·人工细胞智能设计再造·

马红武 现任中国科学院天津工业生物技术研究所研究员、生物设计中心主任。2001 年毕业于天津大学化工学院，获生物化工博士学位，曾先后在德国生物技术中心(GBF)和英国爱丁堡大学信息学院从事研究工作，2011 年底回国。主要研究方向包括代谢网络分析和途径设计、代谢工程及合成生物学、计算生物学软件工具开发、人工智能生物功能预测和设计等，发表 SCI 论文 100 余篇，引用 5 000 余次。曾获科睿唯安“全球高被引科学家奖”和天津市自然科学奖特等奖。

蛋白表达系统的机理模型和人工智能模型研究进展

杨毅[1,2]，杜军[1,2]，杨春贺[1,2]，马红武[1,2*]

1 中国科学院天津工业生物技术研究所 低碳合成工程生物学(全国)重点实验室 生物设计中心，天津 300308
2 国家合成生物技术创新中心，天津 300308

杨毅，杜军，杨春贺，马红武. 蛋白表达系统的机理模型和人工智能模型研究进展[J]. 生物工程学报，2025，41(3): 1079-1097.
YANG Yi, DU Jun, YANG Chunhe, MA Hongwu. Research progress in mechanism models and artificial intelligence models for protein expression systems[J]. Chinese Journal of Biotechnology, 2025, 41(3): 1079-1097.

摘 要：蛋白质是生命活动的基础，研究蛋白表达机制对于揭示细胞组织规律与促进生物技术发展至关重要。蛋白质表达是一个涵盖转录、翻译、折叠、转运与翻译后修饰等精密调控的复杂过程，结合蛋白表达数据构建其模型对理解蛋白表达的各种细胞因素和调控机制具有重要意义。本文重点评述了近年来蛋白表达过程机理模型构建和通过人工智能方法分析各种因素对蛋白表达的影响。化学反应网络模型可从转录翻译的底层过程对蛋白表达进行数学建模，可分析各种胞内成分如聚合酶、tRNA 等对蛋白表达的影响，但模型参数数量巨大，难以直接实验确定，参数拟合是一个需要解决的难题。与之相对，数据驱动的人工智能模型主要研究目标蛋白的氨基酸序列和相应基因及调控区核苷酸序列对蛋白表达的影响，进而指导通过序列设计提高蛋白表达量。将机理模型和人工智能模型相结合，综合考虑胞内因素和表达序列特征的影响，有望进一步加深对蛋白表达系统的理解，为高价值目标蛋白的高效表达和细胞中不同蛋白的协调表达调控提供理论和技术支持。

关键词：蛋白表达系统；化学反应网络；人工智能；深度学习；转录；翻译

资助项目：中国科学院战略性先导科技专项(XDB0480000)
This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0480000).
*Corresponding author. E-mail: ma_hw@tib.cas.cn
Received: 2024-11-26; Accepted: 2025-03-06; Published online: 2025-03-07

Research progress in mechanism models and artificial intelligence models for protein expression systems

YANG Yi[1,2], DU Jun[1,2], YANG Chunhe[1,2], MA Hongwu[1,2*]

1 Biodesign Center, Key Laboratory of Engineering Biology for Low-Carbon Manufacturing, 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: Proteins are the basic building blocks of life. Studying the protein expression mechanism is essential for understanding the cellular organization principles and the development of biotechnology. Protein expression, involving transcription, translation, folding, and post-translational modification, is a complicatedly regulated process affected by various cellular components and sequence features of the expressed protein. Establishing protein expression models based on expression data is of great significance for probing into the regulatory factors and mechanisms of protein expression. Here we review the recent research progress in the mechanism models for quantitatively simulating the protein expression process and the prediction algorithms based on artificial intelligence for analyzing the regulatory factors. Chemical reaction network models have been developed to mathematically describe the elementary processes in protein expression and simulate the influences of various cellular components such as RNA polymerase and tRNA. However, the experimental determination of the huge number of model parameters is a big challenge. The main objective of data-driven AI models is to study the effects of protein/DNA sequences of the target protein on its expression, and subsequently optimize the sequences to improve protein expression. Methods combining mechanism models and AI models have the potential to deepen our understanding of protein expression processes, providing theoretical and technical support for the efficient production of high-value proteins and coordinate the regulation of different proteins.

