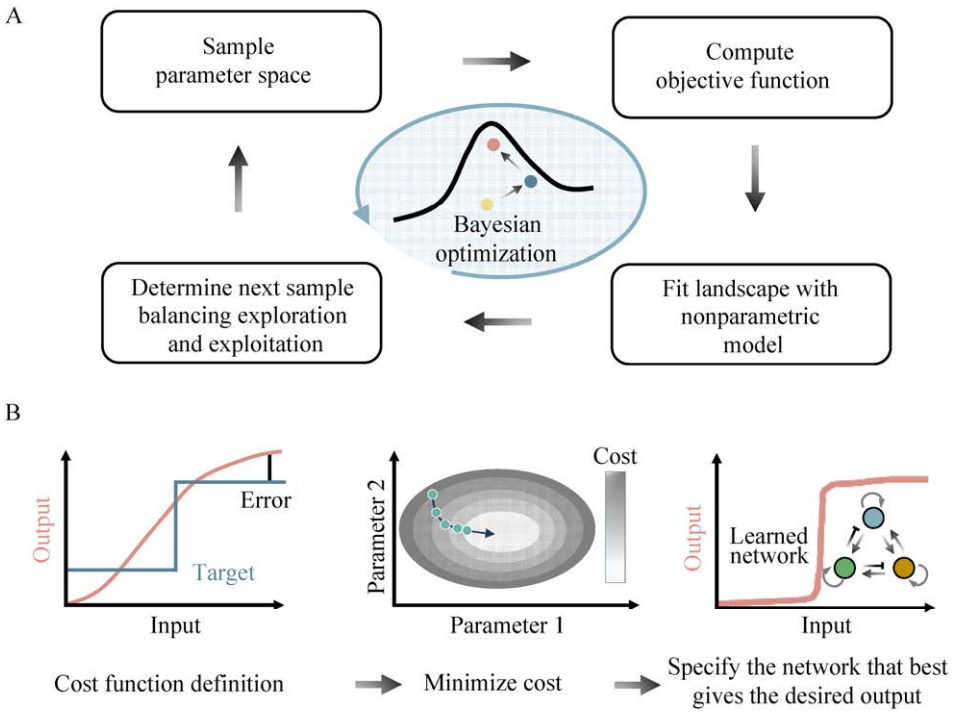


图 6　两种基因线路智能设计算法的工作流程　A：基于贝叶斯优化技术学习基因线路性能景观的形状，并迭代地在参数组合空间中搜寻最优线路[151]。B：GeneNet 算法工作流程[152]。首先，根据基因线路的实际输出与目标输出之间的差异，定义成本函数(左侧)；然后，通过梯度下降法，不断迭代调整基因的参数，以最小化成本函数(中间)；最后，分析学习到的参数网络，找到基因线路的最佳参数(右侧)。

Figure 6　Workflow of two intelligent design algorithms for genetic circuits. A: Learning the shape of the performance landscape of genetic circuits based on Bayesian optimization techniques and iteratively searching for the optimal circuit in the parameter combination space[151]. B: Workflow of the GeneNet algorithm[152]. Firstly, a cost function is defined based on the difference between the actual output of the genetic circuit and the target output (left); Then, through the application of gradient descent, the parameters of the genes are iteratively adjusted to minimize the aforementioned cost function (middle); Finally, the learned parameter network is analyzed to identify the optimal parameters for the genetic circuit (right).

目前，机器学习在基因线路设计领域的应用尚处于起步阶段，但其展现出的巨大潜力已经引起合成生物学家的广泛关注。研究人员期望可以在实验前进行有效的计算机筛选，以提前淘汰掉表现不佳或鲁棒性差的线路模式，从而大大节省后续实验的时间和资源，提高整体研究效率。随着机器学习技术与合成生物学的加速融合以及相关研究的持续推进，有理由相信，机器学习将在基因线路设计领域发挥越来越重要的作用，推动该领域迈向新的高度。

4 挑战

正如前文所述，人工智能与合成生物学的深度融合已为元件和基因线路设计领域带来巨大变革，其应用潜力引起了广泛关注。然而，这一融合过程也面临着来自多个维度的挑战。本文将这些挑战归纳为数据、模型和社会影响这 3 个核心层面，并对此进行简要概述，同时展望未来克服这些挑战的可能方向(图 7)。

4.1 数据挑战

4.1.1 数据规模挑战

高效的机器学习模型需要大量的数据进行训练。例如，在图像处理领域，模型可以利用数百万个样本进行训练[153]。但在合成生物学中，获取大规模的数据集通常非常困难，大多数实验涉及的元件或线路数量远少于 100 个，无法满足机器学习对大数据的需求。此外，合成生物学数据集中普遍存在的不平衡性也限制了机器学习的应用[154]。如，在处理基于不平衡数据集的分类任务时，模型可能过度关注多数类特征，忽视少数类，从而降低整体预测性能。

为解决这些问题，合成生物学家和计算科学家需共同努力。计算科学家可以通过数据重采样、设置类别权重或发现更具区分度的特征等策略，减轻数据稀缺和不平衡对模型性能的影响。例如，Zhang 等[18]通过实施分层等比例抽样策略和引入样本权重机制，成功削弱了样本不平衡对模型训练的不利影响，显著提升了

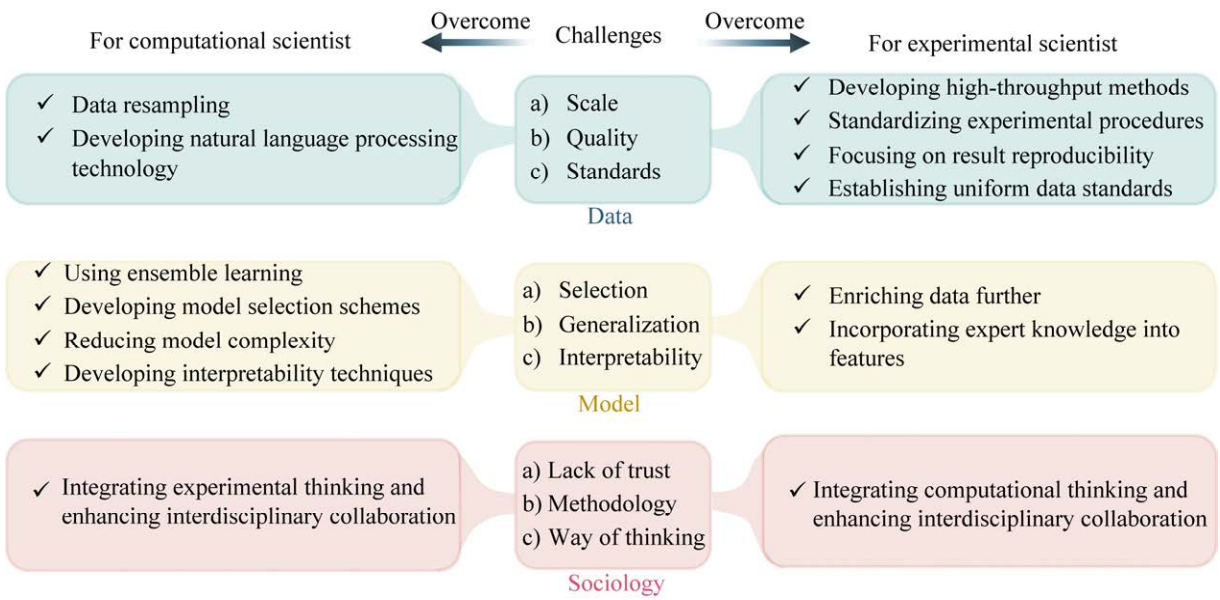


图 7　机器学习与合成生物学融合所面临的挑战及对应解决方案

Figure 7　Challenges and potential solutions for the integration of machine learning and synthetic biology.

