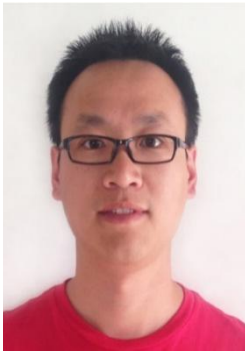
·生物元器件智能设计合成·

**张学礼** 中国科学院天津工业生物技术研究所研究员。长期致力于生物制造和基因编辑的研究工作，创建微生物细胞工厂生产生物基化学品，开发基因编辑技术用于遗传疾病的基因治疗。研究成果发表在 *Nature Biotechnology* 和 *Nature Communications* 等学术期刊。获国家自然科学基金杰出青年和优秀青年基金资助，入选国家高层次人才特殊支持计划。



**孙喆** 中国科学院天津工业生物技术研究所研究员。通过设计和利用高通量测序方法，同时结合代谢工程和高通量筛选等手段，开发基因转录和翻译元件、调控因子和功能酶元件的高通量挖掘和改造技术。研究成果发表在 *Nature Communications* 和 *Nucleic Acids Research* 等学术期刊。

# 生物制造中核酸元件的智能设计

王金盛[1]，孙喆[1,2*]，张学礼[1,2*]

1 中国科学院天津工业生物技术研究所，天津 300308
2 国家合成生物技术创新中心，天津 300308

王金盛，孙喆，张学礼. 生物制造中核酸元件的智能设计[J]. 生物工程学报, 2025, 41(3): 968-992.
WANG Jinsheng, SUN Zhe, ZHANG Xueli. Intelligent design of nucleic acid elements in biomanufacturing[J]. Chinese Journal of Biotechnology, 2025, 41(3): 968-992.

**摘 要：**核酸元件是重要的功能性核酸序列，在生物制造中通过基因表达调控、代谢途径优化和基因编辑等方面来影响目标产物的合成，因此核酸元件的设计和优化对细胞工厂的构建有重要作用。通过人工智能技术可以准确有效地预测功能性核酸元件，设计和优化功能稳定的元件序列，同时解析其作用机制，为生物制造提供强大的技术支持。近年来，人工智能技术在生物制造中通过设计启动子、核糖体结合位点和终止子等核酸元件及其组合，可以大幅

资助项目：国家重点研发计划(2022YFC2106200)
This work was supported by the National Key Research and Development Program of China (2022YFC2106200).
*Corresponding authors. E-mail: SUN Zhe, sunzhe@tib.cas.cn; ZHANG Xueli, zhang_xl@tib.cas.cn
Received: 2024-07-23; Accepted: 2025-02-06; Published online: 2025-02-07

度减少实验工作量，加快生物制造进程。但是由于生物系统的复杂性和高质量训练数据不足等问题，导致核酸元件的智能设计在生物制造中的应用较为单一。本文综述了应用于生物制造的各种 DNA 和 RNA 核酸元件，基于人工智能算法构建的核酸元件预测和设计工具及人工智能技术在生物制造中的应用案例。未来通过整合人工智能技术、合成生物学和高通量技术等，有望开发更高效准确的核酸元件设计方法，加速其在生物制造中的应用。
**关键词：**核酸元件；智能设计；机器学习；合成生物学；代谢工程

# Intelligent design of nucleic acid elements in biomanufacturing

**WANG Jinsheng[1], SUN Zhe[1,2*], ZHANG Xueli[1,2*]**

1 Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
2 National Center of Technology Innovation for Synthetic Biology, Tianjin 300308, China

**Abstract:** Nucleic acid elements are essential functional sequences that play critical roles in regulating gene expression, optimizing pathways, and enabling gene editing to enhance the production of target products in biomanufacturing. Therefore, the design and optimization of these elements are crucial in constructing efficient cell factories. Artificial intelligence (AI) provides robust support for biomanufacturing by accurately predicting functional nucleic acid elements, designing and optimizing sequences with quantified functions, and elucidating the operating mechanisms of these elements. In recent years, AI has significantly accelerated the progress in biomanufacturing by reducing experimental workloads through the design and optimization of promoters, ribosome-binding sites, terminators, and their combinations. Despite these advancements, the application of AI in biomanufacturing remains limited due to the complexity of biological systems and the lack of highly quantified training data. This review summarizes the various nucleic acid elements utilized in biomanufacturing, the tools developed for predicting and designing these elements based on AI algorithms, and the case studies showcasing the applications of AI in biomanufacturing. By integrating AI with synthetic biology and high-throughput techniques, we anticipate the development of more efficient tools for designing nucleic acid elements and accelerating the application of AI in biomanufacturing.
**Keywords:** nucleic acid elements; intelligent design; machine learning; synthetic biology; metabolic engineering

核酸元件是在各种生物过程中发挥重要作用的功能性核酸(DNA 或 RNA)序列，是生物体生命活动的基础，也是生物制造高效生产的关键。DNA 核酸元件包括多种功能元件和调节元件，如启动子、核糖体结合位点和调控因子结合位点等，可以直接控制或间接调节基因的转录和翻译强度，生物制造中经常通过调节启动子和核糖体结合位点强度来增加产物的产量。DNA 转录出的 RNA 核酸元件根据其结构和功能可分为多种，如位于 RNA 的 5′端非编码区的核糖开关，可以通过调节 RNA 的翻译效率、稳定性等来控制代谢流。除此之外，一些非编码

RNA 可以起到调节基因表达的作用，如向导 RNA (small guide RNA, sgRNA)、微小 RNA (microRNA, miRNA)、小干扰 RNA (small interfering RNA, siRNA)、环状 RNA (circular RNA, circRNA)[1]等，也可以归类为 RNA 核酸元件并在生物制造中发挥重要作用。

利用人工智能技术，可以快速和准确地设计核酸元件，并增加核酸元件功能的多样性，推动生物制造领域的发展。基于机器学习的优化算法，如遗传算法和粒子群算法，适用于核酸元件分类和代谢途径设计等问题[2-3]。卷积神经网络和生成对抗网络等深度学习算法在处理大规模、高维度和非线性数据方面有优势，可用于核酸元件的作用机制解析和序列生成等任务[4-5]。本文围绕生物制造中常用的核酸元件，人工智能对核酸元件的预测、设计及其应用进行综述，以期为生物制造中利用人工智能设计核酸元件的研究人员提供参考。

# 1 生物制造中使用的核酸元件

核酸元件广泛应用于生物制造，各种 DNA 元件(启动子、增强子、转录因子结合位点、核糖体结合位点/Kozak 序列、终止子等)和 RNA 元件(sgRNA、miRNA、siRNA、核糖开关、circRNA)从转录和翻译等不同层面调节生物细胞工厂的功能。下面将分别介绍生物制造中常用核酸元件的结构与功能。

## 1.1 DNA 元件

生物体依次通过转录和翻译过程将基因中的遗传信息转化为蛋白质以发挥其功能[6-7]。位于基因 5′端的启动子是一段与 RNA 聚合酶结合以起始基因转录的 DNA 序列，决定了目标基因的转录强度(图 1A)。在原核生物中，RNA 聚合酶通过 σ 因子特异性识别启动子序列，其中大肠杆菌的管家 σ 因子为 $\sigma^{70}$，$\sigma^{70}$ 启动子包括转录起始位点、–10 区和–35 区及其他调节元件[8-10]。真核启动子结构更复杂，包含转录起始位点、TATA 框、起始子元件、下游启动子元件和上游调控元件等[11]。原核生物中的核糖体结合位点(ribosome binding site, RBS)和真核生物中的 Kozak 序列(Kozak consensus sequence)在基因的翻译起始中有重要作用，可以从翻译水平上影响基因的表达(图 1B)。核糖体结合位点(多为 A 或 G)通过碱基互补配对与核糖体 16S rRNA 的 3′端结合，影响蛋白翻译强度[12]。Kozak 序列(保守序列为 GCCRCCAUGG)通过促进核糖体结合到信使核糖核酸(messenger RNA, mRNA)并准确识别起始密码子来提高翻译起始效率[13]。

启动子和核糖体结合元件可以直接调节基因的转录和翻译量，此外还有多种核酸元件可以调节基因的表达。如真核生物中距离基因转录起始位点 2 kb 甚至更远的增强子，可以通过与转录复合体的结合来增强基因的转录水平，使基因的转录频率增加 10–200 倍(图 1C)[14-15]。距离启动子较近的转录因子结合位点可以通过与特定的转录因子结合，调节基因在特定时间和空间的表达模式(图 1D)[16]。位于 3′非翻译区的终止子可以形成发卡结构阻止 RNA 聚合酶在 DNA 模板上的移动，从而促进转录终止(图 1E)[17-18]。

## 1.2 RNA 元件

相较于 DNA 元件，多种 RNA 元件在基因转录后调节其表达水平。如某些启动子区域存在的包含适体区和基因表达区的核糖开关，可以与被感应的配体结合来改变起始密码子附近的茎环结构以控制翻译的启动和暂停[19-22](图 1F)。多种非编码 RNA，如小干扰 RNA 和微小 RNA 也可以调控基因的表达。siRNA 是一段长 19–25 nt 的外源双链 RNA 分子，通过完全互

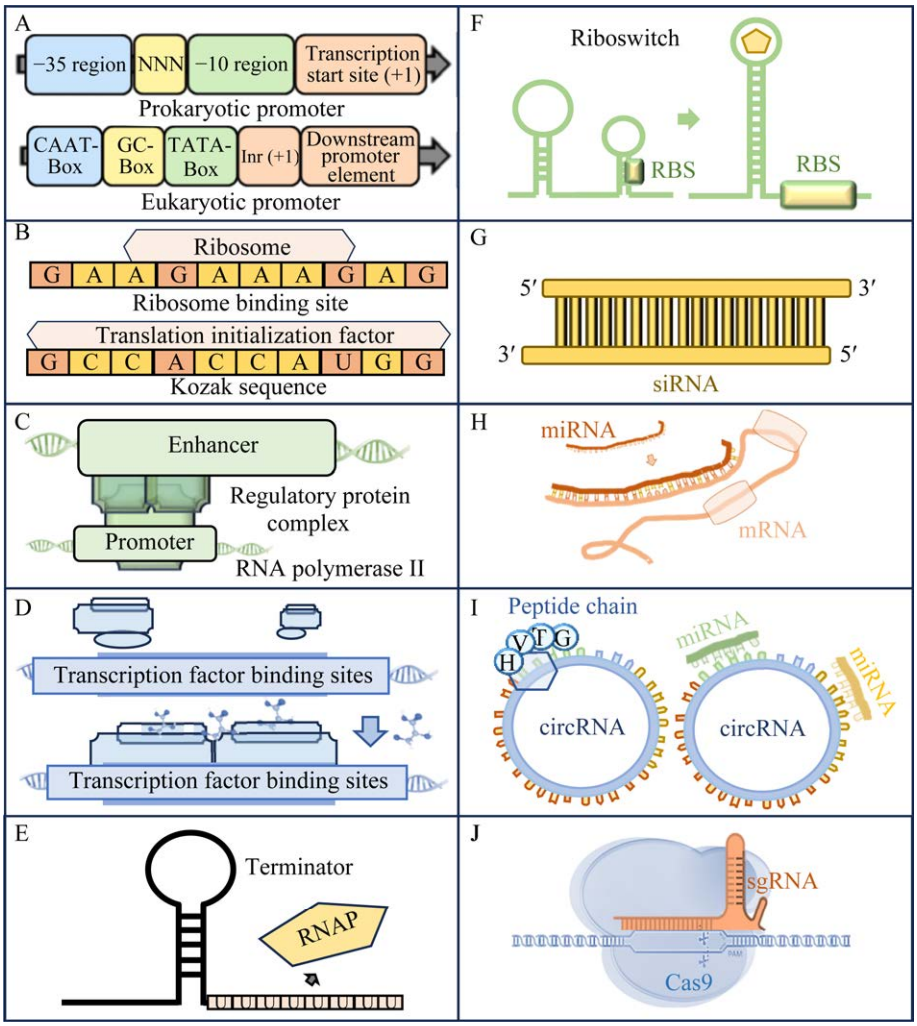



**图 1　生物制造中使用的核酸元件示意图**　A：启动子调节基因的转录。原核启动子元件主要包括 –10 区、–35 区、转录起始位点。真核启动子元件主要包括 CAAT 盒、GC 盒、TATA 盒、起始子元件(Inr)、下游启动子元件。B：核糖体结合位点/Kozak 序列调节基因的翻译起始。C：增强子远距离增强基因的转录。D：转录因子能够感知特定的配体，通过改变与转录结合位点的亲和力来调节基因的表达。E：终止子促进 RNA 聚合酶从模板 DNA 上脱离以实现终止基因的转录。F：核糖开关感应配体，调节基因的表达。G：siRNA 促进 mRNA 降解，沉默靶基因。H：miRNA 促进 mRNA 降解，沉默靶基因。I：环状 RNA 翻译蛋白或与 miRNA 结合调节基因表达。J：sgRNA (红色)与 Cas 蛋白结合，可以实现对基因的精准编辑和调控。

Figure 1　Diagram of nucleic acid elements used in biomanufacturing. A: The promoter regulates gene transcription. The main elements of prokaryotic *Escherichia coli* $\sigma^{70}$ promoters include the –10 region, the –35 region, and the transcription start site. The main elements of eukaryotic promoters include the CAAT box, the GC box, the TATA box, the initiator element (Inr), and the downstream promoter element. B: The ribosome binding site (RBS)/Kozak sequence regulates gene translation initiation. C: Enhancers can remotely enhance the transcription of genes. D: Transcription factors can sense specific ligands and regulate gene expression by altering their affinity for transcription-binding sites. E: Terminators facilitate the dissociation of RNA polymerase (RNAP) from the template DNA to terminate gene transcription. F: Riboswitch can sense specific ligands and regulate gene expression. G: siRNA promotes mRNA degradation and silences target genes. H: miRNA promotes mRNA degradation and silences target. I: circRNA can translate proteins or bind miRNAs to regulate gene expression. J: sgRNA (in red) binds to the Cas protein, enabling precise gene editing and regulation.