Keywords: protein expression system; chemical reaction network; artificial intelligence; deep learning; transcription; translation

蛋白质是生命系统的基本组件，是生物体执行各种功能的关键分子，不仅参与细胞结构的构建，还在催化代谢反应、调节信号通路和维持细胞生理过程中发挥核心作用。由 DNA 转录得到 mRNA 再通过翻译生成蛋白质这一过程是生命系统的中心法则，因此研究蛋白表达系统对揭示生命活动的基本规律具有重要意义。同时，蛋白质表达技术在工业酶制剂、生物医药、生物材料、食用饲用蛋白和生物修复等领域也被广泛应用(图 1)。重组蛋白药物、疫苗以及工业酶的生产均需要在特定底盘细胞中大量表达目标蛋白，食用或饲用单细胞蛋白的生产需要提升菌体中的蛋白含量，因此蛋白表达系统也是现代生物技术领域的重要研究课题。此外，合成生物学和代谢工程研究中常需要对细胞中不同蛋白(酶)的表达量进行精确调控以控制胞内代谢流分布，从而得到更多目标产品。虽然通过启动子、核糖体结合位点(ribosome

binding site, RBS)等调控元件和一些简单基因调控回路，研究人员已经可以对蛋白表达量进行调节[1]，但由于蛋白表达过程中很多过程机理仍然未知，这种调节很难达到可精确定量预测控制的水平。

蛋白表达是生物体内一个复杂而精细的过程，而且在原核和真核生物中有较大差异。因此常需要根据目标蛋白的特征选择合适的底盘细胞进行表达。目前常用的原核表达系统包括大肠杆菌(*Escherichia coli*)[2]、枯草芽孢杆菌(*Bacillus subtilis*)[3]等，其具有生长快速、表达效率高、成本低等优点，但缺乏蛋白后翻译修饰过程，容易形成包涵体而无法得到具有正常生理功能的纯化蛋白。常用的真核生物表达系统有毕赤酵母(*Pichia pastoris*)[4]、酿酒酵母(*Saccharomyces cerevisiae*)[5]、昆虫细胞及哺乳动物细胞[6]等；其更适于表达具有复杂结构的高分子量蛋白，特别是重组蛋白药物和疫苗的生产只能使用特定的哺乳动物细胞如HEK293、中国仓鼠卵巢细胞(Chinese hamster ovary cells, CHO)等；其缺点是表达量较低、培养条件复杂、表达稳定性差、培养成本高。

除了上述基于细胞的蛋白表达系统外，近年来研究人员还开发了无细胞蛋白表达系统(图1)，其使用细胞裂解液中的转录、翻译系统直接进行蛋白表达[7]。无细胞蛋白表达系统的优点在于不受细胞生长周期、宿主毒性或蛋白质折叠问题的限制，使得难以表达的蛋白质(如膜蛋白、毒性蛋白、不稳定蛋白)得以有效生产[8]。此外，该系统还允许进行更高效的蛋白质工程操作，例如快速突变[9]、标记氨基酸[10]以及通过共转录翻译表达多种蛋白[11-12]。由于无需进行复杂的细胞培养，无细胞系统能够在数小时内完成蛋白表达，大大缩短了实验周期[13-14]。除了基于细胞裂解液的无细胞蛋白表达系统外，由纯化的重组蛋白因子和核糖体组成的高度简化的重组元件蛋白合成(protein synthesis using recombinant elements, PURE)系统也被广泛应用[15-16](图1)。与细胞裂解液系统相比，PURE系统成分明确并且排除了细胞提取物中的干扰成分，因此更加精确和可控，非常适合作为蛋白表达系统的过程机理的研究对象[17-19]。此外该系统不含蛋白酶、核酸酶等潜在的降解因子，因此特别适合生产高纯度的功能蛋白[20]。