酶分类预测器的性能。同时，合成生物学家应专注于开发适用于不同场景的高通量实验方法，如 MPRA[86]、uASPIre[155]和 CLASSIC[145]等。这些方法旨在高效、经济地产生高质量、低噪声的训练数据集，以供机器学习算法训练。值得注意的是，随着机器人辅助实验自动化技术的进步，这一目标有望更快实现[156]。如，美国 OpenTrons 公司开发的 OT-2 液体处理机器人和外部热循环仪已协同实现了 DNA 组装自动化流程，用户仅需简单地调用 Python 应用程序接口，即可轻松完成整个操作过程[157]。国内，浙江大学杭州国际科创中心的合成生物学自动化装置(iBioFoundry)是全球合成生物学研究领域规模最大、集成度最高的一套全流程、高通量、自动化科学装置，能够大幅度提高人工细胞构建效率(平均 2 000 个/d)，为数据的快速生成、智能分析和精确调控提供了有力支持。

4.1.2 数据质量挑战

机器学习对数据质量有严格要求。高质量数据应具备高可重复性，即在相同实验条件下获得一致的结果。然而，生物学研究中存在严重的不可重复性问题。2016 年 *Nature* 杂志的调查显示，超过 70%的生物学研究结果无法被重复，且近 60%的研究人员无法重复自己的发现[158]。这一问题削弱了生物学研究的可靠性和可信度，因为可重复性是科学方法有效性的根本前提。当人工智能技术介入时，使用不可重复的数据进行训练会降低模型的泛化能力，甚至削弱对计算与实验结合研究的信心。

为解决这一问题，研究人员需提高实验设计的严谨性，采用标准化的方法和流程，确保数据收集和分析的一致性。科研机构可以建立内部质量控制机制，鼓励或要求研究人员分享详细的实验方法和数据，以提高研究的透明度。期刊应要求作者提供详细的实验细节和数据，便于其他研究者验证和复现研究结果。资助机构也可以在资助条件中加入对研究可重复性的要求，鼓励开放科学实践，从而提升研究的可重复性和模型训练数据的质量。

4.1.3 数据标准挑战

尽管可以从多个来源收集大量生物数据，但由于缺乏统一的数据标准，这些数据往往难以直接用于机器学习[159]。一方面，从文献中提取的元件或线路样本通常包含许多变量。以内含肽的活性表征为例，样本数据可能来源于不同的生物系统(如细菌、真核生物和病毒等)，或使用不同的表征方法(如蛋白质印迹或荧光表征)[118,160]。另一方面，不同的研究组和数据库往往以不同的模式展示数据，使数据整合和结构化变得困难。这导致计算研究者需要花费大量时间(约 50%到 80%[161])进行数据收集、清洗、对齐和转换，从而分散了对科学问题深入探究的精力。

为了解决这一问题，科研社区应制定可广泛接受的数据标准，涵盖从实验设计到数据采集与分析的各个环节，确保数据的质量和标准化程度，并鼓励研究人员和机构分享数据。同时，数据库上传或管理者需着力提升各数据库间的互操作性，例如，通过构建数据映射机制，确保不同来源的数据能够无缝对接与整合。其中，自然语言处理技术在此过程中有望扮演加速器的角色，协助实现数据的自动和半自动化处理。

4.2 模型挑战

4.2.1 模型选择挑战

在机器学习领域，选择最合适的模型进行预测是一项挑战，需要对问题本身、数据特性及模型性能有深入理解。此外，即使是同一种模型，参数的选择也会显著影响模型性能。

为应对这一挑战，可以采取 2 种策略：首

先，使用集成学习方法，如随机森林和梯度提升树，这些方法通过结合多个基础模型的预测，通常能提供比单一模型更优的预测性能。其次，借鉴新药研发领域的虚拟筛选策略，可制定有效的模型筛选方案，根据问题类型(分类、回归、聚类)、数据特征(大小、维度、分布)和预期目标(准确性、运行速度、可解释性)进行初步筛选。例如，对于小规模数据集，线性模型或决策树可能适用；而对于大规模、高维数据，深度学习模型可能更合适。然后，利用网格搜索、随机搜索或贝叶斯优化等方法系统地调整模型参数，以开发出最适合特定问题和数据集的模型。

4.2.2 模型泛化能力挑战

模型的泛化能力是指其在未见过的数据上的预测能力，衡量的是模型从训练数据中学习到的知识是否能够有效地应用于新的、未知的数据集。一个具有良好泛化能力的模型能够在不同的数据集上保持稳定的性能，而不仅仅是对训练数据过拟合。然而，在生物学研究中，复杂多变的生物系统对预测模型的泛化能力提出了严峻挑战。例如，合成生物元件的性能不仅由其序列决定，还受宿主细胞类型、生长条件、外部环境等多种因素影响，导致同一生物元件在不同遗传、理化环境中可能表现出不同的活性。因此，在实际应用中，模型往往在特定实验条件下表现优异，一旦迁移至新环境，其性能可能大幅下降，这种现象被称为“数据泄露”[162]。

为解决这一问题，可以从数据层面入手，通过广泛收集涵盖多种宿主细胞、生长条件及外部环境的数据集，为模型提供全面、多样化的训练数据，使其在多变的环境中仍能保持稳定的预测性能。在算法设计上，可以降低模型的复杂度，如减少神经网络的层数与节点数，或使用正则化技术来约束模型参数，避免过度拟合训练数据中的噪声与异常值，关注数据的普遍特征，以增强其泛化能力。

4.2.3 模型可解释性挑战

模型可解释性是指机器学习模型对其决策过程的解释能力，这对于合成生物元件或基因线路设计尤为重要。这些任务不仅要求模型能够提供准确的预测结果，还需要提供充分的解释，以便研究人员理解模型的决策逻辑，从而优化设计。然而，像深度学习这样的先进模型，其复杂的内部结构和庞大的参数集往往构成“黑箱”，限制了直观解释的可能性。例如，虽然神经网络可以高精度预测酶活性，但往往无法解释哪些具体因素导致酶活性变化，这阻碍了研究人员对酶功能的进一步调整和优化。

为解决这一问题，研究人员正在积极开发多种可解释机器学习模型工具和技术[163]。这些方法包括：特征重要性分析[164]，有助于确定影响预测结果的关键因素；Local Interpretable Model-agnostic Explanations (LIME)[165]，通过在特定预测实例的局部邻域内生成扰动样本并拟合简化模型来解释单个结果；以及 SHapley Additive exPlanations (SHAP)[166]，利用博弈论中的 Shapley 值量化特征对模型输出的贡献。此外，除了算法层面，将专家经验融入特征选择过程也是提高模型可解释性的有效途径。例如，在酶活性预测中，除了使用蛋白质序列，还可以根据生物学领域的专业知识选择与酶活性相关的特征作为模型输入，这不仅能提升模型的预测性能，还能帮助捕捉酶活性变化的内在规律，提高模型的可解释性[167]。