补的方式与靶 mRNA 结合，通过促进 mRNA 降解以抑制基因表达(图 1G)[23]。miRNA 是 21–24 nt 的内源性单链 RNA 分子，由内源基因通过转录和加工生成，可以通过部分配对结合的方式与 mRNA 结合，使 mRNA 降解以调控目标基因的表达(图 1H)。其他种类的 RNA，如环状 RNA (circular RNA, circRNA)和信使核糖核酸相互竞争 miRNA，实现翻译层次的复杂调节(图 1I)[24-26]。除此之外，常用于生物制造的核酸元件还有 CRISPR/Cas 中的 sgRNA，通过将 Cas 蛋白定位至目标 DNA，对基因进行精准编辑和调控(图 1J)[27-29]。

天然核酸元件的类型和功能有限，存在数量较少、诱导或表达强度较低以及强度分布不均匀等问题。此外，天然核酸元件还受到宿主复杂调控网络的影响，难以在复杂条件下达到相同的表达强度，实验重复性较差。同时，天然核酸元件的功能较为单一，难以满足不同发酵条件下目标基因精确调控的需求，以上缺点限制了其在生物制造中的应用[30-31]。为了实现代谢流精细调节以提高目标产物的产量、产率和纯度的目标，研究人员可以利用人工智能技术，挖掘和设计功能更丰富、机制更明确的核酸元件[32-36]。下文将对人工智能技术在各种核酸元件的预测和设计方面的应用进行分类总结。

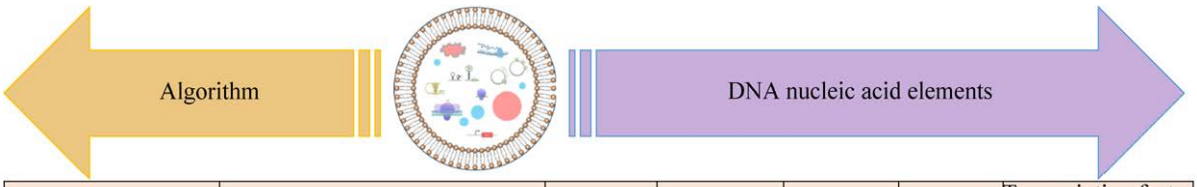
## 2 DNA 元件的智能预测与设计

在现代生物制造中，细胞工厂的构建必然要用到启动子、核糖体结合位点/Kozak 序列和终止子等 DNA 元件，核酸元件的合理设计和优化是提高产品的产量和转化率的关键。已有较多研究利用人工智能技术来挖掘、预测和设计 DNA 元件(图 2 和表 1)，在生物制造中逐渐显露出应用价值。

### 2.1 启动子

启动子的结构具有保守性，根据启动子的分布、核苷酸偏好性和 GC 含量等序列特征，可以从基因组上快速地识别启动子并挖掘新的启动子元件。目前，已经开发了多种工具用于挖掘启动子，如 Sigma70pred (适用于大肠杆菌的 sigma70 启动子)[52]、G4PromFinder (适用于 GC 含量丰富的细菌，如天蓝色链霉菌和铜绿假单胞菌)[53]和 CNNProm (适用于人、鼠、拟南芥、大肠杆菌和枯草芽孢杆菌)[39]等。通过对 iPro70-FMWin[54]、70ProPred[37]和 iPromoter-2L[55-56]等针对细菌启动子表现较好的预测工具进行分析，发现将 DNA 序列与其物理化学特性相结合的特征提取方式可有效地提高预测的敏感性和准确率[57]。此外，研究人员也比较了不同算法对细菌启动子预测的影响，相比于适合小样本数据的支持向量机和需要整合多个决策树的随机森林算法，基于梯度提升框架的极端梯度提升算法(XGBoost)在结合树模型优势的同时进行了多种优化，使训练速度快且性能更高。其中表现较优的基于机器学习和双链稳定性的启动子预测工具(machine learning and duplex stability based promoter prediction in prokaryotes, MLDSPP)可以利用原始下游序列作为对照来预测细菌的启动子，在提高预测准确性的同时，增强了预测结果的可解释性[38]。

除了预测出的天然启动子，生物制造还需要使用具有一定表达强度的启动子进行基因表达。传统实验方法一般通过突变或不同元件组合来获取新的启动子，例如，Alper 等[58]利用易错 PCR 和荧光筛选的方法来构建噬菌体 $P_L$-λ 启动子的随机突变体文库，但缺点是设计成本高且花费时间长。人工智能技术通过学习天然启动子序列可以预测合成启动子的转录强度，能够减少实验工作量，快速获得特定强度的启




| Model classification | Artificial intelligence algorithms | Promoter | RBS/Kozak | Terminator | Enhancer | Transcription factor binding sites |
|---|---|---|---|---|---|---|
| Traditional machine learning models | Linear regression model | | ✓ | | | |
| | Support vector machine | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Random forest | ✓ | | | ✓ | ✓ |
| | XGBoost | ✓ | | ✓ | | ✓ |
| | Clustering algorithm | | | | | ✓ |
| | Bayes algorithm | | | | ✓ | ✓ |
| | Gaussian process regression | | ✓ | | | |
| | Ensemble learning | | | | ✓ | ✓ |
| | Markov chain | | | | ✓ | ✓ |
| Deep learning models | Artificial neural network | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Convolutional neural network | ✓ | ✓ | | ✓ | ✓ |
| | Recurrent neural networks | | | | ✓ | ✓ |
| | Transformer | | | | | ✓ |
| Multi-Task learning | Task adaptive parameter sharing | | | | ✓ | |
| Generative model | Generative adversarial network | ✓ | | | | |

**图 2　DNA 核酸元件预测和设计中使用的人工智能算法**

Figure 2　Summary of artificial intelligence algorithms applied in the prediction and design of DNA nucleic acid elements.

动子。Zhao 等[59]以 Trc 启动子为研究对象，利用 3 665 个突变体构建和优化了不同的机器学习模型，其中基于 XGBoost 的模型性能最佳，可以预测启动子序列和转录强度之间的关系。

基于 XGBoost 算法的判定模型预测的启动子序列往往与天然启动子类似，而生成模型则可以产生更新颖的启动子序列。Wang 等[40]将深度生成模型与预测模型相结合，设计了针对大肠杆菌的 Deep_promoter，通过学习天然启动子的序列特征并捕获核苷酸之间的相互作用生成新的启动子；经过两轮优化及实验测定后发现生成的启动子中 70.8%具有功能，生成的启动子与天然启动子的 K-mer 频率、−10 区和−35 区序列有较大差距。将特定条件与深度生成模型结合，不仅可以生成具有生物功能的启动子，还能赋予其相关特性以使生产的启动子更适用于生物制造。例如条件生成对抗网络模型 DeepSEED 可以根据输入的种子序列(如转录因子结合位点)生成含种子序列的完整启动子。该模型可以应用于大肠杆菌组成型启动子和 IPTG 诱导型启动子及哺乳动物细胞多西环素诱导型启动子的设计，生成的启动子在保留 K-mer 频率和 DNA 构型等关键特征的同时表现出高度的序列多样性[41]。

## 2.2　核糖体结合位点/Kozak 序列

除了改变基因的转录水平，生物制造也需要从翻译水平调节基因表达。人工智能技术可以鉴定原核和真核生物的翻译调控元件，同时从序列结构信息中预测翻译强度[60-61]。如 Salis

**表 1 DNA 核酸元件预测和设计中使用的代表性人工智能算法简介**

Table 1 Introduction to representative artificial intelligence algorithms used in DNA nucleotide component prediction and design

| Nucleic acid elements | Representative tools | Main algorithm or architecture of the model | Applicable species | Function | References |
|---|---|---|---|---|---|
| Promoter | 70ProPred | Support vector machine | Sigma-70 promoters in prokaryotes | Identify sigma70 promoter regions in DNA sequences based on sequence features (single-stranded characteristic and electron-ion potential values) | [37] |
| Promoter | MLDSPP | XGBoost | Bacteria | Using the original downstream sequence as a control to predict promoter regions in bacteria helps to understand gene regulation in different bacterial systems | [38] |
| Promoter | CNNProm | Convolutional neural networks | Human, mouse, plant (*Arabidopsis*), and two bacteria (*Escherichia coli* and *Bacillus subtilis*) | Classifying eukaryotic and prokaryotic promoter and non-promoter sequences | [39] |
| Promoter | Deep_promoter | Deep generative model | *Escherichia coli* | To study the sequence characteristics of natural promoters and capture the interactions between nucleotides to generate new promoters | [40] |
| Promoter | DeepSEED | Generative adversarial networks | *Escherichia coli*, mammalian cells | Generate a complete promoter containing the seed sequence (such as a transcription factor binding site) based on the input seed sequence | [41] |
| RBS | RBS calculator | Simulated annealing algorithm | *Escherichia coli* | Predict the translation initiation rate by quantifying the interaction strength between the 30S ribosomal subunit complex and the mRNA molecule | [42] |
| RBS | TITER | Deep convolutional and recurrent neural networks, linear regression models | HEK293 cell line | Quantitative translation initiation sequencing (QTI-seq) data will predict translation start sites across the whole genome and identify Kozak sequences | [43] |
| RBS | SAPIEN | Residual neural networks | *Escherichia coli* | Quantitative and predictive relationships between RBS sequences and translation strength | [44] |
| Terminator | Support vector machine model | Support vector machine, random context-free grammar, Cocke-younger-kasaami | *Escherichia coli* | Inferring whether the sequence forms a hairpin structure, categorizing rho-independent transcription terminators in the *E. coli* genome | [45] |
| Terminator | XGBoost model | XGBoost | *Escherichia coli* | Classifying terminators of *Escherichia coli* | [46] |
| Enhancer | DEEP | Support vector machine, artificial neural network | 36 different types of human cell tissues | DNA regions are classified as enhancers or non-enhancers | [47] |
| Enhancer | McEnhancer | Interpolation Markov chain, logistic regression classifier | Fruit flies (CRM2893, CRM6053, CRM5481, etc.) | Predict enhancers and discover chromosomal interactions and histone modifications associated with enhancers | [48] |
| Transcription factor binding site | BinDNase | Logistic regression | K562 cell line, HepG2 cell line | Predict the transcription factor and DNA interaction sites of cell line K562/HepG2 using DNase-seq data | [49] |
| Transcription factor binding site | Chained machine learning models | Clustering algorithm, ensemble learning (neural networks, decision trees, and naive Bayes classification) | Human | Classify promoter regions based on the location and strength of transcription factor binding sites, and detect target genes of transcription factors | [50] |
| Transcription factor binding site | D-AEDNet | Encoder-Decoder structure, convolutional neural network | Human | Identifying the locations of transcription factor DNA-binding sites in DNA sequences through chromatin immunoprecipitation sequencing(ChIP-seq) | [51] |

等[42]利用模拟退火算法在大肠杆菌中设计了RBS的统计热力学模型RBS Calculator，通过量化30S核糖体亚基复合物与mRNA分子之间的相互作用强度来预测翻译起始速率；实验验证表明该模型在10万倍翻译强度差异范围内具有较高的准确性，误差仅在2.3倍以内。基于随机森林的机器模型Scikit-ribo可以利用Ribo-seq和RNA-seq数据预测真核生物的Kozak序列和翻译效率[62]。而利用深度卷积和循环神经网络混合的深度学习模型TITER可以利用定量翻译起始测序(quantitative translation initiation sequencing, QTI-seq)数据在HEK293细胞系全基因组中预测翻译起始位点，识别Kozak序列[43]。此外，深度学习算法无需翻译过程的先验知识(如分子间相互作用强度)，仅需序列即可准确预测翻译强度。如通过学习大肠杆菌约30万个核糖体结合位点的翻译动力学数据，SAPIEN工具利用残差神经网络定量和预测RBS序列与翻译强度间的关系，其决定系数$R^2$高达0.927，同时平均绝对误差仅为0.039，优于过往模型[44]。细胞工厂的代谢途径加强经常需要搭配强RBS以提高关键基因的表达，因此人工智能需要设计具有更高翻译强度的RBS。Zhang等[63]利用置信区间上界算法和高斯过程回归来设计优势突变体，可以灵活处理不同类型输入数据，在短时间内使RBS翻译强度增加了34%。