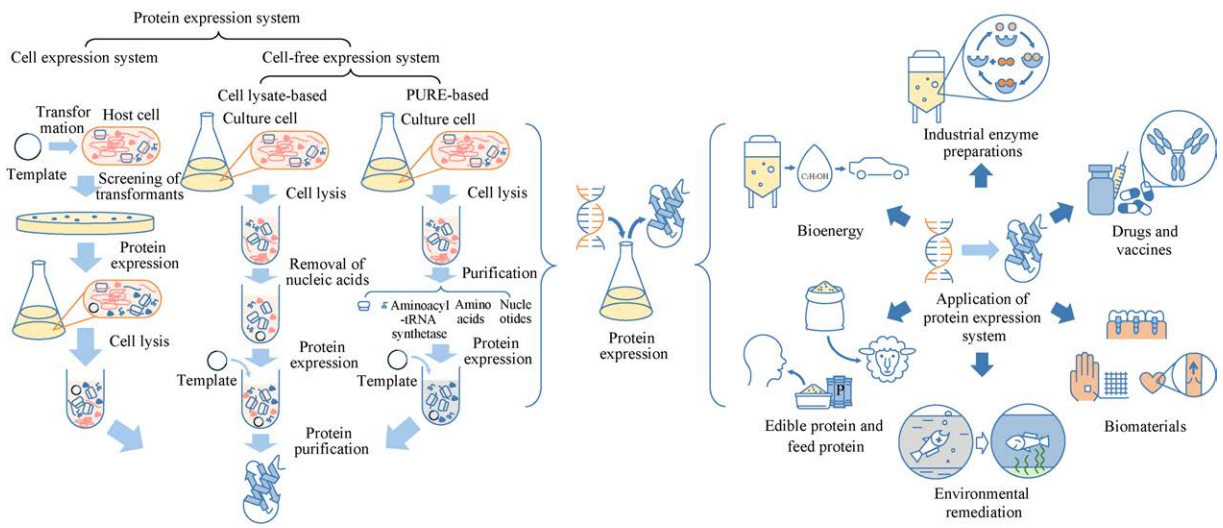

图1　蛋白表达系统的类型和应用

Figure 1　Types and applications of protein expression systems.

定量解析蛋白表达过程的调控机制并由此构建蛋白表达系统的数学模型对实现可预测可调控的蛋白表达非常必要。在基因组规模细胞模型构建方面，研究人员开发了 ME 模型(models of metabolism and macromolecular expression)，尝试将代谢模型与细胞中的大分子表达模型整合[21-23]，但其仍基于计量学模型，可以模拟细胞中的通量变化但无法模拟蛋白和代谢物浓度的变化。在此基础上研究人员开发了全细胞模型(whole cell model)，其整合了细胞中的转录和翻译过程，可以模拟在不同条件下细胞内不同蛋白表达量的变化[24-25]。另一方面，研究人员针对不同基因线路构建了动力学模型并模拟其中蛋白表达量随时间的变化，以解释其复杂动态调控行为，如振荡、开关、逻辑门、模式识别等[26-29]。但基因回路中对蛋白表达的模拟更侧重于蛋白表达量与基因线路整体动力学的关系，因此对蛋白表达的模拟是相当粗略的，往往不考虑或只关注少数因素对蛋白表达的影响，而在真实细胞中影响因素更为复杂。由于蛋白表达的影响因素繁多、机理复杂且存在很多未知情况，构建全面准确的机理模型将是一项困难的长期工作。

数据驱动的人工智能机器学习预测方法为蛋白表达研究提供了新的思路。通过精巧的实验设计和高通量分析方法，研究人员可以有针对性地产生大量蛋白表达实验数据，利用这些数据训练人工智能模型可预测各种因素对蛋白表达的影响。人工智能(artificial intelligence, AI)算法还可以预测蛋白质序列的突变对其表达量或稳定性的影响，这对于工业应用中的大规模蛋白生产尤为重要。