4.3 社会学挑战

在人工智能与合成生物学的融合过程中，社会学挑战往往比技术挑战更为复杂且容易被忽视，这些挑战主要源于实验科学家和计算科学家这 2 个不同群体之间的专业知识融合需求。

首先，这 2 个群体在思维方式上存在本质

区别。实验科学家通常采取从具体到抽象的思考方式，偏好对生物现象进行定性解释，并通过不断调整实验条件来逼近真理，强调实验证据的直接性；相比之下，计算科学家的思维方式更加抽象和理论化，倾向于利用计算模拟和统计分析来预测和理解复杂系统的行为，致力于实现生物现象的自动化、标准化定量预测，即使有时无法提供充分的实验证据。这种思维上的差异可能导致双方在合作中出现沟通不畅和理解困难的问题。

其次，方法论上的差异也是跨学科合作面临的重要挑战。实验科学家多采用实证主义方法，通过观察、实验和归纳来获取结果；而计算科学家则倾向于假设-演绎的方法，通过提出假设、建立模型和验证假设来推动研究。这 2 种方法论在本质上是互补的，但在实践操作中常常难以协调而导致分歧。

此外，跨学科合作中还存在信任与认可的问题。实验科学家可能对计算科学家提出的模型和预测结果持怀疑态度，认为这些结果缺乏实验验证的可靠性和直接性，甚至可能视为资源的浪费；而计算科学家可能认为实验科学家的研究方法过于保守和局限，未能充分利用数据和信息技术的优势，导致研究效率低下。这种相互间的不信任和不认可进一步增加了合作的难度。

为克服这些社会学障碍，促进人工智能与合成生物学的有效融合，需要加强跨学科教育，提高研究人员对对方领域的认可和理解，促进思维方式的融合。此外，应鼓励组建由实验科学家和计算科学家共同参与的研究团队，通过团队合作来克服单一学科的局限性，推动跨学科创新的实现。

5 总结与展望

合成生物学作为一个前沿的跨学科领域，正经历着由传统实验范式向智能化设计范式的深刻转变。面对生物系统的复杂性和高维度的生化特征带来的挑战，机器学习等人工智能技术为生物元件和基因线路的设计提供了创新性的解决方案。这些技术能够有效降低实验成本，提升研发效率和成功率。尽管实践中仍面临来自数据、模型和社会层面的多重挑战，但通过推动标准化数据平台和工具的开发、优化算法，并加强跨学科合作，这些问题有望得到逐步解决。

值得注意的是，机器学习与合成生物学融合的重大突破往往依赖于高通量实验方法的成功开发。同时，基于大数据的深度学习模型在预测性能上已显著优于经典机器学习算法。因此，对于合成生物学家而言，开发高通量实验方法或建立自动化生物工厂以降低获取庞大表型数据库的成本成为关键任务。然而，在这些新技术尚未普及之前，计算科学家仍可以通过创新思路，如从多维角度提取数据特征，来获得可靠的预测结果。正如哥伦比亚大学的计算生物学家 Mohammed AlQuraishi 所预测，将物理先验知识和计算模拟高效嵌入到机器学习框架中，可能会成为计算合成生物学的下一个突破口[168]。近期，基于序列、结构、热力学和动力学等多维特征的预测模型的成功开发，正是这一趋势的有力证明[19,169-173]。

长远来看，机器学习与合成生物学的深度融合，将促进生物元件与基因线路的设计从一种“试错”主导的过程转变为“预测-验证”的迭代优化模式。这不仅意味着更快的创新周期和更高的设计成功率，也将开启一个全新的研究范式，其中计算机辅助设计将成为标准实践，推动合成生物学向更加个性化、智能化和自动化的方向发展。

先进的计算模拟、人工智能和高通量实验

技术有望提高合成生物系统的可预测性，实现对生物系统的精准设计和控制，从而更好地理解和开发人造生命系统。这将为解决能源危机、环境保护和医疗健康等全球性挑战开辟新道路，引领生物科技迈向充满无限可能的新纪元。

作者贡献声明

毛瑞超：方案设计、实验操作、数据管理、初稿写作；王宝俊：方案设计、提供材料、经费支持、监督指导、稿件润色修改。

作者利益冲突公开声明

作者声明没有任何可能会影响本文所报告工作的已知经济利益或个人关系。

REFERENCES

[1] BUSON F, GAO YL, WANG BJ. Genetic parts and enabling tools for biocircuit design[J]. ACS Synthetic Biology, 2024, 13(3): 697-713.

[2] WANG BJ, BARAHONA M, BUCK M. A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals[J]. Biosensors & Bioelectronics, 2013, 40(1): 368-376.

[3] LIU Y, PINTO F, WAN XY, YANG ZG, PENG SG, LI MX, COOPER JM, XIE Z, FRENCH CE, WANG BJ. Reprogrammed tracrRNAs enable repurposing of RNAs as crRNAs and sequence-specific RNA biosensors[J]. Nature Communications, 2022, 13: 1937.

[4] LI HS, ISRANI DV, GAGNON KA, GAN KA, RAYMOND MH, SANDER JD, ROYBAL KT, JOUNG JK, WONG WW, KHALIL AS. Multidimensional control of therapeutic human cell function with synthetic gene circuits[J]. Science, 2022, 378(6625): 1227-1234.

[5] WANG BJ, BARAHONA M, BUCK M. Amplification of small molecule-inducible gene expression *via* tuning of intracellular receptor densities[J]. Nucleic Acids Research, 2015, 43(3): 1955-1964.

[6] CHEN Y, KIM JK, HIRNING AJ, JOSIĆ K, BENNETT MR. Emergent genetic oscillations in a synthetic microbial consortium[J]. Science, 2015, 349(6251): 986-989.

[7] TAN X, LETENDRE JH, COLLINS JJ, WONG WW. Synthetic biology in the clinic: engineering vaccines, diagnostics, and therapeutics[J]. Cell, 2021, 184(4): 881-898.

[8] PATRA P, DISHA BR, KUNDU P, DAS M, GHOSH A. Recent advances in machine learning applications in metabolic engineering[J]. Biotechnology Advances, 2023, 62: 108069.

[9] PRESNELL KV, ALPER HS. Thermodynamic and first-principles biomolecular simulations applied to synthetic biology: promoter and aptamer designs[J]. Molecular Systems Design & Engineering, 2018, 3(1): 19-37.

[10] VOLK MJ, LOURENTZOU I, MISHRA S, VO LT, ZHAI CX, ZHAO HM. Biosystems design by machine learning[J]. ACS Synthetic Biology, 2020, 9(7): 1514-1533.

[11] RAI K, WANG YD, O'CONNELL RW, PATEL AB, BASHOR CJ. Using machine learning to enhance and accelerate synthetic biology[J]. Current Opinion in Biomedical Engineering, 2024, 31: 100553.

[12] GOSHISHT MK. Machine learning and deep learning in synthetic biology: key architectures, applications, and challenges[J]. ACS Omega, 2024, 9(9): 9921-9945.

[13] MERZBACHER C, OYARZÚN DA. Applications of artificial intelligence and machine learning in dynamic pathway engineering[J]. Biochemical Society Transactions, 2023, 51(5): 1871-1879.