通过人工智能可以优化RBS序列，进而调节代谢流以消除生物制造中的代谢瓶颈[64-65]。Farasat等[66]利用遗传算法和线性回归模型来预测RBS的翻译强度，成功将链孢红素生产速度由3.3 μg/(gDCW·h)提升至196 μg/(gDCW·h)。人工智能技术的另一个优势是可以减少实验工作量，加快实验进度。如Jervis等[67]利用支持向量机对RBS序列和表型之间的关系进行建模，仅测试3%的样本就将柠檬烯的产量提高了60%。除了控制RBS的翻译强度来增加产量，还可以将RBS与其他核酸元件结合起来实现更复杂的功能。例如深度学习模型CLM-RDR利用卷积神经网络学习了含有7 053个不同RBS的葡萄糖二酸生物传感器文库，可以预测生物传感器的动态范围并具有较好的泛化能力(如预测乙醇酸/阿拉伯糖生物传感器)[68]。

## 2.3 终止子

细菌中按照转录机制可以分为2类终止子，分别为Rho依赖型和Rho非依赖型终止子[69-72]。Rho依赖型终止子依赖Rho蛋白沿着mRNA滑动并结合至Rut位点来解离转录复合体以实现转录终止，因此可以根据其保守结合序列预测终止子[73]。如RhoTermPredict模型和P&TIT模型可以分别利用Rho依赖型转录终止子富含C和缺乏G的70–80 nt共有基序来识别大肠杆菌、枯草芽孢杆菌和肠道沙门氏菌及产二素链霉菌中的Rho依赖性终止子[74-75]。Rho非依赖型终止子的序列不保守，但是其具有发卡结构并紧邻一段连续的U尾区，因此可以通过算法模拟序列结构来预测终止子。Francis-Lyon等[45]利用动态规划算法推断序列是否形成发卡结构，然后利用支持向量机判断序列是否属于终止子元件，最终获得的大肠杆菌Rho非依赖型终止子分类器的成功率达96.4%。此外，采用不同的编码方法对核苷酸序列进行编码也可以提升终止子预测的准确率，例如使用核苷酸碱基的电子-离子相互作用电位值对核苷酸序列进行编码后训练人工神经网络，可以在减小模型大小的同时使预测准确率超过95%[76]。

除了预测终止子，人工智能也可以设计具有不同终止强度的终止子并解析转录终止机制。如Zhai等[46]设计了基于XGBoost的机器学习模型用于大肠杆菌终止子的分类，分类模型在5倍交叉验证中的平均准确率为0.956，并发

现更长的茎、更紧凑的环和完美的连续 U 尾区更有利于终止子转录终止。

## 2.4 增强子

真核生物中的增强子可调节 DNA 构象并结合转录调控因子，可以有效增加调节基因的转录水平[15,77-78]。增强子的预测根据不同的特征表示方法和算法有多种工具，如 ChromHMM[79]、Segway[80]、CSI-ANN[81]、ChromaGenSVM[82]、RFECS[83]、EnhancerFinder[84]和 MCSE-enhancer[85]等，部分模型的准确度可超过 90%。2015 年，Kleftogiannis 等[47]开发了增强子分类模型 DEEP，该模型首先训练出多个支持向量机模型，然后使用人工神经网络对多个模型进行整合并推导最终决策，能够有效地将 DNA 区域分类为增强子或非增强子，在 36 个不同细胞组织中达到约 90%的准确率，适用于多种组织细胞。2017 年，Hafez 等[48]设计了根据序列特征预测增强子的插值马尔可夫链模型 McEnhancer，其原理是如果一个序列被分类为增强子，并且该序列与训练集中的增强子序列相似，那么它可能会增强相应基因的表达；该模型对果蝇胚胎的表达模式预测准确率为 73%−98%，并发现了增强子相关的染色体相互作用和组蛋白修饰。

除了预测增强子，人工智能技术也可以用于研究增强子的作用机制。例如基于多任务学习技术的可解释性深度学习模型 HEAP，其使用所有细胞类型数据集训练一个基础模型，然后通过添加子集层来适应特定的任务。HEAP 模型可以结合 DNA 序列和表观遗传修饰信息来预测增强子活性，在分析表观遗传修饰重要性的同时识别基序，可评价每个核苷酸在增强子中的作用[86]。

## 2.5 转录因子结合位点

转录因子结合位点可以与转录因子结合以调节基因的表达[16]，其长度通常较短(5–20 bp)，可以通过如 DNA 足迹分析、凝胶阻滞实验和染色质免疫共沉淀等实验来检测结合位点的位置及与转录因子结合的亲和力[87-89]。除此之外，也可以利用深度测序数据预测转录因子结合位点及其调控关系，如逻辑回归模型 BinDNase 利用 DNase-seq 数据预测细胞系 K562 和 HepG2 中多个转录因子与 DNA 的相互作用位点，能推广到多种细胞类型且优于其他方法[49,90-92]。如果没有深度测序数据，人工智能算法也可以根据已有的转录因子结合位点来预测新结合位点的位置。如 Dinakarpandian 等[50]使用聚类算法对 PXR/RXRα 调节基因启动子的序列进行打分，然后通过集成神经网络、决策树和朴素贝叶斯分类算法训练串联机器学习模型，根据转录因子结合位点的位置和强度对启动子区域进行分类，在检测转录因子靶标基因中具有超过 70%的准确度。转录因子和结合位点之间的亲和力也影响着转录因子对基因的调节，支持向量回归模型 DNAffinity 不仅可以预测转录因子和结合位点之间的亲和力，还可以进一步预测碱基错配和突变对其相互作用的影响[93]。

经典机器学习算法需要手工设计特征或者进行特征选择，但是特征的选用会影响模型的准确度。Zhou 等[94]在解析转录因子与 DNA 的结合时，发现在基于支持向量回归算法的模型中整合 DNA 的 3D 形状特征可以降低特征空间的维度，使模型优于仅依赖于序列输入的模型，也为研究转录因子与基因组结合后产生的功能提供了新方法。相比于机器学习，深度学习无需人为设计特征，所以 BPNet[95]、MAResNet[96]和 DeepD2V[97]等工具选择了深度学习算法对转录因子结合位点进行预测设计。而相比于使用单一算法，组合不同的深度学习算法并利用其各自优势来构建集成框架，可以提高模型的性能和泛化能力。如 D-AEDNet 工具利用染色质免疫沉淀测序(Chromatin Immunoprecipitation Sequencing, ChIP-seq)数据识

别转录因子结合位点时，采用卷积神经网络学习编码器中的核苷酸位置信息，通过注意力门消除信息融合过程中的噪声响应，同时利用基于滑动窗口的转录因子结合位点发现方法可以在预测位点的同时理解预测结果的生物学意义，0.889 的模型准确率高于 DeepSNR 等[51]。这种综合利用特征表示和多算法集成的方法有助于推动核酸元件设计领域的发展和应用。

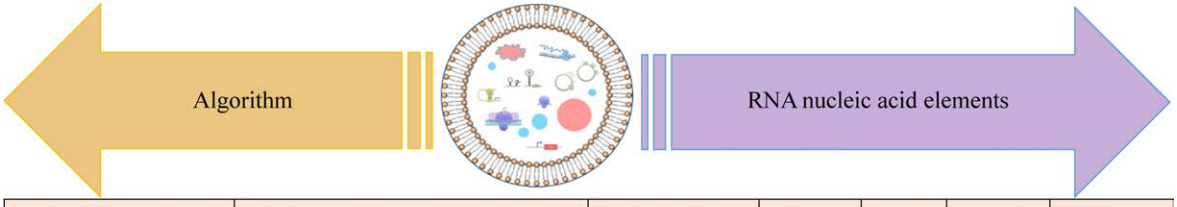
## 3 RNA 元件的智能预测与设计

除了上述 DNA 元件，核糖开关、sgRNA、miRNA、siRNA 和 circRNA 等 RNA 元件也具有重要的调控功能，为生物制造提供了丰富的精细调节工具。人工智能在这些 RNA 元件的挖掘预测和优化设计中发挥了重要作用(图 3 和表 2)，推动了 RNA 元件在生物制造中的广泛应用。

### 3.1 核糖开关

核糖开关可以直接结合小分子代谢物以调节下游基因的转录和翻译，具有结构简单、反应迅速等优点，广泛应用于生物制造领域[110]。通过人工智能技术可以有效预测核糖开关的功能，如使用卷积神经网络和双向递归神经网络构建的 Riboflow 可以对 32 种细菌核糖开关进行分类，准确率高达 99%[98]。此外，Angenent-Mari 等[99]利用多层感知机、卷积神经网络和长短期记忆循环神经网络所构成的模型来预测 23 个病毒和人类基因组中的核糖开关；通过 91 534 个核糖开关数据的训练，该深度神经网络模型的表现 ($R^2$=0.43–0.70) 优于热力学和动力学模型 ($R^2$=0.04–0.15)，同时引入注意力可视化技术 VIS4Map，可通过预测核糖开关的二级结构来准确预测其开关模式。

人工智能技术在解释核糖开关的机制方面或进一步设计新的核糖开关方面也有重要的作用。如深度神经网络模型 TISnet 可以预测植物和人类细胞 mRNA 上的翻译起始位点，发现位




| Model classification | Artificial intelligence algorithms | Riboswitches | sgRNA | miRNA | siRNA | circRNA |
|---|---|---|---|---|---|---|
| Traditional machine learning models | Linear regression model | | ✓ | | | |
| | Support vector machine | | ✓ | ✓ | ✓ | ✓ |
| | Random forest | | | ✓ | ✓ | ✓ |
| | Ensemble learning | | | ✓ | ✓ | |
| Deep learning models | Artificial neural network | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Convolutional neural network | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Recurrent neural networks | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Transformer | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Graph neural network | ✓ | | | ✓ | |

**图 3 RNA 核酸元件预测和设计中使用的人工智能算法**

Figure 3 Summary of artificial intelligence algorithms applied in the prediction and design of RNA nucleic acid elements.

表 2 RNA 核酸元件预测和设计中使用的代表性人工智能算法简介

Table 2 Introduction to representative artificial intelligence algorithms used in RNA nucleotide component prediction and design

| Nucleic acid elements | Representative tools | Main algorithm or architecture of the model | Applicable species or targets | Function | References |
|---|---|---|---|---|---|
| Riboswitches | Riboflow | Convolutional neural networks and bidirectional recurrent neural networks | Bacteria | Design genetic circuits using customized riboswitch aptamers to achieve precise translational control | [98] |
| Riboswitches | Deep neural network model | Logistic regression, multilayer perceptron, convolutional neural networks, and long short-term memory networks | Viral and human transcription factors | To accurately predict the switching pattern by predicting the secondary structure of riboswitches | [99] |
| Riboswitches | TISnet | Deep neural networks | *Arabidopsis*, *Nicotiana benthamiana*, human HEK293FT cells, yeast, etc. | Predicting translation start sites on plant and human cell mRNA, it is found that riboswitches can regulate translation by selecting different start codons | [100] |
| Riboswitches | NuSpeak, STORM | Convolutional neural networks, recurrent neural networks, multilayer perceptron, transfer learning | Human genome and viral genomes (such as SARS-CoV-2) | Design and optimization of riboswitches | [101] |
| sgRNA | autoBioSeqpy | Convolutional neural networks, recurrent neural networks | Type III secreted proteins, protein subcellular localization, and CRISPR/Cas9 sgRNA activity | The prediction of type III secreted proteins, protein subcellular localization, and CRISPR/Cas9 sgRNA activity | [102] |
| sgRNA | Iterative random forest model | Random intersection trees | *Escherichia coli*, human | Explain and predict sgRNA efficiency and enhance understanding of the complex quantum biological processes involved in the CRISPR-Cas9 mechanism | [103] |
| miRNA | TEC-miTarget | Encoder and convolutional neural networks | Human | Predicting the interaction between microRNA and its candidate target sites | [104] |
| miRNA | BrumiR | Random forest | Animals, plants | Predicting miRNA in animals and plants from sRNA-seq data | [105] |
| siRNA | OptiRNA | Support vector machine, bayesian sampling | Human | Predicting the Inhibition Efficiency of human-derived cell siRNA | [106] |
| siRNA | GNN4 siRNA | Graph neural networks, convolutional neural networks, long short-term memory networks, and dense neural networks | Human | Predicting siRNA-mRNA silencing efficiency related to cancer | [107] |
| siRNA | BERT-siRNA | BERT, multilayer perceptron, transfer learning | Human genome siRNA (Jurkat cell line) | Predict siRNA target gene knockdown efficiency | [108] |
| circRNA | circLGB-circMRT | LightGBM, support vector machine, random forest, stochastic gradient descent, gaussian naive bayes, circLGB, circMRT | Human | Predict the regulatory information of circRNA, including their interactions with microRNA, RNA-binding proteins, and transcriptional regulation | [34] |
| circRNA | CRIECNN | Convolutional neural networks, bidirectional long short-term memory networks, and self-attention mechanism | Human | Effectively predicting circRNA-RBP binding sites in full-length sequences | [109] |