机理模型与人工智能模型为蛋白表达研究提供了重要工具，但二者在方法论和应用场景上的核心差异与互补性尚未被系统阐释。机理模型通过模拟转录、翻译等底层生化反应，能够解析 tRNA 丰度、能量代谢等胞内组分对表达量的影响，但只考虑了分子复合体组织、解离以及相关的代谢层面的因素，未考虑蛋白编码和调控序列层面特征的影响。相比之下，人工智能模型通过挖掘序列特征与表达量间的非线性关系，可预测编码序列或调控序列的优化方向，然而，由于这些模型未考虑底盘系统的影响，只能基于特定底盘的数据进行训练，并仅适用于该底盘，因此通用性较差且难以揭示生物学机制。在此背景下，机理模型与人工智能模型的结合为蛋白表达预测提供了新的思路，但目前缺乏成熟的方法。

本文的重点在于通过对蛋白表达的机理模型特别是化学反应网络(chemical reaction network, CRN)模型以及人工智能模型相关研究的评述，对二者在蛋白表达预测中的特点(如 CRN 模型在机制解析方面的优势和 AI 模型在序列优化方面的优势)进行分析和比较，并探讨二者协同建模的可能性。例如，基于 AI 模型的序列特征预测结果可以为机理模型添加序列层面的约束，基于机理层面影响因素相关数据的训练也有助于提高 AI 模型的通用性。本文旨在探讨通过整合二者优势推进蛋白表达系统的跨尺度建模方法，并为蛋白表达的理性优化提供方法论支持。

1 蛋白表达基本过程及化学反应网络模型描述

蛋白表达过程始于基因的转录，即 DNA 序列被转录成 mRNA 分子。随后，mRNA 与核糖体结合，开始蛋白质的翻译过程。在翻译过程中，核糖体按照 mRNA 上的遗传密码，依次将氨基酸连接成多肽链。最后经过一系列的蛋白质修饰和加工，形成具有生物活性的蛋白质分

子。早在 20 世纪 60 年代，研究人员就已经开始使用简单常微分方程(ordinary differential equation, ODE)模型来描述蛋白质表达的动力学过程，例如 Goodwin[30]构建的经典振荡器模型用 ODE 模拟具有负反馈的基因调控网络中的蛋白表达的振荡模式，为理解蛋白质表达中的反馈机制奠定了基础。这种粗粒化的蛋白表达的数学描述方法目前仍广泛应用于基因调控回路的动力学模型构建。但此类模型只反映蛋白对调控因子的响应，不包含对转录翻译的具体过程的描述，使其无法用于分析蛋白序列、tRNA 组成等对蛋白表达过程的影响。因此本文将聚焦于从更底层机理角度对蛋白表达过程进行数学描述的模型。

CRN 理论基于 20 世纪 70 年代早期 Feinberg[31]、Horn 等[32]和 Krambeck[33]的开创性工作逐渐发展起来。CRN 模型通过一系列相互关联的反应描述化学物质的相互转化，并通过较简单的数学方程描述各种影响因素对反应速率的影响，例如质量作用定律、Michaelis-Menten 方程或 Hill 方程，通过求解相关的常微分方程组实现对相关系统动态行为的模拟。常用的分析方法包括稳态分析[34]、稳定性和分岔分析[35]、敏感性分析[36]等。CRN 具有较高通用性，很多生物过程如酶促反应、基因表达调控、信号传导等均可以通过 CRN 模型表示[27,37-39]，也成为模拟和理解蛋白表达过程的重要方法[40-44]。作为所有组分定量化的无细胞蛋白质合成系统，PURE 系统可以通过调节组分方便地对反应条件进行微调，与组分复杂且批次间可重复性低的其他蛋白质合成体系相比，在生理机制研究和蛋白质工程方面具有很大的优势[45]，因此也成为蛋白表达 CRN 建模的重要对象[17,46]。为了构建蛋白表达全面准确的机理模型，需要考虑多种影响因素，整合相关的变量和参数构建细粒化的模型，如分析不同种类氨基酸的添加反应对蛋白合成的影响就需要模型中包含肽链延长过程中每个氨基酸的添加反应。接下来对面向 PURE 系统的细粒化 CRN 模型进行介绍。