[14] VAISHNAV ED, DE BOER CG, MOLINET J, YASSOUR M, FAN L, ADICONIS X, THOMPSON DA, LEVIN JZ, CUBILLOS FA, REGEV A. The evolution, evolvability and engineering of gene regulatory DNA[J]. Nature, 2022, 603(7901): 455-463.

[15] ZHAO M, YUAN ZQ, WU LT, ZHOU SH, DENG Y. Precise prediction of promoter strength based on a *de novo* synthetic promoter library coupled with machine learning[J]. ACS Synthetic Biology, 2022, 11(1): 92-102.

[16] SAMPLE PJ, WANG B, REID DW, PRESNYAK V, MCFADYEN IJ, MORRIS DR, SEELIG G. Human 5' UTR design and variant effect prediction from a massively parallel translation assay[J]. Nature Biotechnology, 2019, 37(7): 803-809.

[17] ZENG XC, WEI Z, DU QX, LI JQ, XIE Z, WANG XW. Unveil *cis*-acting combinatorial mRNA motifs by interpreting deep neural network[J]. Bioinformatics, 2024, 40(Suppl 1): i381-i389.

[18] ZHANG QF, ZHENG WL, SONG ZD, ZHANG Q, YANG LR, WU JP, LIN JP, XU G, YU HR. Machine learning enables prediction of pyrrolysyl-tRNA synthetase substrate specificity[J]. ACS Synthetic Biology, 2023, 12(8): 2403-2417.

[19] ELIA VENANZI NA, BASCIU A, VARGIU AV, KIPARISSIDES A, DALBY PA, DIKICIOGLU D. Machine learning integrating protein structure, sequence, and dynamics to predict the enzyme activity of bovine enterokinase variants[J]. Journal of Chemical Information and Modeling, 2024, 64(7): 2681-2694.

[20] SEO E, CHOI YN, SHIN YR, KIM D, LEE JW. Design of synthetic promoters for cyanobacteria with generative deep-learning model[J]. Nucleic Acids Research, 2023, 51(13): 7071-7082.

[21] LINDER J, BOGARD N, ROSENBERG AB, SEELIG G. A generative neural network for maximizing fitness and diversity of synthetic DNA and protein

sequences[J]. Cell Systems, 2020, 11(1): 49-62.
[22] ZHANG PC, WANG HC, XU HW, WEI L, LIU LY, HU ZR, WANG XW. Deep flanking sequence engineering for efficient promoter design using DeepSEED[J]. Nature Communications, 2023, 14: 6309.
[23] VALERI JA, COLLINS KM, RAMESH P, ALCANTAR MA, LEPE BA, LU TK, CAMACHO DM. Sequence-to-function deep learning frameworks for engineered riboregulators[J]. Nature Communications, 2020, 11: 5058.
[24] WAN FP, KONTOGIORGOS-HEINTZ D, DE LA FUENTE-NUNEZ C. Deep generative models for peptide design[J]. Digital Discovery, 2022, 1(3): 195-208.
[25] TUCS A, TRAN DP, YUMOTO A, ITO Y, UZAWA T, TSUDA K. Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks[J]. ACS Omega, 2020, 5(36): 22847-22851.
[26] ASPERTI A, EVANGELISTA D, LOLI PICCOLOMINI E. A survey on variational autoencoders from a green AI perspective[J]. SN Computer Science, 2021, 2(4): 301.
[27] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model[J]. Artificial Intelligence Review, 2020, 53(8): 5929-5955.
[28] CLIFTON J, LABER E. Q-learning: theory and applications[J]. Annual Review of Statistics and Its Application, 2020, 7: 279-301.
[29] PETERS J. Policy gradient methods[J]. Scholarpedia, 2010, 5(11): 3698.
[30] FRANÇOIS-LAVET V, HENDERSON P, ISLAM R, BELLEMARE MG, PINEAU J. An introduction to deep reinforcement learning[J]. Foundations and Trends® in Machine Learning, 2018, 11(3/4): 219-354.
[31] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C]. Proceedings of the 33rd annual meeting on Association for Computational Linguistics. USA: Association for Computational Linguistics, 1995.
[32] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]. Proceedings of the eleventh annual conference on Computational learning theory. Madison Wisconsin USA, 1998.
[33] MIYATO T, MAEDA SI, KOYAMA M, ISHII S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.
[34] TARVAINEN A, VALPOLA H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. Advances in Neural Information Processing Systems, 2017, 30: 1195-1204.
[35] NGUYEN VL, SHAKER MH, HÜLLERMEIER E. How to measure uncertainty in uncertainty sampling for active learning[J]. Machine Learning, 2022, 111(1): 89-122.
[36] HUANG SJ, JIN R, ZHOU ZH. Active learning by querying informative and representative examples[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(10): 1936-1949.
[37] CAI WB, ZHANG MH, ZHANG Y. Active learning for ranking with sample density[J]. Information Retrieval Journal, 2015, 18(2): 123-144.
[38] YOSINSKI J, CLUNE J, BENGIO Y, LIPSON H. How transferable are features in deep neural networks?[J]. Advances in Neural Information Processing Systems, 2014, 4: 3320-3328.
[39] KIM J, OWEN-SMITH J. Model reuse in machine learning for author name disambiguation: an exploration of transfer learning[J]. IEEE Access, 2020, 8: 188378-188389.
[40] 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.
[41] KOTOPKA BJ, SMOLKE CD. Model-driven generation of artificial yeast promoters[J]. Nature Communications, 2020, 11: 2113.
[42] KLEPIKOVA AV, KASIANOV AS, GERASIMOV ES, LOGACHEVA MD, PENIN AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling[J]. The Plant Journal, 2016, 88(6): 1058-1070.
[43] TRELOAR NJ, FEDOREC AJH, INGALLS B, BARNES CP. Deep reinforcement learning for the control of microbial co-cultures in bioreactors[J]. PLoS Computational Biology, 2020, 16(4): e1007783.
[44] GOPAKUMAR V, TIWARI S, RAHMAN I. A deep learning based data driven soft sensor for bioprocesses[J]. Biochemical Engineering Journal, 2018, 136: 28-39.
[45] MELLOR J, GRIGORAS I, CARBONELL P, FAULON JL. Semisupervised Gaussian process for automated enzyme search[J]. ACS Synthetic Biology, 2016, 5(6): 518-528.
[46] BORKOWSKI O, KOCH M, ZETTOR A, PANDI A, BATISTA AC, SOUDIER P, FAULON JL. Large scale active-learning-guided exploration for *in vitro* protein production optimization[J]. Nature Communications, 2020, 11: 1872.
[47] KRAUS OZ, GRYS BT, BA J, CHONG Y, FREY BJ, BOONE C, ANDREWS BJ. Automated analysis of high-content microscopy data with deep learning[J]. Molecular Systems Biology, 2017, 13(4): 924.
[48] DOENCH JG, FUSI N, SULLENDER M, HEGDE M, VAIMBERG EW, DONOVAN KF, SMITH I, TOTHOVA Z, WILEN C, ORCHARD R, VIRGIN HW, LISTGARTEN J, ROOT DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9[J]. Nature Biotechnology, 2016, 34: 184-191.
[49] SCHMITZ M, BALLESTIN JB, LIANG JS, TOMAS F, FREIST L, VOIGT K, Di VENTURA B, ALI ÖZTÜRK M. Int&in: a machine learning-based web server for active split site identification in inteins[J]. Protein Science, 2024, 33(6): e4985.
[50] OPGENORTH P, COSTELLO Z, OKADA T, GOYAL