于上游起始密码子下游的核糖开关的调控机制是通过介导不同起始密码子启动翻译进行的[100]。在设计新的核糖开关方面，Valeri 等[101]构建的语言模型 NuSpeak 可以在保证碱基互补配对的条件下针对病毒和人类基因组设计核糖开关；使用卷积神经网络模型 STORM 进一步优化核糖开关的性能，并利用迁移学习扩展模型适用范围，最终针对 20 种不同病毒基因组的平均马修斯相关系数约为 0.50。通过结合不同类型的神经网络模型和方法，可以更好地理解核糖开关的功能和调节机制，为生物制造提供有效的工具和方法。

## 3.2 sgRNA

生物制造离不开对基因的编辑和调节，其中 CRISPR/Cas 系统广泛应用于细胞工厂的基因组编辑[111-112]。sgRNA 是引导 Cas 蛋白对基因进行编辑的关键元件，其中预测 sgRNA 靶标和对应编辑效率的人工智能工具众多，如 CRISPRscan[113]、CRISPcut[114]、DeepCas9variant[115]、CRISep[116]、DeepHF[117]、DeepCRISPR[118]和 CRISPOR[119]等都可以高效预测 sgRNA 介导的编辑效率。研究人员全面比较了 36 种分别基于经典机器学习和深度学习算法的靶向 sgRNA 设计工具，发现考虑了序列和染色质特征的机器学习模型表现更好[120]。Zhang 等[121]则在 12 个公开数据集上比较了 8 种 sgRNA 靶标和脱靶活性预测工具，发现大多数工具在中等和大规模数据都有较高的预测准确度。因为训练集的不同，部分 sgRNA 的预测工具有物种特异性，如 CRISPRpred[122]和 sgRNA Scorer 2.0[123]适用于化脓性链球菌、嗜热链球菌和金黄色葡萄球菌；DeepGuide[124]适用于解脂耶氏酵母；CNN_5layers[125]，EuPaGDT[126]和 CRISPR-P[127]分别适用于细菌、真核病原体和植物等。

特征提取也会影响预测 sgRNA 编辑效率模型的准确度。如生物序列分类工具 autoBioSeqpy 使用卷积网络、循环网络和卷积循环网络分别构建模型，通过独热编码、字典编码及不同长度的 K-mer 作为序列特征，发现字典编码在所有数据集中都显著优于独热编码($P$<0.05)，同时发现将卷积网络和循环网络混合的 CNN-biLSTM 模型表现最好，准确率约为 91.8%[102]。也有研究从量子化学角度优化 sgRNA 的特征表示方法，如 Noshay 等[103]针对大肠杆菌和智人开发的基于迭代随机森林算法训练的模型，利用碱基对寡聚体描述 sgRNA 结构中的核苷酸位置，用二元编码和量子化学特性评估 sgRNA 的子序列结构信息，预测精度为 0.51，特征工程凸显了量子力学的作用。特征提取对 sgRNA 的模型编码效率有较大影响，尝试不同的特征编码方式可以有效优化模型的性能，加深对核酸元件设计和生物学过程的理解。

## 3.3 miRNA

人工智能在 miRNA 的研究和应用中也发挥重要作用，包括挖掘新的 miRNA，预测 miRNA 靶基因及 miRNA 与 mRNA 的相互作用等。PHDcleav[128]、LBSizeCleav[129]、ReCGBM[130]、DeepMirCut[131]、DiCleave[132]和 miRanalyzer[133]等机器学习模型可以预测 miRNA 或其靶标。同样由于训练集的不同，部分机器学习模型只针对特定物种，如使用随机森林算法的 BrumiR[105]和 Mirnovo[134]可以利用 sRNA-Seq 数据从动物和植物中挖掘未知的 miRNA。BrumiR 使用随机森林算法，能够直接从 sRNA-seq 数据中发现 miRNA，同时由于其无需对齐和使用基于图形的方法，具有运算速度快(比 miRDeep2 快 21 倍，比 MiR-PREFeR 快 6 倍)和准确率高(97%)的优势[105]。由于 miRNA 的结构特征包括其二级结构、碱基对和稳定性等信息，而自由能则反映了 miRNA 分子的稳定性和互补配对等特性。引

入 miRNA 的结构特征和自由能可以有效提高机器学习模型预测 miRNA 的准确度，如 LBSizeCleav 利用环/凸起结构的长度特征改进了 miRNA 预测，相比于 PHDcleav，其准确率达到了 85.1%[129]。

随着人工智能技术的发展，多种基于深度学习算法的 miRNA 预测工具也被开发出来，包括基于卷积神经网络和循环神经网络的 miTAR[135]，基于 Transformer 和卷积神经网络的 TEC-miTarget[104]，由自动编码器和前馈网络组成的深度神经网络模型 miRAW[136]，基于残差神经网络的 TargetNet[137]和基于图神经网络的 PDMDA[138]等。这些模型使用不同的神经网络算法或其组合对结构和特征进行表示，能够有效地预测 miRNA 的功能，解析 miRNA 与靶基因相互作用的调控机制。其中 TEC-miTarget 利用 Transformer 编码器处理 RNA 序列，使用自注意力机制和四层卷积神经网络提取接触图的特征，从而预测 miRNA 与其候选目标位点之间的相互作用；在 miRaw 数据集上的准确度为 96.47%，灵敏度为 95.85%，F1 分数为 97.40%，优于 PITA、mirSVR、miRDB、microT 和 Targetscan 等工具[104]。

数据库可以为人工智能提供数据，人工智能也可以帮助并建立了多个 miRNA 数据库[139-141]。如基于支持向量机模型 MirTarget 开发的在线数据库 miRDB，可以用于 miRNA 的靶点预测和功能注释；miRDB 包含 5 个物种中 7 000 个左右的 miRNA 及其对应的约 350 万个靶点的预测数据[142]。这些数据库在 miRNA 的功能注释和验证、种间比较及新 miRNA 的发现中都具有重要意义。

## 3.4 siRNA

siRNA 的设计对靶基因的沉默效果具有显著影响，因此已经开发了多种机器学习模型来对 siRNA 进行分类、预测其沉默效率并进行设计。如随机森林模型 CoRAL[143]、集成学习模型 ILGBMSH[144]及随机森林和支持向量机的组合模型等[145-147]，其中支持向量机算法适合处理自由能、靶位点可及性和核苷酸特征等高维数据。例如 Ladunga[106]设计的支持向量机模型 OptiRNA，该模型从 2 200 多个 siRNA 的 572 个特征(包括热力学性质、可及性以及发夹结构等)中进行学习，能够预测具有最佳抑制效果的 siRNA，在交叉验证实验中准确率达到了 92.3%。相比之下，随机逻辑回归在解释模型方面更有优势，可以寻找 siRNA 的关键特征。如 Klingelhoefer 等[148]使用随机逻辑回归算法来识别与 siRNA 抑制效率相关的特征，算法揭示了包括反义的 UCU 和 ACGA 序列等多个与 siRNA 沉默效率相关的特征，为研究 siRNA 介导的沉默机制提供了工具。

除了经典的机器学习，多种深度学习算法因其独特的算法优势也被用于 siRNA 的沉默效率预测和设计中。利用人工神经网络在学习复杂非线性关系方面表现突出的特点，人工神经网络模型 BIOPREDsi 有效预测了 249 个 siRNA 的沉默效率，同时利用该模型有效筛选出了靶向低氧反应元件(hypoxic response element, HRE)、低氧诱导因子(Hypoxia-inducible factor 1α, HIF1A)和芳香烃受体核转位蛋白(aryl hydrocarbon receptor nuclear translocator, ARNT)的 siRNA[149]。相比于人工神经网络算法，图神经网络在捕捉复杂图结构中的信息方面具有显著优势，适合于分子结构及分子间相互作用的建模[150]。有研究人员设计了用于分析 siRNA-mRNA 相互作用的图神经网络工具 GNN4_siRNA，它可以学习 siRNA 序列特征和热力学特征并用以预测 siRNA 沉默特定靶基因的效率，其在基准数据集上优于 BIOPREDsi[149]、DSIR[151]和 siRNApred[152-153]等机器学习算法，其皮尔逊相关系数约 73.6%[107]。但是图神经网络计算复杂度高且训练时间长，

基于 Transformer 架构的 BERT 算法能够快速捕捉上下文信息和复杂的序列依赖关系，适合处理序列数据。如由 DNA-BERT 预训练模块和多层感知器模块构成的 BERT-siRNA 工具应用迁移学习概念来避免小样本量的偏差和大量的预处理过程，在独立公共 siRNA 数据集的测试中明显优于其他 siRNA 沉默效率预测模型[108]。各种人工智能模型在 siRNA 研究中均有广泛应用，需根据具体研究需求、数据特征和计算资源进行权衡来选择合适的模型。

### 3.5 circRNA

circRNA 形成的稳定闭环结构使其不容易被 RNA 外切酶降解，有助于提高蛋白产量，所以在蛋白表达、mRNA 治疗和 RNA 疫苗中广泛应用。同时 circRNA 也可以通过与 miRNA 或代谢相关蛋白结合来调节细胞代谢网络[154-158]。目前，已经开发了多种 circRNA 预测工具，例如基于机器学习算法的 circLGB 和 circMRT 可以区分 circRNA 和 lncRNA，并预测 circRNA 与 miRNA、RNA 结合蛋白和转录因子的相互作用[34]。A-to-I 脱氨基作用、A-to-I 密度和内部核糖体进入位点等多个特征的增加提高了 circLGB 的分类准确度，通过 191 个特征训练的 circLGB 其受试者工作特征曲线下面积值达到 0.999，优于 circDeep[159]和 PredcircRNA[160]等预测工具。基于深度学习的算法也被应用于环状 RNA 分类，其中采用多层卷积神经网络和双向长短期记忆网络来学习高阶特征的 CircDC 可以预测 circRNA，并将 circRNA 与 lncRNA 进行分类，模型准确率和灵敏度分别达到 0.861 4 和 0.838 1，优于 WebCircRNA[161]和 CirRNAPL[162]。

此外，多种深度学习算法也用于预测 circRNA 或其与蛋白的相互作用，如用于预测人类 circRNA 反向剪接的卷积神经网络模型 DeepCirCode[163]和预测 circRNA 与 RNA 结合蛋白互作的多算法组合模型 CRIECNN[109]等。其中 DeepCirCode 与支持向量机和随机森林等传统机器学习算法相比表现出优越的性能，成功预测出 RNA 剪接、转录或翻译的基序[163]。综合利用多种深度学习算法可进一步提高模型准确度，如 CRIECNN 集成了卷积神经网络、双向长短时记忆网络和自注意力机制，测试了 K-mer、Doc2Vec、BERT 和 EIIP 等提取特征的方法，可在全长序列中有效预测 circRNA 与 RNA 结合蛋白的结合位点；在预测准确率方面，CRIECNN 在所有 37 个子数据集上都表现优异，优于 CRIP、HCRNet、PASSION 和 iCircRBP-DHN 等模型，平均受试者工作特征曲线下的面积(area under receiver operating characteristic curve, AUC)为 0.957 7[109]。

## 4 核酸元件的智能设计在生物制造中的应用

人工智能模型可以学习生物制造中产物产量、底物浓度、温度、pH、碳源、氮源和氧气条件等代谢数据以及菌株基因组、代谢关键基因的启动子、核糖体结合位点、终止子和基因拷贝数等遗传数据，进而对核酸元件进行优化[66,164-166]。目前，人工智能技术在生物制造中的应用主要是优化启动子和核糖体结合位点等核酸元件，以更短的时间和更少的工作量优化代谢流，提高目标产品产量。下面将分别阐述生物制造中核酸元件常用的智能优化方法、机器学习算法以及数据集的有效利用(表 3)。