2017 年，Shimizu 团队[47-48]基于质量作用动力学构建了一个细粒化的 PURE 翻译系统的 CRN 模型。该模型由 968 个反应和 241 个代谢物组成，涵盖了 PURE 系统所有组分参与的所有反应，可以实现对 fMGG 三肽(fMet-Gly-Gly)合成的模拟[47]。另外该模型还可以模拟所有组分的动力学过程，为模拟和探索 PURE 系统蛋白质合成提供了一种有价值的方法。但其建模对象为 fMGG 三肽，无法实现对其他蛋白的模拟。此外，由于没有对转录进行建模，该模型对 PURE 系统的描述不完整，无法进行实验验证。Murray 团队[49]在 Shimizu 模型的基础上整合了转录模型[43]，并将翻译模型中的氨酰-tRNA 合成和肽链延长子系统相关反应扩展至任意氨基酸，构建了面向任意蛋白表达的全局细粒化 PURE 系统 CRN 模型。该模型可以实现对任意蛋白表达的模拟，能够准确预测具有 T7 启动子和强 RBS 的 MGapt 和 deGFP 蛋白的表达，误差在 10%以内[49]。Murray 模型为模拟和分析多种细胞内因素，特别是编码区和调控区序列特征对蛋白表达的影响奠定了基础。接下来基于 Murray 模型对蛋白表达基本过程及其 CRN 模型描述进行介绍。

1.1 基因转录过程及相关化学反应网络模型描述

基因转录过程主要由 RNA 聚合酶驱动，可分为 3 个阶段：起始、延伸和终止。在起始阶段，RNA 聚合酶(RNA polymerase, RNAP)首先识别并结合到基因启动子区域，双链 DNA 在启动子处局部解开，使模板链暴露出来。延伸阶段，RNAP 将与模板互补的核糖核苷酸逐个连

接到 RNA 链上合成 RNA 分子。链延长直到遇到 DNA 上的终止信号，RNA 链从 DNA 模板上释放，转录过程结束[50]。原核生物中所有类型的 RNA(包括 mRNA、rRNA 和 tRNA)由同一种 RNAP 合成。而真核生物则有 3 种 RNAP，分别负责不同类型 RNA 的合成。此外，生成的初级 mRNA 还需经过进一步加工(如剪接、5′端加帽、3′端加多聚 A 尾等)后成为成熟的 mRNA，才能通过核孔转运到细胞质进行翻译[51]。由于真核生物转录过程涉及更多因素，更为复杂，目前 CRN 模型主要针对原核生物的转录过程展开。

Murray 模型中的转录模块参考了 Tuza 等[43]的工作，分为起始、延伸和终止 3 个部分，各部分分别包含相关分子间的结合、水解以及解离等反应，能够实现对转录全局的定性和定量模拟。由于模型中的核苷酸添加反应是相互独立的，因此可以通过解析相关的反应参数的变化对转录动力学的影响分析影响转录的编码区序列特征。此外，模型未考虑启动子、终止子以及其他调控序列的序列特征和空间结构、调控因子对转录效率的影响，也未考虑多个 RNAP 同时转录的情况，因此如需分析相关因素对转录的影响，需要在模型中添加相应的代谢物、反应和参数。