G, CHEN Y, GIN J, BENITES V, de RAAD M, NORTHEN TR, DENG K, DEUTSCH S, BAIDOO EEK, PETZOLD CJ, HILLSON NJ, GARCIA MARTIN H, BELLER HR. 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.
[51] DE FERRARI L, MITCHELL JBO. From sequence to enzyme mechanism using multi-label machine learning[J]. BMC Bioinformatics, 2014, 15: 150.
[52] PETERSON L. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.
[53] ELMELIGY A, MEHRANI P, THIBAULT J. Artificial neural networks as metamodels for the multiobjective optimization of biobutanol production[J]. Applied Sciences, 2018, 8(6): 961.
[54] ZHOU YK, LI G, DONG JK, XING XH, DAI JB, ZHANG C. 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.
[55] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep Learning[M]. London: MIT Press, 2016.
[56] JANIESCH C, ZSCHECH P, HEINRICH K. Machine learning and deep learning[J]. Electronic Markets, 2021, 31(3): 685-695.
[57] SHARIFANI K, AMINI M. Machine learning and deep learning: a review of methods and applications[J]. World Information Technology and Engineering Journal, 2023, 10(7): 3897-3904.
[58] WANG BJ, KITNEY RI, JOLY N, BUCK M. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology[J]. Nature Communications, 2011, 2: 508.
[59] WANG BJ, BARAHONA M, BUCK M. Engineering modular and tunable genetic amplifiers for scaling transcriptional signals in cascaded gene networks[J]. Nucleic Acids Research, 2014, 42(14): 9484-9492.
[60] MENG HL, WANG JF, XIONG ZQ, XU F, ZHAO GP, WANG Y. Quantitative design of regulatory elements based on high-precision strength prediction using artificial neural network[J]. PLoS One, 2013, 8(4): e60288.
[61] VAN BREMPT M, CLAUWAERT J, MEY F, STOCK M, MAERTENS J, WAEGEMAN W, DE MEY M. Predictive design of sigma factor-specific promoters[J]. Nature Communications, 2020, 11: 5822.
[62] SAHU B, HARTONEN T, PIHLAJAMAA P, WEI B, DAVE K, ZHU FJ, KAASINEN E, LIDSCHREIBER K, LIDSCHREIBER M, DAUB CO, CRAMER P, KIVIOJA T, TAIPALE J. Sequence determinants of human gene regulatory elements[J]. Nature Genetics, 2022, 54(3): 283-294.
[63] JORES T, TONNIES J, WRIGHTSMAN T, BUCKLER ES, CUPERUS JT, FIELDS S, QUEITSCH C. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters[J]. Nature Plants, 2021, 7(6): 842-855.
[64] WANG HC, DU QX, WANG Y, XU HW, WEI Z, WANG XW. GPro: generative AI-empowered toolkit for promoter design[J]. Bioinformatics, 2024, 40(3): btae123.
[65] GHANDI M, MOHAMMAD-NOORI M, GHAREGHANI N, LEE D, GARRAWAY L, BEER MA. gkmSVM: an R package for gapped-kmer SVM[J]. Bioinformatics, 2016, 32(14): 2205-2207.
[66] LIU B, FANG LY, LONG R, LAN X, CHOU KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition[J]. Bioinformatics, 2016, 32(3): 362-369.
[67] BUTT AH, ALKHALIFAH T, ALTURISE F, KHAN YD. A machine learning technique for identifying DNA enhancer regions utilizing CIS-regulatory element patterns[J]. Scientific Reports, 2022, 12(1): 15183.
[68] DONI JAYAVELU N, JAJODIA A, MISHRA A, HAWKINS RD. Candidate silencer elements for the human and mouse genomes[J]. Nature Communications, 2020, 11: 1061.
[69] ZENG WW, CHEN SQ, CUI XJ, CHEN XY, GAO ZJ, JIANG R. SilencerDB: a comprehensive database of silencers[J]. Nucleic Acids Research, 2021, 49(D1): D221-D228.
[70] HUANG D, OVCHARENKO I. Enhancer-silencer transitions in the human genome[J]. Genome Research, 2022, 32(3): 437-448.
[71] HUANG D, PETRYKOWSKA HM, MILLER BF, ELNITSKI L, OVCHARENKO I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression[J]. Genome Research, 2019, 29(4): 657-667.
[72] ZHANG TJ, LI LY, SUN HL, XU DL, WANG GH. DeepICSH: a complex deep learning framework for identifying cell-specific silencers and their strength from the human genome[J]. Briefings in Bioinformatics, 2023, 24(5): bbad316.
[73] AGARWAL V, KELLEY DR. The genetic and biochemical determinants of mRNA degradation rates in mammals[J]. Genome Biology, 2022, 23(1): 245.
[74] ANGENENT-MARI NM, GARRUSS AS, SOENKSEN LR, CHURCH G, COLLINS JJ. Deep learning for RNA synthetic biology[J]. Biorxiv, 2019: 872077.
[75] ABADI S, YAN WX, AMAR D, MAYROSE I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action[J]. PLoS Computational Biology, 2017, 13(10): e1005807.
[76] KIM HK, MIN S, SONG M, JUNG S, CHOI JW, KIM Y, LEE S, YOON S, KIM HH. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity[J]. Nature Biotechnology, 2018, 36(3): 239-241.
[77] LIN JC, WONG KC. Off-target predictions in CRISPR-Cas9 gene editing using deep learning[J]. Bioinformatics, 2018, 34(17): i656-i663.
[78] KIM GB, GAO Y, PALSSON BO, LEE SY. 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.