### 4.1 人工智能优化关键基因的转录与翻译

通过人工智能技术优化关键基因的表达强度可以有效提高细胞工厂目标产物的产量。以基因的转录水平优化为例，为了增加高温条件下酿酒酵母的乙醇产量，Khamwachirapithak 等[167]选择了酿酒酵母的乙醇生产限速酶基因 *ADH1*、

**表 3 核酸元件在生物制造应用中的代表性案例**

Table 3 Representative cases of nucleic acid elements in bio-manufacturing applications

| Product | Nucleic acid element | Dataset size (strain library size) | Algorithm | Production improvement | References |
|---|---|---|---|---|---|
| Ethanol | Promoters of the ethanol production rate-limiting enzyme genes *ADH1* and *PDC1*, the heat and pressure response gene *TPS1* | 216 | Linear regression, generalized linear regression, decision tree, random forest, support vector machine, XGBoost | The ethanol yield at 40 °C increased by 7.4% | [167] |
| Dodecyl alcohol | RBSs of the sulfotransferase, acyl-CoA reductase, and acyl-CoA synthetase genes | 60 | Random forest, polynomial regression, neural regression, and tree-based pipeline optimization tool | The yield of dodecanol increased by 21% in the first training cycle and increased sixfold (0.83 g/L) in the second training cycle | [168] |
| Coumaric acid | Promoters and terminators of key genes in the phenylalanine pathway and the tyrosine pathway | 440 | Multiple linear regression, support vector regression, kernel ridge regression, and random forest | The yield of coumaric acid increased by 68% (0.52 g/L) | [169] |
| Carotenoids | Key gene promoters in the methyl erythritol phosphate pathway | 163 | 24 types of machine learning methods | The optimized strain produced 3.18 mg/L/$OD_{600}$ of carotenoids, a 4.3-fold increase over the parent strain | [170] |
| Limonene | RBSs of key gene for the limonene synthesis and the mevalonate pathway | 5 184 | Support vector machine, feedforward neural network | Boosted production by over 60% after screening only 3% of the library | [67] |
| Terpenoids | Promoters of five genes in the mevalonate pathway | 243 | Random forest | The geraniol, humulene and squalene titers were increased by 94-fold, 60-fold, and 35-fold, respectively | [171] |
| Tryptophan | Promoters for the tryptophan production pathway genes (*PCK1*, *TAL1*, *TKL1*, *CDC19*, *PFK1*) | 7 776 | Bayesian ensemble, bayesian optimization algorithm | The tryptophan production increased by 74% | [172] |
| Threonine | 16 genes related to threonine synthesis | 385 | Deep learning model (consisting of a batch normalization layer followed by 2 to 9 feedforward layers) | The L-threonine titer increased from 2.7 g/L to 8.4 g/L | [173] |
| Violaquinone | The promoters of the five genes in the biosynthetic pathway of violacein (*vioA*, *vioB*, *vioE*, *vioD*, *vioC*) | 91 | Linear regression model | High-accuracy Pearson correlation coefficients: 0.80 for violacein, 0.77 for deoxyviolacein, 0.83 for proviolacein and 0.92 for prodeoxyviolacein | [174] |
| Carotene, violaxanthin | Production pathways of β-carotene and violaxanthin | 24 | Artificial neural network, ensemble learning | The β-carotene yield increased by 64.4%, and the violaxanthin yield increased by 2.42 times | [175] |

*PDC1* 和热压力应答基因 *TPS1* 作为优化靶标，利用不同强度启动子组合表达这 3 个基因，构建了一个由 216 个菌株组成的酿酒酵母文库；利用启动子强度和 30 °C 条件下乙醇产量作为输入信号，对比了 6 种机器学习算法，发现 XGBoost 在模型调整和正则化中更为灵活有效，在第二轮测试中就使 40 °C 时的乙醇产量增加了 7.4%。除了通过改变启动子强度来优化代谢途径，人工智能技术也可调节关键基因的翻译强度来提高菌株的产量。Opgenorth 等[168]对硫酯酶、酰基辅酶 A 还原酶、酰基-CoA 合成酶的 RBS 设计了含有 60 个菌株的大肠杆菌文库，利用随机森林、多项式回归、神经回归和基于树的管道优化工具构建集成学习模型，在第 1 个测试周期将十二醇滴度提高了 21%，并进一步在第 2 个周期将十二醇产量提升了 6 倍，达到了 0.83 g/L。

除了单独改变启动子和 RBS 的转录和翻译强度，还可以利用算法优化不同核酸元件的组合以提高菌株性能。香豆酸是一种来源于植物的重要酚酸，其合成途径主要包括苯丙氨酸途径和酪氨酸途径。Moreno-Paz 等[169]将苯丙氨酸途径和酪氨酸途径中的关键基因分别用不同的启动子和终止子组合来进行表达，然后随机测定了 440 株菌的香豆酸产量及对应的启动子和终止子序列；使用获得的数据对多元线性回归、支持向量回归、核岭回归和随机森林这 4 种机器学习模型进行训练，后 3 种方法都有比较好的表现；最终香豆酸在酿酒酵母中的产量提高了 68%，达到了 0.52 g/L 的水平。

## 4.2 生物制造中机器学习算法的选择

由于在所有机器学习算法中，没有一个算法能够在所有场景下都表现得比其他算法更好，所以研究人员往往需要同时比较多种机器学习算法，寻找在特定数据集上表现最优的算法用于迭代。如 Shimazaki 等[170]通过对 24 种机器学习方法的测试，成功预测了酿酒酵母合成类胡萝卜素时甲羟戊酸途径中 *tHMG1*、*ERG20*、*ERG19* 和 *ERG13* 等关键基因的表达水平，模型指导构建菌株的类胡萝卜素生产力为亲本菌株的 4.3 倍。Kang 等[176]比较了多层感知机、随机森林、支持向量机和极端梯度提升等机器学习算法，最终选择多层感知器结合遗传算法从 2 047 个可能的组合中预测潜在的过表达靶点，通过机器学习的辅助对预测基因进行微调，分别使改造后菌株 LY04 的番茄红素产量提高了 8 倍和 6 倍，最高产量为 1.25 g/L。

不同的机器学习算法各有优势，如支持向量机算法的参数相对较少、调参相对简单且泛化能力强，对小样本数据表现良好[177]，适用于难以获得大量实验数据的代谢工程实验。Jervis 等[67]分别选用 12 个不同强度的 RBS 来组合表达柠檬烯合成关键基因 *GPPS* 和 *LimS*，分别测定柠檬烯的产量后来训练支持向量机模型，发现模型可以准确预测 RBS 序列所对应菌株的柠檬烯产量；然后用相同的方法对甲羟戊酸途径中的 *mvaS*、*mvaE*、*mvaK1* 和 *IDI* 基因构建 RBS 组合表达文库，进而对 RBS 序列和柠檬烯的产量关系进行建模，仅筛选不到 3%的文库就能够使产量提高 60%以上。相比于支持向量机，随机森林由多个决策树组成，具有可解释性、不易过拟合且预测准确性高等优点。Mukherjee 等[171]针对甲羟戊酸途径的 5 个基因构建了由不同启动子进行表达的由 243 个酿酒酵母菌株组成的文库；随机森林模型显示，编码甲羟戊酸激酶的 *ERG12* 是除 *HMG1* 和 *IDI1* 外最关键的基因，进一步将甲醛戊酸途径基因定位到细胞质和过氧化物酶体可以分别使香叶醇、蛇麻烯和角鲨烯的产量提高了 94 倍、60 倍和 35 倍。相比于支持向量机和随机森林算法，贝叶斯集成

算法具有连续学习、自适应调整和快速适应新信息的能力。贝叶斯优化算法则能够在高维度和复杂搜索空间中寻找最优解,比传统的优化算法(如梯度下降)搜索效率更高。如 Zhang 等[172]设计利用 6 个启动子分别表达戊糖磷酸途径和糖酵解途径的 5 个靶基因，然后通过同源重组以及 CRISPR/Cas9 技术对设计的 7 776 个组合菌株进行构建，进而通过色氨酸生物传感器对菌株的产量进行测定；在收集数据的基础上，利用基于贝叶斯集成的 ART 和贝叶斯优化的 EVOLVE 对酿酒酵母的色氨酸代谢进行学习和设计预测，最终最优设计菌株的色氨酸产量提升了 74%。

相比于经典的机器学习算法，深度学习能够通过反向传播机制自动学习特征，不需要手动进行特征提取，因而适合处理复杂数据并在大规模数据集上表现出色。如 Hanke 等[173]利用深度学习模型 DeepLearning4J 进行训练并预测和开发高产苏氨酸菌株；首先根据苏氨酸合成相关的 16 个基因构建了基因敲除、低表达和高表达的 385 个组合菌株，进一步发酵测定了各个菌株的苏氨酸产量，然后通过调整特定基因组合的权重来改进模型，仅经过 3 轮迭代组合和模型预测，就将 L-苏氨酸产量从 2.7 g/L 提升到了 8.4 g/L。不同的机器学习算法对实验数据的适用性各有差异，除此之外数据集的质量和处理方式也对生物制造中核酸元件的智能设计起着至关重要的作用。

### 4.3 生物制造中数据集的构建原则

相比之下，生物制造中菌株的构建成本高，因此有必要探索训练模型所需的最小数据集。研究人员发现通过控制数据的多样性和分布的均匀性，较小的数据集也可以训练出较准确的模型。如 Nikolados 等[178]为了研究实验数据训练模型的有效性和准确性，测定了绿色荧光蛋白翻译序列和表达量的关系，进而比较了岭回归、多层感知器、支持向量回归和随机森林回归等 4 种机器学习模型和卷积神经网络对蛋白表达量的预测准确度；通过分析发现，训练机器学习模型只需 1 000–2 000 个序列即可达到较高的准确度，而在 2 000 条序列条件下，卷积神经网络比机器学习模型中突变序列的中值预测分数高 10%，这也为研究人员合理设计文库，利用小数据集训练模型，获得高产菌株提供了可能。同样，Lee 等[174]使用线性回归模型来预测紫罗兰素生物合成途径中 *vioA*、*vioB*、*vioE*、*vioD*、*vioC* 这 5 个基因的启动子强度，从而优化紫罗兰素、脱氧紫罗兰素、原紫罗兰素和原脱氧紫罗兰素的生产；利用 91 个样本的启动子强度和产物产量对模型进行训练后，其中紫罗兰素、脱氧紫罗兰素、紫罗兰素和原紫罗兰素的皮尔逊相关系数均达到了较高的精度，分别为 0.80、0.77、0.83 和 0.92；同时发现,在样本量为 5 和 10 时的训练精度快速提升，但当将训练集增加到 50 或 91 个样本时，训练精度仅有适度的改进，这表明 1%–2%这种相对较低的采样率足以训练模型并达到较高的准确率。除此之外，研究人员设计了人工神经网络算法 MiYA 用于组合优化酿酒酵母中导入的异源代谢途径；在仅搜索 2%–5%的组合空间时，MiYA 即可精确调整 β-胡萝卜素生物合成途径；同时，MiYA 利用 24 株紫罗兰素生产菌株的数据就成功预测并构建了产量提高 2.42 倍的菌株，也表明较小的数据集可以训练出较好的模型[175]。

## 5 总结与展望

DNA 和 RNA 核酸元件广泛应用于生物制造领域，因此多种人工智能工具也被开发出来用于核酸元件预测和设计。除了预测和挖掘基

因组中已有的核酸元件，通过输入核酸元件的稳定性、特异性和亲和性等信息，人工智能技术可以通过模型分析序列与功能之间的关系，理解核酸元件工作机制。此外，通过整合生物信息学、机器学习和进化算法等技术，人工智能还可以实现对核酸元件的结构、功能和稳定性等多方面的优化，精准设计具有多种功能的核酸元件，并帮助生成具有目的特异性和序列多样性的新元件。

人工智能模型的训练通常依赖于大量高质量的数据，但是由于生物的多样性和复杂性，模型的通用性经常不足，限制了人工智能在生物制造中的应用。同时，由于部分人工智能算法具有“黑箱”特性，不透明的决策过程导致难以将核酸元件的功能机制和模型参数联系起来，这也为人工智能在生物制造中的应用带来了不确定性。此外，利用人工智能模型来设计核酸元件需要实验人员具有丰富的计算经验，但是生物数据的高维度和多样性往往会增加模型设计和调试的难度。以上原因也导致了核酸元件的智能设计在生物制造领域中的应用案例较少。由于通过代谢工程实验来获取数据的成本比较高，且生物学特征和实验条件复杂，这使得适用于处理结构化数据和特征较少的传统机器学习算法(如支持向量机、随机森林、贝叶斯网络)在代谢工程中的应用更多[179]。尽管深度学习在图像识别和语言处理领域中取得了更大的成功，但能够在小训练集上表现良好且稳定的机器学习方法仍为生物制造应用的首选。