1.2 翻译过程及相关化学反应网络模型

与转录类似，以 mRNA 为模板的蛋白翻译过程也主要分为起始、延伸和终止 3 个阶段。但与转录相比，参与蛋白翻译过程的蛋白、tRNA 等大分子更多，机制更为复杂。翻译过程从 mRNA 与核糖体小亚基结合开始，随后 mRNA 上的起始密码子(通常为 AUG)被起始 tRNA 识别，有多种起始因子蛋白(如大肠杆菌中的 IF1、IF2、IF3)辅助 tRNA 和 mRNA 的结合。起始氨基酸在原核生物中是甲酰甲硫氨酸(fMet)，而真核生物中则是甲硫氨酸(Met)[52]。在肽链延伸阶段，由大小亚基结合形成的核糖体在延伸因子(EF-Tu 和 EF-G)的帮助下逐步沿着 mRNA 移动，每次读取 1 个密码子。肽链的延伸原料为携带特定氨基酸的 tRNA 即氨酰-tRNA，其合成由氨酰-tRNA 合成酶催化。特定的氨酰-tRNA 进入核糖体氨酰位(aminoacyl site, A site)，识别 mRNA 上的密码子，新的氨基酸连接到多肽链上后 tRNA 移至核糖体肽酰位(peptidyl site, P site)，将 A 位空出准备接受新的 tRNA[53-54]。当核糖体遇到终止密码子(UAA、UAG 或 UGA)时，释放因子(如 RF-1、RF-2)进入 A 位，促使新合成的多肽链从核糖体释放出来[55]。

Murray 模型中的翻译模型分为氨酰-tRNA 合成、翻译起始、翻译延伸、翻译终止以及能量代谢 5 个子模块，各模块均包含了相关生物学过程的几乎全部细节化的反应(表 1)，能够实现对翻译全局的定性和定量模拟。下面以氨酰-tRNA 合成模块为例，对 CRN 模型的构建和模拟的范式进行简要介绍。

在氨酰-tRNA 合成过程中，氨酰-tRNA 合成酶(aminoacyl-tRNA synthetase, aaRS)通过一系列生化反应将特定氨基酸与其对应的 tRNA 分子共价连接，形成负载氨基酸的 tRNA (氨酰-tRNA)。例如，aaRS 首先与氨基酸结合形成复合体，随后 ATP 水解提供能量将氨基酸活化，最后活化的氨基酸被转移至 tRNA 的 3′端，完成负载(图 2A)。构建 CRN 模型时需要将此过程分解为各个独立的反应，并明确参与反应的代谢物及其计量学关系(图 2B)。然后基于质量作用定律假设，考虑某代谢物的所有生成和消耗方式，构建动力学方程。再通过反应动力学实验确定动力学参数(反应速率常数)。包含全部代谢物的动力学方程组能够描述 CRN 系统的动力学演化(图 2C)。给定初始条件就可以借助

CRN 模拟工具(如 BioCRNpyler[56]、bioscrape[57])对 CRN 中各代谢物的动力学进行模拟(图 2D),进而可以分析不同变量(如 tRNA 和氨酰-tRNA 合成酶丰度)和参数(如氨酰-tRNA 合成酶结合氨基酸的速率)的变化对的蛋白表达的影响。例如,在 PURE 系统中,通过调整模型中的 tRNA 补充策略,可优化低频密码子的翻译效率。

尽管 CRN 模型能够从分子机制层面解析蛋白表达过程,其实际应用仍面临显著挑战。首先,现有 CRN 模型的规模庞大,例如 Shimizu 模型包含 968 个反应[47],Murray 模型进一步扩展至数千个反应[49](表 1)。这些模型的预测准确性取决于大量动力学参数值,而此类参数测定的实验通常难以高通量实现。其次,当前 CRN 模型对编码区和调控区的序列特征(如密码子偏好性、mRNA 二级结构、启动子强度)的整合能力有限,Shimizu 模型和 Murray 模型中均不包含序列特征相关的反应或参数(表 1)。而这些 CRN 模型的局限性却能够被人工智能模型弥补。

1.3 蛋白折叠转运分泌过程及相关模型

从翻译得到的多肽链形成具有特定功能的

表 1 Murray 模型中翻译模型的模块划分和相关特征

Table 1 Module division and related features of the translation part of the Murray model