[79] BARISSI S, SALA A, WIECZÓR M, BATTISTINI F, OROZCO M. DNAffinity: a machine-learning approach to predict DNA binding affinities of transcription factors[J]. Nucleic Acids Research, 2022, 50(16): 9105-9114.
[80] APGAR J, ROSS M, ZUO X, DOHLE S, STURTEVANT D, SHEN BZ, DELAVEGA H, LESSARD P, LAZAR G, RAAB RM. A predictive model of intein insertion site for use in the engineering of molecular switches[J]. PLoS One, 2012, 7(5): e37355.
[81] JENSEN K, ALPER H, FISCHER C, STEPHANOPOULOS G. Identifying functionally important mutations from phenotypically diverse sequence data[J]. Applied and Environmental Microbiology, 2006, 72(5): 3696-3701.
[82] DE MEY M, MAERTENS J, LEQUEUX GJ, SOETAERT WK, VANDAMME EJ. Construction and model-based analysis of a promoter library for *E. coli*: an indispensable tool for metabolic engineering[J]. BMC Biotechnology, 2007, 7: 34.
[83] LIU DY, MAO ZT, GUO JX, WEI LY, MA HW, TANG YJ, CHEN T, WANG ZW, ZHAO XM. Construction, model-based analysis, and characterization of a promoter library for fine-tuned gene expression in *Bacillus subtilis*[J]. ACS Synthetic Biology, 2018, 7(7): 1785-1797.
[84] WANG Y, WANG HC, WEI L, LI SL, LIU LY, WANG XW. Synthetic promoter design in *Escherichia coli* based on a deep generative network[J]. Nucleic Acids Research, 2020, 48(12): 6403-6412.
[85] KIRYU H, OSHIMA T, ASAI K. Extracting relations between promoter sequences and their strengths from microarray data[J]. Bioinformatics, 2005, 21(7): 1062-1068.
[86] YUAN M, LI H, WANG SZ. Massively parallel reporter assay: a novel technique for analyzing the regulation of gene expression[J]. Hereditas (Beijing), 2023, 45(10): 859-873.
[87] HEINTZMAN ND, REN B. Finding distal regulatory elements in the human genome[J]. Current Opinion in Genetics & Development, 2009, 19(6): 541-549.
[88] BOYLE AP, SONG LY, LEE BK, LONDON D, KEEFE D, BIRNEY E, IYER VR, CRAWFORD GE, FUREY TS. High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells[J]. Genome Research, 2011, 21(3): 456-464.
[89] ERWIN GD, OKSENBERG N, TRUTY RM, KOSTKA D, MURPHY KK, AHITUV N, POLLARD KS, CAPRA JA. Integrating diverse datasets improves developmental enhancer prediction[J]. PLoS Computational Biology, 2014, 10(6): e1003677.
[90] YANG BT, LIU F, REN C, OUYANG ZY, XIE ZW, BO XC, SHU WJ. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone[J]. Bioinformatics, 2017, 33(13): 1930-1936.
[91] NIU K, LUO XM, ZHANG SM, TENG ZX, ZHANG TJ, ZHAO YM. iEnhancer-EBLSTM: identifying enhancers and strengths by ensembles of bidirectional long short-term memory[J]. Frontiers in Genetics, 2021, 12, 665498.
[92] NGUYEN QH, NGUYEN-VO TH, LE NQK, DO TTT, RAHARDJA S, NGUYEN BP. iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks[J]. BMC Genomics, 2019, 20(Suppl 9): 951.
[93] BASITH S, HASAN MM, LEE G, WEI LY, MANAVALAN B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome[J]. Briefings in Bioinformatics, 2021, 22(6): bbab252.
[94] NGAN CY, WONG CH, TJONG H, WANG WB, GOLDFEDER RL, CHOI C, HE H, GONG L, LIN JY, URBAN B, CHOW J, LI MH, LIM J, PHILIP V, MURRAY SA, WANG HY, WEI CL. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development[J]. Nature Genetics, 2020, 52: 264-272.
[95] GISSELBRECHT SS, PALAGI A, KURLAND JV, ROGERS JM, OZADAM H, ZHAN Y, DEKKER J, BULYK ML. Transcriptional silencers in *Drosophila* serve a dual role as transcriptional enhancers in alternate cellular contexts[J]. Molecular Cell, 2020, 77(2): 324-337.
[96] PANG BX, SNYDER MP. Systematic identification of silencers in human cells[J]. Nature Genetics, 2020, 52(3): 254-263.
[97] ZHU Y, SUN L, CHEN Z, WHITAKER JW, WANG T, WANG W. Predicting enhancer transcription and activity from chromatin modifications[J]. Nucleic Acids Research, 2013, 41(22): 10032-10043.
[98] CREYGHTON MP, CHENG AW, WELSTEAD GG, KOOISTRA T, CAREY BW, STEINE EJ, HANNA J, LODATO MA, FRAMPTON GM, SHARP PA, BOYER LA, YOUNG RA, JAENISCH R. Histone H3K27ac separates active from poised enhancers and predicts developmental state[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(50): 21931-21936.
[99] YARRA SS, ASHOK G, MOHAN U. Toehold switches; a foothold for synthetic biology[J]. Biotechnology and Bioengineering, 2023, 120(4): 932-952.
[100] PARDEE K, GREEN AA, TAKAHASHI MK, BRAFF D, LAMBERT G, LEE JW, FERRANTE T, MA D, DONGHIA N, FAN M, DARINGER NM, BOSCH I, DUDLEY DM, O'CONNOR DH, GEHRKE L, COLLINS JJ. Rapid, low-cost detection of Zika virus using programmable biomolecular components[J]. Cell, 2016, 165(5): 1255-1266.
[101] GREEN AA, KIM J, MA D, SILVER PA, COLLINS JJ, YIN P. Complex cellular logic computation using ribocomputing devices[J]. Nature, 2017, 548(7665): 117-121.
[102] ZADEH JN, WOLFE BR, PIERCE NA. Nucleic acid sequence design *via* efficient ensemble defect optimization[J]. Journal of Computational Chemistry, 2011, 32(3): 439-452.
[103] TO ACY, CHU DHT, WANG AR, LI FCY, CHIU AWO, GAO DY, CHOI CHJ, KONG SK, CHAN TF, CHAN KM, YIP KY. A comprehensive web tool for toehold