未来生物制造中核酸元件的智能设计需要在以下几个方面继续提升:(1) 数据质量和数量的提升。通过收集和整理更多高质量的核酸元件数据，建立标准化的数据集，从而提升人工智能模型的训练效果和预测准确性。(2) 模型的可解释性。提升人工智能算法的透明度与可解释性，可以帮助研究人员理解核酸元件的作用机制，同时提高模型的可信度。(3) 多模态数据融合。结合包括基因组、转录组和代谢组等多种生物数据源，提升模型的综合分析能力，为生物制造设计更精准的核酸元件。(4) 计算资源的优化。通过算法优化和硬件加速来提升人工智能模型的训练效率和计算性能，降低资源消耗和训练成本。未来人工智能技术在生物制造领域的应用需要增强各学科间的合作，同时综合利用各种算法优势，从多个角度挖掘数据中的潜在规律和信息，为核酸元件的设计和优化提供更多有效的工具，进一步加速生物制造领域的发展。

## 作者贡献声明

王金盛：初稿写作；孙喆、张学礼：经费支持、监督指导和稿件润色修改。

## 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] KAIKKONEN MU, LAM MT, GLASS CK. Non-coding RNAs as regulators of gene expression and epigenetics[J]. Cardiovascular Research, 2011, 90(3): 430.
[2] CHUANG LY, TSAI JH, YANG CH. Binary particle swarm optimization for operon prediction[J]. Nucleic Acids Research, 2010, 38(12): e128.
[3] JIANG YR, LUO J, HUANG DQ, LIU Y, LI DD. Machine learning advances in microbiology: a review of methods and applications[J]. Frontiers in Microbiology, 2022, 13: 925454.
[4] YANG RT, WU F, ZHANG CJ, ZHANG LN. iEnhancer-GAN: a deep learning framework in combination with word embedding and sequence generative adversarial net to identify enhancers and their strength[J]. International Journal of Molecular Sciences, 2021, 22(7): 3589.
[5] QUANG D, XIE XH. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences[J]. Nucleic Acids Research, 2016, 44(11): e107.

[6] VORLÄNDER MK, PACHECO-FIALLOS B, PLASCHKA C. Structural basis of mRNA maturation: time to put it together[J]. Current Opinion in Structural Biology, 2022, 75: 102431.
[7] SOLLER M. Pre-messenger RNA processing and its regulation: a genomic perspective[J]. Cellular and Molecular Life Sciences, 2006, 63(7/8): 796-819.
[8] BHUKYA R, KUMARI A, AMILPUR S, DASARI CM. PPred-PCKSM: a multi-layer predictor for identifying promoter and its variants using position based features[J]. Computational Biology and Chemistry, 2022, 97: 107623.
[9] STRUHL K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes[J]. Cell, 1999, 98(1): 1-4.
[10] RANGEL-CHAVEZ C, GALAN-VASQUEZ E, MARTINEZ-ANTONIO A. Consensus architecture of promoters and transcription units in *Escherichia coli*: design principles for synthetic biology[J]. Molecular BioSystems, 2017, 13(4): 665-676.
[11] KRISHNAMURTHY S, HAMPSEY M. Eukaryotic transcription initiation[J]. Current Biology, 2009, 19(4): R153-R156.
[12] RINALDI AJ, LUND PE, BLANCO MR, WALTER NG. The Shine-Dalgarno sequence of riboswitch-regulated single mRNAs shows ligand-dependent accessibility bursts[J]. Nature Communications, 2016, 7: 8976.
[13] KOZAK M. Structural features in eukaryotic mRNAs that modulate the initiation of translation[J]. Journal of Biological Chemistry, 1991, 266(30): 19867-19870.
[14] BULGER M, GROUDINE M. Functional and mechanistic diversity of distal transcription enhancers[J]. Cell, 2011, 144(3): 327-339.
[15] KLEFTOGIANNIS D, KALNIS P, BAJIC VB. Progress and challenges in bioinformatics approaches for enhancer identification[J]. Briefings in Bioinformatics, 2016, 17(6): 967-979.
[16] TOGNON M, GIUGNO R, PINELLO L. A survey on algorithms to characterize transcription factor binding sites[J]. Briefings in Bioinformatics, 2023, 24(3): bbad156.
[17] FISCHER V, SCHUMACHER K, TORA L, DEVYS D. Global role for coactivator complexes in RNA polymerase II transcription[J]. Transcription, 2019, 10(1): 29-36.
[18] LAWSON MR, BERGER JM. Tuning the sequence specificity of a transcription terminator[J]. Current Genetics, 2019, 65(3): 729-733.
[19] SALVAIL H, BREAKER RR. Riboswitches[J]. Current Biology, 2023, 33(9): R343-R348.
[20] SCHMIDT CM, SMOLKE CD. RNA switches for synthetic biology[J]. Cold Spring Harbor Perspectives in Biology, 2019, 11(1): a032532.
[21] KAVITA K, BREAKER RR. Discovering riboswitches: the past and the future[J]. Trends in Biochemical Sciences, 2023, 48(2): 119-141.
[22] GREEN AA, SILVER PA, COLLINS JJ, YIN P. Toehold switches: de-novo-designed regulators of gene expression[J]. Cell, 2014, 159(4): 925-939.
[23] HANNON GJ. RNA interference[J]. Nature, 2002, 418(6894): 244-251.
[24] SALMENA L, POLISENO L, TAY Y, KATS L, PANDOLFI PP. A *ceRNA* hypothesis*:* the Rosetta stone of a hidden RNA language?[J]. Cell, 2011, 146(3): 353-358.
[25] TAY Y, RINN J, PANDOLFI PP. The multilayered complexity of *ceRNA* crosstalk and competition[J]. Nature, 2014, 505(7483): 344-352.
[26] SEN R, GHOSAL S, DAS S, BALTI S, CHAKRABARTI J. Competing endogenous RNA: the key to posttranscriptional regulation[J]. The Scientific World Journal, 2014, 2014: 896206.
[27] HESS GT, TYCKO J, YAO D, BASSIK MC. Methods and applications of CRISPR-mediated base editing in eukaryotic genomes[J]. Molecular Cell, 2017, 68(1): 26-43.
[28] NISHIDA K, KONDO A. CRISPR-derived genome editing technologies for metabolic engineering[J]. Metabolic Engineering, 2021, 63: 141-147.
[29] KILDEGAARD KR, TRAMONTIN LRR, CHEKINA K, LI MJ, GOEDECKE TJ, KRISTENSEN M, BORODINA I. CRISPR/Cas9-RNA interference system for combinatorial metabolic engineering of *Saccharomyces cerevisiae*[J]. Yeast, 2019, 36(5): 237-247.
[30] LIU D, MANNAN AA, HAN YC, OYARZÚN DA, ZHANG FZ. Dynamic metabolic control: towards precision engineering of metabolism[J]. Journal of Industrial Microbiology & Biotechnology, 2018, 45(7): 535-543.
[31] ZHANG YP, SUN JB, MA YH. Biomanufacturing: history and perspective[J]. Journal of Industrial Microbiology & Biotechnology, 2017, 44(4-5): 773-784.
[32] LIEBAL UW, KÖBBING S, NETZE L, SCHWEIDTMANN AM, MITSOS A, BLANK LM. Insight to gene expression from promoter libraries with the machine learning workflow Exp2Ipynb[J]. Frontiers in Bioinformatics, 2021, 1: 747428.
[33] JACOBS TM, YUMEREFENDI H, KUHLMAN B, LEAVER-FAY A. SwiftLib: rapid degenerate-*codon*-library optimization through dynamic programming[J]. Nucleic Acids Research, 2015, 43(5): e34.
[34] ZHANG GS, DENG YY, LIU QY, YE BX, DAI ZM, CHEN YW, DAI XH. Identifying circular RNA and predicting its regulatory interactions by machine learning[J]. Frontiers in Genetics, 2020, 11: 655.
[35] de BOER CG, VAISHNAV ED, SADEH R, ABEYTA EL, FRIEDMAN N, REGEV A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters[J]. Nature Biotechnology, 2020, 38(1): 56-65.
[36] LIU XY, GUPTA STP, BHIMSARIA D, REED JL, RODRÍGUEZ-MARTÍNEZ JA, ANSARI AZ, RAMAN S. *De novo* design of programmable inducible promoters[J]. Nucleic Acids Research, 2019, 47(19): 10452-10463.
[37] HE WY, JIA CZ, DUAN YC, ZOU Q. 70ProPred: a

predictor for discovering sigma70 promoters based on combining multiple features[J]. BMC Systems Biology, 2018, 12(Suppl 4): 44.
[38] 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.
[39] UMAROV RK, SOLOVYEV VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks[J]. PLoS One, 2017, 12(2): e0171410.
[40] 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.
[41] 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(1): 6309.
[42] 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.
[43] ZHANG S, HU HL, JIANG T, ZHANG L, ZENG JY. TITER: predicting translation initiation sites by deep learning[J]. Bioinformatics, 2017, 33(14): i234-i242.
[44] 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(1): 3551.
[45] FRANCIS-LYON P, CRISTIANINI N, HOLBROOK S. *Terminator* detection by support vector machine utilizing a stochastic context-free grammar[C]//2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. April 1–5, 2007, Honolulu, HI, USA. IEEE, 2007: 170-177.
[46] ZHAI WJ, DUAN YT, ZHANG XM, XU GQ, LI H, SHI JS, XU ZH, ZHANG XJ. Sequence and thermodynamic characteristics of terminators revealed by FlowSeq and the discrimination of terminators strength[J]. Synthetic and Systems Biotechnology, 2022, 7(4): 1046-1055.
[47] KLEFTOGIANNIS D, KALNIS P, BAJIC VB. DEEP: a general computational framework for predicting enhancers[J]. Nucleic Acids Research, 2015, 43(1): e6.
[48] HAFEZ D, KARABACAK A, KRUEGER S, HWANG YC, WANG LS, ZINZEN RP, OHLER U. McEnhancer: predicting gene expression *via* semi-supervised assignment of enhancers to target genes[J]. Genome Biology, 2017, 18(1): 199.
[49] KÄHÄRÄ J, LÄHDESMÄKI H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data[J]. Bioinformatics, 2015, 31(17): 2852-2859.
[50] DINAKARPANDIAN D, RAHEJA V, MEHTA S, SCHUETZ EG, ROGAN PK. Tandem machine learning for the identification of genes regulated by transcription factors[J]. BMC Bioinformatics, 2005, 6: 204.
[51] ZHANG YQ, WANG ZX, ZENG YQ, ZHOU JL, ZOU Q. High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method[J]. Briefings in Bioinformatics, 2021, 22(6): bbab273.
[52] PATIYAL S, SINGH N, ALI MZ, PUNDIR DS, RAGHAVA GPS. Sigma70Pred: a highly accurate method for predicting sigma70 promoter in *Escherichia coli* K-12 strains[J]. Frontiers in Microbiology, 2022, 13: 1042127.
[53] Di SALVO M, PINATEL E, TALÀ A, FONDI M, PEANO C, ALIFANO P. G4PromFinder: an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs[J]. BMC Bioinformatics, 2018, 19(1): 36.
[54] RAHMAN MS, AKTAR U, JANI MR, SHATABDA S. iPro70-FMWin: identifying Sigma70 promoters using multiple windowing and minimal features[J]. Molecular Genetics and Genomics, 2019, 294(1): 69-84.
[55] LIU B, LI K. iPromoter-2L2.0: identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features[J]. Molecular Therapy-Nucleic Acids, 2019, 18: 80-87.
[56] LIU B, YANG F, HUANG DS, CHOU KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC[J]. Bioinformatics, 2018, 34(1): 33-40.
[57] CASSIANO MHA, SILVA-ROCHA R. Benchmarking bacterial promoter prediction tools: potentialities and limitations[J]. mSystems, 2020, 5(4): e00439-20.
[58] ALPER H, FISCHER C, NEVOIGT E, STEPHANOPOULOS G. Tuning genetic control through promoter engineering[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(36): 12678-12683.
[59] 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.
[60] NAIR TM. Calliper randomization: an artificial neural network based analysis of *E. coli* ribosome binding sites[J]. Journal of Biomolecular Structure & Dynamics, 1997, 15(3): 611-617.
[61] BISANT D, MAIZEL J. Identification of ribosome binding sites in *Escherichia coli* using neural network models[J]. Nucleic Acids Research, 1995, 23(9): 1632-1639.
[62] FANG H, HUANG YF, RADHAKRISHNAN A, SIEPEL A, LYON GJ, SCHATZ MC. Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution[J]. Cell Systems, 2018, 6(2): 180-191.e4.
[63] ZHANG MY, HOLOWKO MB, HAYMAN ZUMPE H,