Submodule	Number of reactions	Key species (metabolites)	Related biological processes	Defects
Aminoacylation	42 M	21 aminoacyl-tRNA synthetases, 21 tRNA, amino acids, ATP	Aminoacyl-tRNA synthetase catalyzes the production of aminoacyl-tRNA	Not consider the effect of aminoacyl-tRNA synthetase type
Translation initiation	195	Initiation factor (IF1, IF2, IF3), ribosome 30S and 50S subunit, mRNA, fMet-tRNA, GTP	Under the mediation of initiation factors, mRNA, fMet-tRNAfMet, 30S and 50S subunits of ribosome consume GTP to generate 70S initiation complex	Not include many important factors affecting translation initiation, such as RBS efficiency (sequence, secondary structure), 5'UTR, start codon type
Translation elongation	46+14 N	Elongation factor (EF-Tu, EF-Ts, EF-G), ribosome, mRNA, aminoacyl-tRNA, GTP	Under the mediation of elongation factors, translation complex consumes GTP to carry out multiple rounds of amino acid addition reactions to generate polypeptide chain	Not consider the effect of codon-anticodon recognition efficiency, tRNA abundance, mRNA stability, coding region length, GC content, secondary structure, etc
Translation termination and ribosome recycling	72	Release factor (RF1, RF2, RF3), ribosome recycling factor (RRF), ribosome 30S and 50S subunit, mRNA, GTP	Under the mediation of release factors, the translation complex consumes GTP, recognizes the stop codon and dissociates	Not consider the effect of termination efficiency of stop codons and secondary structure of mRNA 3'UTR
Energy regeneration	61	Creatine kinase (CK), nucleoside diphosphate kinase (NDK), myokinase (MK), pyrophosphatase (PPiase), ATP, GTP, creatine	ATP and GTP regeneration catalyzed by CK, NDK, MK, and PPiase	None

M: Number of types of aminoacyl-tRNA synthetases; N: Number of amino acids in target protein.