switch design[J]. Bioinformatics, 2018, 34(16): 2862-2864.
[104] GREEN AA, SILVER PA, COLLINS JJ, YIN P. Toehold switches: *de-novo*-designed regulators of gene expression[J]. Cell, 2014, 159(4): 925-939.
[105] TAKAHASHI MK, TAN X, DY AJ, BRAFF D, AKANA RT, FURUTA Y, DONGHIA N, ANANTHAKRISHNAN A, COLLINS JJ. A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers[J]. Nature Communications, 2018, 9: 3347.
[106] WILSON D, CHAROENSAWAN V, KUMMERFELD SK, TEICHMANN SA. DBD: taxonomically broad transcription factor predictions: new content and functionality[J]. Nucleic Acids Research, 2008, 36(suppl_1): D88-D92.
[107] LAMBERT SA, JOLMA A, CAMPITELLI LF, DAS PK, YIN YM, ALBU M, CHEN XT, TAIPALE J, HUGHES TR, WEIRAUCH MT. The human transcription factors[J]. Cell, 2018, 175(2): 598-599.
[108] HERTZ GZ, STORMO GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences[J]. Bioinformatics, 1999, 15(7): 563-577.
[109] SCHNEIDER TD, STEPHENS RM. Sequence logos: a new way to display consensus sequences[J]. Nucleic Acids Research, 1990, 18(20): 6097-6100.
[110] LUSCOMBE NM, LASKOWSKI RA, THORNTON JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level[J]. Nucleic Acids Research, 2001, 29(13): 2860-2874.
[111] ROHS R, WEST SM, SOSINSKY A, LIU P, MANN RS, HONIG B. The role of DNA shape in protein–DNA recognition[J]. Nature, 2009, 461(7268): 1248-1253.
[112] ROHS R, JIN XS, WEST SM, JOSHI R, HONIG B, MANN RS. Origins of specificity in protein-DNA recognition[J]. Annual Review of Biochemistry, 2010, 79(1): 233-269.
[113] ZHANG QH, SHEN Z, HUANG DS. Predicting *in-vitro* transcription factor binding sites using DNA sequence + shape[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(2): 667-676.
[114] PARK S, KOH Y, JEON H, KIM H, YEO Y, KANG J. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism[J]. Scientific Reports, 2020, 10(1): 13413.
[115] FU LY, ZHANG LH, DOLLINGER E, PENG QK, NIE Q, XIE XH. Predicting transcription factor binding in single cells through deep learning[J]. Science Advances, 2020, 6(51): eaba9031.
[116] CHEN C, HOU J, SHI XW, YANG H, BIRCHLER JA, CHENG JL. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks[J]. BMC Bioinformatics, 2021, 22: 38.
[117] DI VENTURA B, MOOTZ HD. Switchable inteins for conditional protein splicing[J]. Biological Chemistry, 2019, 400(4): 467-475.
[118] PINTO F, THORNTON EL, WANG BJ. An expanded library of orthogonal split inteins enables modular multi-peptide assemblies[J]. Nature Communications, 2020, 11: 1529.
[119] WALDHAUER MC, SCHMITZ SN, AHLMANN-ELTZE C, GLEIXNER JG, SCHMELAS CC, HUHN AG, BUNNE C, BÜSCHER M, HORN M, KLUGHAMMER N, KREFT J, SCHÄFER E, BAYER PA, KRÄMER SG, NEUGEBAUER J, WEHLER P, MAYER MP, EILS R, Di VENTURA B. Backbone circularization of *Bacillus subtilis* family 11 xylanase increases its thermostability and its resistance against aggregation[J]. Molecular BioSystems, 2015, 11(12): 3231-3243.
[120] PAN D, XUAN BQ, SUN YM, HUANG SW, XIE MR, BAI YD, XU WJ, QIAN ZK. An intein-mediated modulation of protein stability system and its application to study human cytomegalovirus essential gene function[J]. Scientific Reports, 2016, 6: 26167.
[121] PURDE V, KUDRYASHOVA E, HEISLER DB, SHAKYA R, KUDRYASHOV DS. Intein-mediated cytoplasmic reconstitution of a split toxin enables selective cell ablation in mixed populations and tumor xenografts[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(36): 22090-22100.
[122] CHOI H, EOM S, KIM HU, BAE Y, JUNG HS, KANG S. Load and display: engineering encapsulin as a modular nanoplatform for protein-cargo encapsulation and protein-ligand decoration using split intein and SpyTag/SpyCatcher[J]. Biomacromolecules, 2021, 22(7): 3028-3039.
[123] PERLER FB. InBase: the intein database[J]. Nucleic Acids Research, 2002, 30(1): 383-384.
[124] STEVENS AJ, BROWN ZZ, SHAH NH, SEKAR G, COWBURN D, MUIR TW. Design of a split intein with exceptional protein splicing activity[J]. Journal of the American Chemical Society, 2016, 138(7): 2162-2165.
[125] HO TYH, SHAO A, LU ZY, SAVILAHTI H, MENOLASCINA F, WANG L, DALCHAU N, WANG BJ. A systematic approach to inserting split inteins for Boolean logic gate engineering and basal activity reduction[J]. Nature Communications, 2021, 12: 2200.
[126] MOOTZ HD, MUIR TW. Protein splicing triggered by a small molecule[J]. Journal of the American Chemical Society, 2002, 124(31): 9044-9045.
[127] OTOMO T, ITO N, KYOGOKU Y, YAMAZAKI T. NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation[J]. Biochemistry, 1999, 38(49): 16040-16044.
[128] LEE YT, SU TH, LO WC, LYU PC, SUE SC. Circular permutation prediction reveals a viable backbone disconnection for split proteins: an approach in identifying a new functional split intein[J]. PLoS One, 2012, 7(8): e43820.
[129] DAGLIYAN O, KROKHOTIN A, OZKAN-DAGLIYAN I, DEITERS A, DER CJ, HAHN

KM, DOKHOLYAN NV. Computational design of chemogenetic and optogenetic split proteins[J]. Nature Communications, 2018, 9: 4042.
[130] SALIS HM, MIRSKY EA, VOIGT CA. Automated design of synthetic ribosome binding sites to control protein expression[J]. Nature Biotechnology, 2009, 27(10): 946-950.
[131] BONIECKI MJ, LACH G, DAWSON WK, TOMALA K, LUKASZ P, SOLTYSINSKI T, ROTHER KM, BUJNICKI JM. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction[J]. Nucleic Acids Research, 2016, 44(7): e63.
[132] HOSSACK EJ, HARDY FJ, GREEN AP. Building enzymes through design and evolution[J]. ACS Catalysis, 2023, 13(19): 12436-12444.
[133] NI B, KAPLAN DL, BUEHLER MJ. ForceGen: End-to-end *de novo* protein generation based on nonlinear mechanical unfolding responses using a language diffusion model[J]. Science Advances, 2024, 10(6): eadl4000.
[134] NOTIN P, ROLLINS N, GAL Y, SANDER C, MARKS D. Machine learning for functional protein design[J]. Nature Biotechnology, 2024, 42(2): 216-228.
[135] GAO YL, WANG L, WANG BJ. Customizing cellular signal processing by synthetic multi-level regulatory circuits[J]. Nature Communications, 2023, 14, 8415.
[136] 高歌, 边旗, 王宝俊. 合成基因线路的工程化设计研究进展与展望[J]. 合成生物学, 2025. https://doi.org/10.12211/2096-8280.2023-096.
GAO G, BIAN Q, WANG BJ. Synthetic genetic circuit engineering: principles, advances and prospects[J]. Synthetic Biology Journal, 2025. https://doi.org/10.12211/2096-8280.2023-096 (in Chinese).
[137] QIN CR, XIANG YH, LIU J, ZHANG RL, LIU ZM, LI TT, SUN Z, OUYANG XY, ZONG YQ, ZHANG HM, OUYANG Q, QIAN L, LOU CB. Precise programming of multigene expression stoichiometry in mammalian cells by a modular and programmable transcriptional system[J]. Nature Communications, 2023, 14: 1500.
[138] WAN XY, PINTO F, YU LY, WANG BJ. Synthetic protein-binding DNA sponge as a tool to tune gene expression and mitigate protein toxicity[J]. Nature Communications, 2020, 11: 5961.
[139] WAN XY, VOLPETTI F, PETROVA E, FRENCH C, MAERKL SJ, WANG BJ. Cascaded amplifying circuits enable ultrasensitive cellular sensors for toxic metals[J]. Nature Chemical Biology, 2019, 15(5): 540-548.
[140] HARTLINE CJ, SCHMITZ AC, HAN YC, ZHANG FZ. Dynamic control in metabolic engineering: theories, tools, and applications[J]. Metabolic Engineering, 2021, 63: 126-140.
[141] 项延会, 李婷婷, 娄春波. 人工基因回路设计原理的进展与挑战[J]. 生命科学, 2021, 33(12): 1445-1451.
XIANG YH, LI TT, LOU CB. Progresses and challenges for the fundamental design principles of genetic circuits[J]. Chinese Bulletin of Life Sciences, 2021, 33(12): 1445-1451 (in Chinese).
[142] ZHU JJ, ZHANG Q, FOROURAGHI B, WANG X. Applications of machine learning techniques in genetic circuit design[C]. 2021 13th International Conference on Machine Learning and Computing. Shenzhen China, 2021.
[143] DANIELS KG, WANG SY, SIMIC MS, BHARGAVA HK, CAPPONI S, TONAI Y, YU W, BIANCO S, LIM WA. Decoding CAR T cell phenotype using combinatorial signaling motif libraries and machine learning[J]. Science, 2022, 378(6625): 1194-1200.
[144] QIN CR, XU T, ZHAO XJ, ZONG YQ, ZHANG HM, LOU CB, OUYANG Q, QIAN L. Functional predictability of universal gene circuits in diverse microbial hosts[J]. Quantitative Biology, 2024, 12(2): 129-140.
[145] RAI K, O'CONNELL R W, MEHTA P, PATEL A, BASHOR C J. CLASSIC: A platform for high throughput mapping of genetic design spaces in mammalian cells and ML guided prediction of gene circuit behavior[C]. ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design, 2024.
[146] CZAR MJ, CAI Y, PECCOUD J. Writing DNA with GenoCAD[J]. Nucleic Acids Research, 2009, 37(suppl_2): W40-W47.
[147] MYERS CJ, BARKER N, JONES K, KUWAHARA H, MADSEN C, NGUYEN NP D. iBioSim: a tool for the analysis and design of genetic circuits[J]. Bioinformatics, 2009, 25(21): 2848-2849.
[148] NIELSEN AAK, DER BS, SHIN J, VAIDYANATHAN P, PARALANOV V, STRYCHALSKI EA, ROSS D, DENSMORE D, VOIGT CA. Genetic circuit design automation[J]. Science, 2016, 352(6281): aac7341.
[149] FRANK SA. Optimization of transcription factor genetic circuits[J]. Biology, 2022, 11(9): 1294.
[150] SHEN JX, LIU F, TU YH, TANG C. Finding gene network topologies for given biological function with recurrent neural network[J]. Nature Communications, 2021, 12, 3125.
[151] MERZBACHER C, MAC AODHA O, OYARZÚN DA. Bayesian optimization for design of multiscale biological circuits[J]. ACS Synthetic Biology, 2023, 12(7): 2073-2082.
[152] HISCOCK TW. Adapting machine-learning algorithms to design gene circuits[J]. BMC Bioinformatics, 2019, 20: 214.
[153] GHEISARI M, EBRAHIMZADEH F, RAHIMI M, MOAZZAMIGODARZI M, LIU Y, DUTTA PRAMANIK PK, ALI HERAVI M, MEHBODNIYA A, GHADERZADEH M, FEYLIZADEH MR, KOSARI S. Deep learning: applications, architectures, models, tools, and frameworks: a comprehensive survey[J]. CAAI Transactions on Intelligence Technology, 2023, 8(3): 581-606.
[154] WANG L, HAN M, LI XJ, ZHANG N, CHENG HD. Review of classification methods on unbalanced data sets[J]. IEEE Access, 2021, 9: 64606-64628.
[155] HÖLLERER S, PAPAXANTHOS L, GUMPINGER AC, FISCHER K, BEISEL C, BORGWARDT K, BENENSON Y, JESCHEK M. Large-scale DNA-based phenotypic recording and deep learning enable highly