ONG CS. Machine learning guided batched design of a bacterial ribosome binding site[J]. ACS Synthetic Biology, 2022, 11(7): 2314-2326.
[64] SPRENGART ML, FUCHS E, PORTER AG. The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*[J]. EMBO Journal, 1996, 15(3): 665-674.
[65] SALIS HM. The ribosome binding site calculator[J]. Methods in Enzymology, 2011, 498: 19-42.
[66] FARASAT I, KUSHWAHA M, COLLENS J, EASTERBROOK M, GUIDO M, SALIS HM. Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria[J]. Molecular Systems Biology, 2014, 10(6): 731.
[67] JERVIS AJ, CARBONELL P, VINAIXA M, DUNSTAN MS, HOLLYWOOD KA, ROBINSON CJ, RATTRAY NJW, YAN CY, SWAINSTON N, CURRIN A, SUNG R, TOOGOOD H, TAYLOR S, FAULON JL, BREITLING R, TAKANO E, SCRUTTON NS. Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli*[J]. ACS Synthetic Biology, 2019, 8(1): 127-136.
[68] DING NN, YUAN ZQ, ZHANG XJ, CHEN J, ZHOU SH, DENG Y. Programmable cross-ribosome-binding sites to fine-tune the dynamic range of transcription factor-based biosensor[J]. Nucleic Acids Research, 2020, 48(18): 10602-10613.
[69] ALLFANO P, RIVELLINI F, LIMAURO D, BRUNI CB, CARLOMAGNO MS. A consensus motif common to all rho-dependent prokaryotic transcription terminators[J]. Cell, 1991, 64(3): 553-563.
[70] CIAMPI MS. Rho-dependent terminators and transcription termination[J]. Microbiology, 2006, 152(Pt 9): 2515-2528.
[71] ROBERTS JW. Mechanisms of bacterial transcription termination[J]. Journal of Molecular Biology, 2019, 431(20): 4030-4039.
[72] PETERS JM, VANGELOFF AD, LANDICK R. Bacterial transcription terminators: the RNA 3′-end Chronicles[J]. Journal of Molecular Biology, 2011, 412(5): 793-813.
[73] LI J, MASON SW, GREENBLATT J. Elongation factor NusG interacts with termination factor rho to regulate termination and antitermination of transcription[J]. Genes & Development, 1993, 7(1): 161-172.
[74] Di SALVO M, PUCCIO S, PEANO C, LACOUR S, ALIFANO P. RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on *Escherichia coli*, *Bacillus subtilis* and *Salmonella enterica* databases[J]. BMC Bioinformatics, 2019, 20(1): 117.
[75] SALVO MD, PINATEL EM, BELLIS GD, TALÀ A, PEANO C, ALIFANO P. P&TIT: a computer tool for predicting prototypical transcription promoter and terminator elements by conserved motifs[C]//2017 International Conference on Bioinformatics and Biomedicine. October 26, 2017.
[76] NAIR TM, TAMBE SS, KULKARNI BD. Application of artificial neural networks for prokaryotic transcription terminator prediction[J]. FEBS Letters, 1994, 346(2-3): 273-277.
[77] PENNACCHIO LA, BICKMORE W, DEAN A, NOBREGA MA, BEJERANO G. Enhancers: five essential questions[J]. Nature Reviews Genetics, 2013, 14(4): 288-295.
[78] RAY-JONES H, SPIVAKOV M. Transcriptional enhancers and their communication with gene promoters[J]. Cellular and Molecular Life Sciences, 2021, 78(19-20): 6453-6485.
[79] ERNST J, KELLIS M. ChromHMM: automating chromatin-state discovery and characterization[J]. Nature Methods, 2012, 9(3): 215-216.
[80] HOFFMAN MM, BUSKE OJ, WANG J, WENG ZP, BILMES JA, NOBLE WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation[J]. Nature Methods, 2012, 9(5): 473-476.
[81] FIRPI HA, UCAR D, TAN K. Discover regulatory DNA elements using chromatin signatures and artificial neural network[J]. Bioinformatics, 2010, 26(13): 1579-1586.
[82] FERNÁNDEZ M, MIRANDA-SAAVEDRA D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines[J]. Nucleic Acids Research, 2012, 40(10): e77.
[83] RAJAGOPAL N, XIE W, LI Y, WAGNER U, WANG W, STAMATOYANNOPOULOS J, ERNST J, KELLIS M, REN B. RFECS: a random-forest based algorithm for enhancer identification from chromatin state[J]. PLoS Computational Biology, 2013, 9(3): e1002968.
[84] 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.
[85] MIR BA, REHMAN MU, TAYARA H, CHONG KT. Improving enhancer identification with a multi-classifier stacked ensemble model[J]. Journal of Molecular Biology, 2023, 435(23): 168314.
[86] LIU YH, WANG ZX, YUAN H, ZHU GQ, ZHANG YQ. HEAP: a task adaptive-based explainable deep learning framework for enhancer activity prediction[J]. Briefings in Bioinformatics, 2023, 24(5): bbad286.
[87] GARNER MM, REVZIN A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system[J]. Nucleic Acids Research, 1981, 9(13): 3047-3060.
[88] VIERSTRA J, STAMATOYANNOPOULOS JA. Genomic footprinting[J]. Nature Methods, 2016, 13(3): 213-221.
[89] DAS PM, RAMACHANDRAN K, VANWERT J, SINGAL R. Chromatin immunoprecipitation assay[J]. BioTechniques, 2004, 37(6): 961-969.
[90] LUO KX, HARTEMINK AJ. Using DNase digestion data to accurately identify transcription factor binding sites[J]. Pacific Symposium on Biocomputing, 2013: 80-91.

[91] SHERWOOD RI, HASHIMOTO T, O'DONNELL CW, LEWIS S, BARKAL AA, van HOFF JP, KARUN V, JAAKKOLA T, GIFFORD DK. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape[J]. Nature Biotechnology, 2014, 32(2): 171-178.
[92] PIQUE-REGI R, DEGNER JF, PAI AA, GAFFNEY DJ, GILAD Y, PRITCHARD JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data[J]. Genome Research, 2011, 21(3): 447-455.
[93] 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.
[94] ZHOU TY, SHEN N, YANG L, ABE N, HORTON J, MANN RS, BUSSEMAKER HJ, GORDÂN R, ROHS R. Quantitative modeling of transcription factor binding specificities using DNA shape[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(15): 4654-4659.
[95] AVSEC Ž, WEILERT M, SHRIKUMAR A, KRUEGER S, ALEXANDARI A, DALAL K, FROPF R, McANANY C, GAGNEUR J, KUNDAJE A, ZEITLINGER J. Base-resolution models of transcription-factor binding reveal soft motif syntax[J]. Nature Genetics, 2021, 53(3): 354-366.
[96] HAN K, SHEN LC, ZHU YH, XU J, SONG JN, YU DJ. MAResNet: predicting transcription factor binding sites by combining multi-scale bottom-up and top-down attention and residual network[J]. Briefings in Bioinformatics, 2022, 23(1): bbab445.
[97] DENG L, WU H, LIU XJ, LIU H. DeepD2V: a novel deep learning-based framework for predicting transcription factor binding sites from combined DNA sequence[J]. International Journal of Molecular Sciences, 2021, 22(11): 5521.
[98] PREMKUMAR KAR, BHARANIKUMAR R, PALANIAPPAN A. Riboflow: using deep learning to classify riboswitches with ~99% accuracy[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 808.
[99] ANGENENT-MARI NM, GARRUSS AS, SOENKSEN LR, CHURCH G, COLLINS JJ. A deep learning approach to programmable RNA switches[J]. Nature Communications, 2020, 11(1): 5057.
[100] XIANG YZ, HUANG WZ, TAN LM, CHEN TY, HE Y, IRVING PS, WEEKS KM, ZHANG QC, DONG XN. Pervasive downstream RNA hairpins dynamically dictate start-codon selection[J]. Nature, 2023, 621(7978): 423-430.
[101] 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(1): 5058.
[102] JING RY, LI YZ, XUE L, LIU FJ, LI ML, LUO JS. autoBioSeqpy: a deep learning tool for the classification of biological sequences[J]. Journal of Chemical Information and Modeling, 2020, 60(8): 3755-3764.
[103] NOSHAY JM, WALKER T, ALEXANDER WG, KLINGEMAN DM, ROMERO J, WALKER AM, PRATES E, ECKERT C, IRLE S, KAINER D, JACOBSON DA. Quantum biological insights into CRISPR-Cas9 sgRNA efficiency from explainable-AI driven feature engineering[J]. Nucleic Acids Research, 2023, 51(19): 10147-10161.
[104] YANG TP, WANG Y, HE YH. TEC-miTarget: enhancing microRNA target prediction based on deep learning of ribonucleic acid sequences[J]. BMC Bioinformatics, 2024, 25(1): 159.
[105] MORAGA C, SANCHEZ E, FERRARINI MG, GUTIERREZ RA, VIDAL EA, SAGOT MF. BrumiR: a toolkit for *de novo* discovery of microRNAs from sRNA-seq data[J]. GigaScience, 2022, 11: giac093.
[106] LADUNGA I. More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature[J]. Nucleic Acids Research, 2007, 35(2): 433-440.
[107] ROSA ML, FIANNACA A, PAGLIA LL, URSO A. A graph neural network approach for the analysis of siRNA-target biological networks[J]. International Journal of Molecular Sciences, 2022, 23(22): 14211.
[108] XU JY, XU N, XIE WX, ZHAO CK, YU L, FENG WX. BERT-siRNA: siRNA target prediction based on BERT pre-trained interpretable model[J]. Gene, 2024, 910: 148330.
[109] LASANTHA D, VIDANAGAMACHCHI S, NALLAPERUMA S. CRIECNN: Ensemble convolutional neural network and advanced feature extraction methods for the precise forecasting of circRNA-RBP binding sites[J]. Computers in Biology and Medicine, 2024, 174: 108466.
[110] SERGANOV A, NUDLER E. A decade of riboswitches[J]. Cell, 2013, 152(1/2): 17-24.
[111] GUPTA D, BHATTACHARJEE O, MANDAL D, SEN MK, DEY D, DASGUPTA A, KAZI TA, GUPTA R, SINHAROY S, ACHARYA K, CHATTOPADHYAY D, RAVICHANDIRAN V, ROY S, GHOSH D. CRISPR-Cas9 system: a new-fangled dawn in gene editing[J]. Life Sciences, 2019, 232: 116636.
[112] JANIK E, NIEMCEWICZ M, CEREMUGA M, KRZOWSKI L, SALUK-BIJAK J, BIJAK M. Various aspects of a gene editing system-CRISPR-Cas9[J]. International Journal of Molecular Sciences, 2020, 21(24): 9604.
[113] MORENO-MATEOS MA, VEJNAR CE, BEAUDOIN JD, FERNANDEZ JP, MIS EK, KHOKHA MK, GIRALDEZ AJ. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*[J]. Nature Methods, 2015, 12(10): 982-988.
[114] DHANJAL JK, RADHAKRISHNAN N, SUNDAR D. CRISPcut: a novel tool for designing optimal sgRNAs for CRISPR/Cas9 based experiments in human cells[J]. Genomics, 2019, 111(4): 560-566.
[115] KIM N, CHOI S, KIM S, SONG M, SEO JH, MIN S, PARK J, CHO SR, KIM HH. Deep learning models to predict the editing efficiencies and outcomes of diverse base editors[J]. Nature Biotechnology, 2024, 42:

484-497.
[116] LIU Y, FAN R, YI JK, CUI QH, CUI CM. A fusion framework of deep learning and machine learning for predicting sgRNA cleavage efficiency[J]. Computers in Biology and Medicine, 2023, 165: 107476.
[117] WANG DQ, ZHANG CD, WANG B, LI B, WANG Q, LIU D, WANG HY, ZHOU Y, SHI LM, LAN F, WANG YM. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning[J]. Nature Communications, 2019, 10(1): 4284.
[118] CHUAI GH, MA HH, YAN JF, CHEN M, HONG NF, XUE DY, ZHOU C, ZHU CY, CHEN K, DUAN B, GU F, QU S, HUANG DS, WEI J, LIU Q. DeepCRISPR: optimized CRISPR guide RNA design by deep learning[J]. Genome Biology, 2018, 19(1): 80.
[119] HAEUSSLER M, SCHÖNIG K, ECKERT H, ESCHSTRUTH A, MIANNÉ J, RENAUD JB, SCHNEIDER-MAUNOURY S, SHKUMATAVA A, TEBOUL L, KENT J, JOLY JS, CONCORDET JP. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR[J]. Genome Biology, 2016, 17(1): 148.
[120] CHUAI GH, WANG QL, LIU Q. *In silico* meets *in vivo*: towards computational CRISPR-based sgRNA design[J]. Trends in Biotechnology, 2017, 35(1): 12-21.
[121] ZHANG GS, LUO Y, DAI XH, DAI ZM. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities[J]. Briefings in Bioinformatics, 2023, 24(6): bbad333.
[122] RAHMAN MK, RAHMAN MS. CRISPRpred: a flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems[J]. PLoS One, 2017, 12(8): e0181943.
[123] CHARI R, YEO NC, CHAVEZ A, CHURCH GM. sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity[J]. ACS Synthetic Biology, 2017, 6(5): 902-904.
[124] BAISYA D, RAMESH A, SCHWARTZ C, LONARDI S, WHEELDON I. Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in *Yarrowia lipolytica*[J]. Nature Communications, 2022, 13(1): 922.
[125] WANG L, ZHANG JH. Prediction of sgRNA on-target activity in bacteria by deep learning[J]. BMC Bioinformatics, 2019, 20(1): 517.
[126] PENG D, TARLETON R. EuPaGDT: a web tool tailored to design CRISPR guide RNAs for eukaryotic pathogens[J]. Microbial Genomics, 2015, 1(4): e000033.
[127] LIU H, DING YD, ZHOU YQ, JIN WQ, XIE KB, CHEN LL. CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants[J]. Molecular Plant, 2017, 10(3): 530-532.
[128] AHMED F, KAUNDAL R, RAGHAVA GPS. PHDcleav: a SVM based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors[J]. BMC Bioinformatics, 2013, 14(Suppl 14): S9.
[129] BAO Y, HAYASHIDA M, AKUTSU T. LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length[J]. BMC Bioinformatics, 2016, 17(1): 487.
[130] LIU PY, SONG JN, LIN CY, AKUTSU T. ReCGBM: a gradient boosting-based method for predicting human dicer cleavage sites[J]. BMC Bioinformatics, 2021, 22(1): 63.
[131] BELL J, HENDRIX DA. Predicting drosha and dicer cleavage sites with DeepMirCut[J]. Frontiers in Molecular Biosciences, 2022, 8: 799056.
[132] MU LX, SONG JN, AKUTSU T, MORI T. DiCleave: a deep learning model for predicting human Dicer cleavage sites[J]. BMC Bioinformatics, 2024, 25(1): 13.
[133] HACKENBERG M, RODRÍGUEZ-EZPELETA N, ARANSAY AM. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments[J]. Nucleic Acids Research, 2011, 39(Web Server issue): W132-W138.
[134] VITSIOS DM, KENTEPOZIDOU E, QUINTAIS L, BENITO-GUTIÉRREZ E, van DONGEN S, DAVIS MP, ENRIGHT AJ. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests[J]. Nucleic Acids Research, 2017, 45(21): e177.
[135] GU TJ, XIE MY, BRAD BARBAZUK W, LEE JH. Biological features between miRNAs and their targets are unveiled from deep learning models[J]. Scientific Reports, 2021, 11(1): 23825.
[136] PLA A, ZHONG XF, RAYNER S. miRAW: a deep learning-based approach to predict microRNA targets by analyzing whole microRNA transcripts[J]. PLoS Computational Biology, 2018, 14(7): e1006185.
[137] MIN S, LEE B, YOON S. TargetNet: functional microRNA target prediction with deep neural networks[J]. Bioinformatics, 2022, 38(3): 671-677.
[138] YAN C, DUAN GH, LI N, ZHANG LS, WU FX, WANG JX. PDMDA: predicting deep-level miRNA-disease associations with graph neural networks and sequence features[J]. Bioinformatics, 2022, 38(8): 2226-2234.
[139] DWEEP H, GRETZ N, STICHT C. miRWalk database for miRNA-target interactions[J]. Methods in Molecular Biology, 2014, 1182: 289-305.
[140] GRIFFITHS-JONES S, GROCOCK RJ, van DONGEN S, BATEMAN A, ENRIGHT AJ. miRBase: microRNA sequences, targets and gene nomenclature[J]. Nucleic Acids Research, 2006, 34(Database issue): D140-D144.
[141] XU P, LI XB, LIANG YJ, BAO ZS, ZHANG FY, GU LL, KOSARI S, LIU WB. PmiRtarbase: a positive miRNA-target regulations database[J]. Computational Biology and Chemistry, 2022, 98: 107690.
[142] CHEN YH, WANG XW. miRDB: an online database for prediction of functional microRNA targets[J]. Nucleic Acids Research, 2020, 48(D1): D127-D131.
[143] RYVKIN P, LEUNG YY, UNGAR LH, GREGORY BD, WANG LS. Using machine learning and

high-throughput RNA sequencing to classify the precursors of small non-coding RNAs[J]. Methods, 2014, 67(1): 28-35.
[144] ZHAO CK, XU N, TAN JW, CHENG Q, XIE WX, XU JY, WEI ZY, YE J, YU L, FENG WX. ILGBMSH: an interpretable classification model for the shRNA target prediction with ensemble learning algorithm[J]. Briefings in Bioinformatics, 2022, 23(6): bbac429.
[145] MYSARA M, ELHEFNAWI M, GARIBALDI JM. MysiRNA: improving siRNA efficacy prediction using a machine-learning model combining multi-tools and whole stacking energy (ΔG)[J]. Journal of Biomedical Informatics, 2012, 45(3): 528-534.
[146] LIU TY, HUANG JY, LUO DL, REN LP, NING L, HUANG J, LIN H, ZHANG Y. Cm-siRPred: Predicting chemically modified siRNA efficiency based on multi-view learning strategy[J]. International Journal of Biological Macromolecules, 2024, 264: 130638.
[147] WANG LJ, HUANG CY, YANG JY. Predicting siRNA potency with random forests and support vector machines[J]. BMC Genomics, 2010, 11(Suppl 3): S2.
[148] KLINGELHOEFER JW, MOUTSIANAS L, HOLMES C. Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency[J]. Bioinformatics, 2009, 25(13): 1594-1601.
[149] HUESKEN D, LANGE J, MICKANIN C, WEILER J, ASSELBERGS F, WARNER J, MELOON B, ENGEL S, ROSENBERG A, COHEN D, LABOW M, REINHARDT M, NATT F, HALL J. Design of a genome-wide siRNA library using an artificial neural network[J]. Nature Biotechnology, 2005, 23(8): 995-1001.
[150] MOHAMED SK, NOUNU A, NOVÁČEK V. Biological applications of knowledge graph embedding models[J]. Briefings in Bioinformatics, 2021, 22(2): 1679-1693.
[151] VERT JP, FOVEAU N, LAJAUNIE C, VANDENBROUCK Y. An accurate and interpretable model for siRNA efficacy prediction[J]. BMC Bioinformatics, 2006, 7: 520.
[152] HAN Y, HE F, CHEN YB, LIU YN, YU HL. SiRNA silencing efficacy prediction based on a deep architecture[J]. BMC Genomics, 2018, 19(Suppl 7): 669.
[153] HAN Y, HE F, TAN X, YU HL. Effective small interfering RNA design based on convolutional neural network[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). November 13-16, 2017, Kansas City, MO. IEEE, 2017: 16-21.
[154] COSTELLO A, LAO NT, BARRON N, CLYNES M. Continuous translation of circularized mRNA improves recombinant protein titer[J]. Metabolic Engineering, 2019, 52: 284-292.
[155] LIU X, ZHANG Y, ZHOU SR, DAIN L, MEI L, ZHU GZ. Circular RNA: an emerging frontier in RNA therapeutic targets, RNA therapeutics, and mRNA vaccines[J]. Journal of Controlled Release, 2022, 348: 84-94.
[156] NIU D, WU YR, LIAN JQ. Circular RNA vaccine in disease prevention and treatment[J]. Signal Transduction and Targeted Therapy, 2023, 8(1): 341.
[157] LIU L, WANG PJ, ZHAO DD, ZHU L, TANG JL, LENG WC, SU JC, LIU Y, BI CH, ZHANG XL. Engineering circularized mRNAs for the production of spider silk proteins[J]. Applied and Environmental Microbiology, 2022, 88(8): e0002822.
[158] SHARMA NK, DWIVEDI P, BHUSHAN R, MAURYA PK, KUMAR A, DAKAL TC. Engineering circular RNA for molecular and metabolic reprogramming[J]. Functional & Integrative Genomics, 2024, 24(4): 117.
[159] CHAABANE M, WILLIAMS RM, STEPHENS AT, PARK JW. circDeep: deep learning approach for circular RNA classification from other long non-coding RNA[J]. Bioinformatics, 2020, 36(1): 73-80.
[160] PAN XY, XIONG K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features[J]. Molecular BioSystems, 2015, 11(8): 2219-2226.
[161] PAN XY, XIONG K, ANTHON C, HYTTEL P, FREUDE KK, JENSEN LJ, GORODKIN J. WebCircRNA: classifying the circular RNA potential of coding and noncoding RNA[J]. Genes, 2018, 9(11): 536.
[162] NIU MT, ZHANG J, LI YJ, WANG CK, LIU ZQ, DING H, ZOU Q, MA Q. CirRNAPL: a web server for the identification of circRNA based on extreme learning machine[J]. Computational and Structural Biotechnology Journal, 2020, 18: 834-842.
[163] WANG J, WANG LJ. Deep learning of the back-splicing code for circular RNA formation[J]. Bioinformatics, 2019, 35(24): 5235-5242.
[164] CZAJKA JJ, OYETUNDE T, TANG YJ. Integrated knowledge mining, genome-scale modeling, and machine learning for predicting *Yarrowia lipolytica* bioproduction[J]. Metabolic Engineering, 2021, 67: 227-236.
[165] DU F, LI ZJ, LI X, ZHANG DD, ZHANG F, ZHANG ZX, XU YS, TANG J, LI YQ, HUANG XX, GU Y, SUN XM, HUANG H. Optimizing multicopy chromosomal integration for stable high-performing strains[J]. Nature Chemical Biology, 2024, 20(12): 1670-1679.
[166] 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.
[167] KHAMWACHIRAPITHAK P, SAE-TANG K, MHUANTONG W, TANAPONGPIPAT S, ZHAO XQ, LIU CG, WEI DQ, CHAMPREDA V, RUNGUPHAN W. Optimizing ethanol production in *Saccharomyces cerevisiae* at ambient and elevated temperatures through machine learning-guided combinatorial promoter modifications[J]. ACS Synthetic Biology, 2023, 12(10): 2897-2908.
[168] 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.

[169] MORENO-PAZ S, van der HOEK R, ELIANA E, ZWARTJENS P, GOSIEWSKA S, MARTINS dos SANTOS VAP, SCHMITZ J, SUAREZ-DIEZ M. Machine learning-guided optimization of *p*-coumaric acid production in yeast[J]. ACS Synthetic Biology, 2024, 13(4): 1312-1322.

[170] SHIMAZAKI S, YAMADA R, YAMAMOTO Y, MATSUMOTO T, OGINO H. Building a machine-learning model to predict optimal mevalonate pathway gene expression levels for efficient production of a carotenoid in yeast[J]. Biotechnology Journal, 2024, 19(1): e2300285.

[171] MUKHERJEE M, BLAIR RH, WANG ZQ. Machine-learning guided elucidation of contribution of individual steps in the mevalonate pathway and construction of a yeast platform strain for terpenoid production[J]. Metabolic Engineering, 2022, 74: 139-149.

[172] ZHANG J, PETERSEN SD, RADIVOJEVIC T, RAMIREZ A, PÉREZ-MANRÍQUEZ A, ABELIUK E, SÁNCHEZ BJ, COSTELLO Z, CHEN Y, FERO MJ, MARTIN HG, NIELSEN J, KEASLING JD, JENSEN MK. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism[J]. Nature Communications, 2020, 11(1): 4880.

[173] HANKE P, PARRELLO B, VASIEVA O, AKINS C, CHLENSKI P, BABNIGG G, HENRY C, FOFLONKER F, BRETTIN T, ANTONOPOULOS D, STEVENS R, FONSTEIN M. Engineering of increased L-threonine production in bacteria by combinatorial cloning and machine learning[J]. Metabolic Engineering Communications, 2023, 17: e00225.

[174] LEE ME, ASWANI A, HAN AS, TOMLIN CJ, DUEBER JE. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay[J]. Nucleic Acids Research, 2013, 41(22): 10668-10678.

[175] 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.

[176] KANG CK, SHIN J, CHA Y, KIM MS, CHOI MS, KIM T, PARK YK, CHOI YJ. Machine learning-guided prediction of potential engineering targets for microbial production of lycopene[J]. Bioresource Technology, 2023, 369: 128455.

[177] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[178] NIKOLADOS EM, WONGPROMMOON A, AODHA OM, CAMBRAY G, OYARZÚN DA. Accuracy and data efficiency in deep learning models of protein expression[J]. Nature Communications, 2022, 13(1): 7755.

[179] van LENT P, SCHMITZ J, ABEEL T. Simulated design-build-test-learn cycles for consistent comparison of machine learning methods in metabolic engineering[J]. ACS Synthetic Biology, 2023, 12(9): 2588-2599.