accurate sequence-function mapping[J]. Nature Communications, 2020, 11: 3551.
[156] MATZKO R, KONUR S. Technologies for design-build-test-learn automation and computational modelling across the synthetic biology workflow: a review[J]. Network Modeling Analysis in Health Informatics and Bioinformatics, 2024, 13(1): 22.
[157] STORCH M, HAINES MC, BALDWIN GS. DNA-BOT: a low-cost, automated DNA assembly platform for synthetic biology[J]. Synthetic Biology, 2020, 5(1): ysaa010.
[158] BAKER M. 1, 500 scientists lift the lid on reproducibility[J]. Nature, 2016, 533(7604): 452-454.
[159] CHEN Y, GUENTHER JM, GIN JW, CHAN LJG, COSTELLO Z, OGORZALEK TL, TRAN HM, BLAKE-HEDGES JM, KEASLING JD, ADAMS PD, GARCÍA MARTÍN H, HILLSON NJ, PETZOLD CJ. Automated “cells-to-peptides” sample preparation workflow for high-throughput, quantitative proteomic assays of microbes[J]. Journal of Proteome Research, 2019, 18(10): 3752-3761.
[160] ARANKO AS, WLODAWER A, IWAÏ H. Nature’s recipe for splitting inteins[J]. Protein Engineering, Design & Selection, 2014, 27(8): 263-271.
[161] LOHR S. For big-data scientists, ‘janitor work’ is key hurdle to insights[J]. New York Times, 2014, 17: B4.
[162] BERNETT J, BLUMENTHAL DB, GRIMM DG, HASELBECK F, JOERES R, KALININA OV, LIST M. Guiding questions to avoid data leakage in biological machine learning applications[J]. Nature Methods, 2024, 21(8): 1444-1453.
[163] CHEN V, YANG MY, CUI WB, KIM JS, TALWALKAR A, MA J. Applying interpretable machine learning in computational biology-pitfalls, recommendations and opportunities for new developments[J]. Nature Methods, 2024, 21(8): 1454-1461.
[164] ARCHER KJ, KIMES RV. Empirical characterization of random forest variable importance measures[J]. Computational Statistics & Data Analysis, 2008, 52(4): 2249-2260.
[165] CHOWDHURY KR, SIL A, SHUKLA SR. Explaining a black-box sentiment analysis model with local interpretable model diagnostics explanation (LIME)[M]. Communications in Computer and Information Science. Cham: Springer International Publishing, 2021: 90-101.
[166] RODRÍGUEZ-PÉREZ R, BAJORATH J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions[J]. Journal of Computer-Aided Molecular Design, 2020, 34(10): 1013-1026.
[167] CORRAL-CORRAL R, BELTRÁN JA, BRIZUELA CA, DEL RIO G. Systematic identification of machine-learning models aimed to classify critical residues for protein function from protein structure[J]. Molecules, 2017, 22(10): 1673.
[168] DOERR A. Protein design: the experts speak[J]. Nature Biotechnology, 2024, 42(2): 175-178.
[169] ESPOSITO C, WANG SZ, LANGE UEW, OELLIEN F, RINIKER S. Combining machine learning and molecular dynamics to predict P-glycoprotein substrates[J]. Journal of Chemical Information and Modeling, 2020, 60(10): 4730-4749.
[170] LI QB, ZHENG Y, SU TY, WANG Q, LIANG QF, ZHANG ZD, QI QS, TIAN J. Computational design of a cutinase for plastic biodegradation by mining molecular dynamics simulations trajectories[J]. Computational and Structural Biotechnology Journal, 2022, 20: 459-470.
[171] DREW BENNETT WF, HE S, BILODEAU CL, JONES D, SUN DL, KIM H, ALLEN JE, LIGHTSTONE FC, INGÓLFSSON HI. Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning[J]. Journal of Chemical Information and Modeling, 2020, 60(11): 5375-5381.
[172] LIU B, HUANG HY, LIAO WX, PAN XY, JIN C, YUAN Y. DeepSipred: a deep-learning-based approach on siRNA inhibition prediction[C]. 2024 4th International Conference on Bioinformatics and Intelligent Computing. Beijing China, 2024.
[173] PAUL S, OLYMON K, MARTINEZ GS, SARKAR S, YELLA VR, KUMAR A. MLDSPP: bacterial promoter prediction tool using DNA structural properties with machine learning and explainable AI[J]. Journal of Chemical Information and Modeling, 2024, 64(7): 2705-2719.