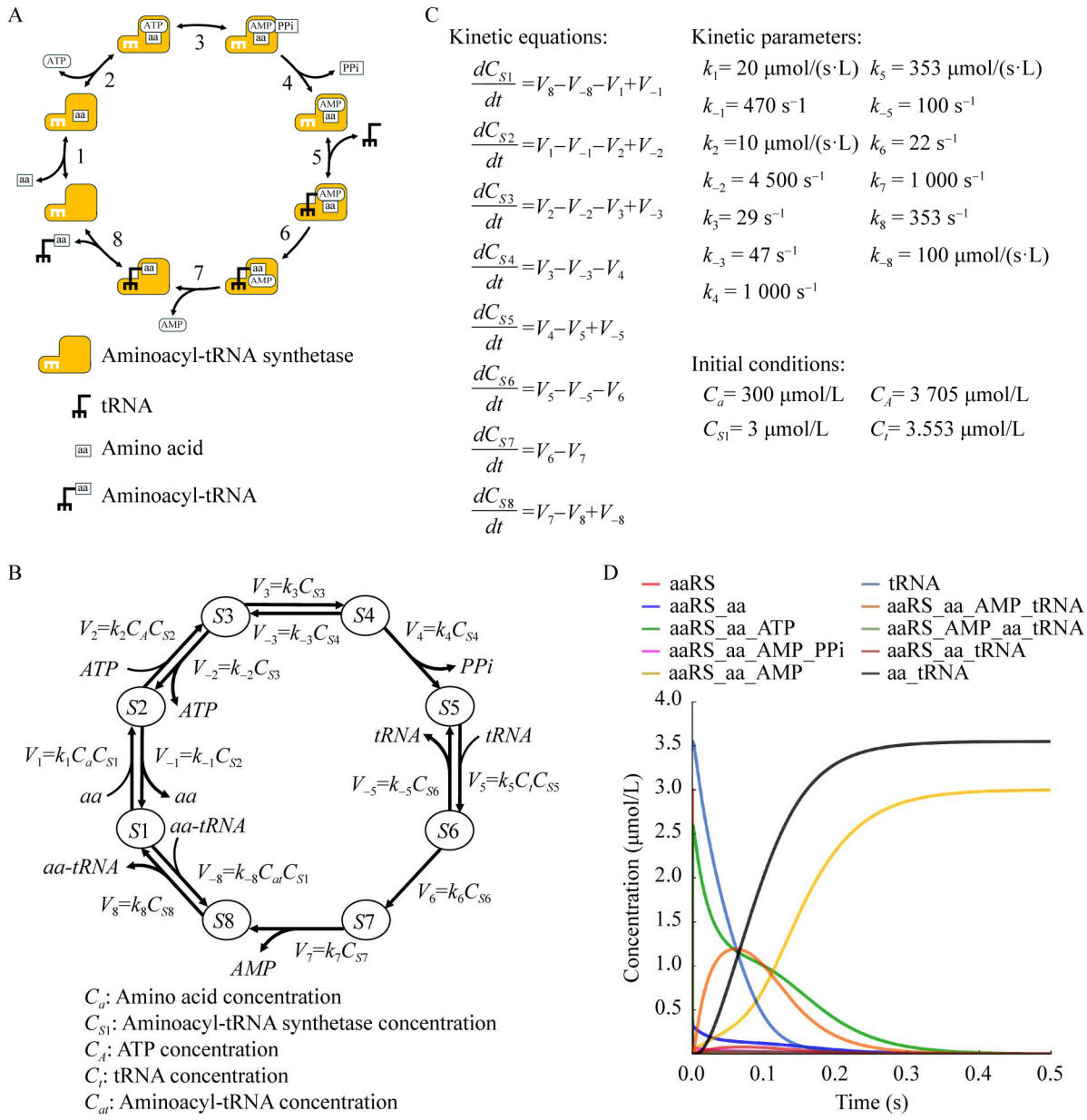

图 2　使用 CRN 模型模拟氨酰-tRNA 合成的流程　A：氨酰-tRNA 合成机制的示意图。B：CRN 模型的图形化表示。V_1–V_8：反应速率；S1–S8：包含氨酰-tRNA 合成酶的中间代谢物；C_{S1}–C_{S8}：包含氨酰-tRNA 合成酶的中间代谢物的浓度。C：CRN 模型，包含动力学方程、动力学参数和初始条件(参考 Murray 模型[49])。各个变量和参数分别对应 B 中的相同元素。D：动力学模拟结果。aaRS：氨酰-tRNA 合成酶；aa：氨基酸。

Figure 2　Simulation of the aminoacyl-tRNA synthesis process using the CRN model. A: Schematic diagram of the aminoacyl-tRNA synthesis mechanism. B: Graphical representation of the CRN model. V_1–V_8: Reaction rate; S1–S8: Intermediate metabolites including aminoacyl-tRNA synthetase; C_{S1}–C_{S8}: Concentration of intermediate metabolites of aminoacyl-tRNA synthetases. C: CRN model, including kinetic equations, kinetic parameters and initial conditions (refer to Murray model[49]). Each variable and parameter corresponds to the same element in B. D: Kinetic simulation results. aaRS: Aminoacyl-tRNA synthetase; aa: Amino acid.

蛋白结构也是一个复杂的、多阶段的过程，涉及蛋白质的折叠和不同亚基的组装等。多肽序列中氨基酸的排列决定了其如何折叠，每个氨基酸的化学性质(如极性、亲水性、疏水性等)均对折叠过程有重要影响[58]。同时细胞环境如pH 值、温度、离子浓度等因素也会影响蛋白质的折叠和稳定性[59]。蛋白翻译后修饰，如磷酸化、糖基化、脂酰化及二硫键的形成对蛋白正确四级结构的形成和稳定性也会产生重要影响[60]。在真核生物中蛋白质还要经过复杂的转运、定位和分泌过程到达细胞膜、特定细胞器内或其膜上才能实现正常功能。这个过程涉及蛋白信号序列的识别、内质网和高尔基体的转运、翻译后修饰以及细胞内各种运输小泡的协同作用[61-63]。

目前，PURE 系统的细粒化 CRN 模型(Shimizu 模型和 Murray 模型)均不包含蛋白折叠、转运以及翻译后修饰相关反应和参数，但可以借鉴其他类型的模型为 CRN 模型添加相应的模块。Li 等[64]在酿酒酵母基因组规模代谢网络模型基础上添加了蛋白质翻译、折叠和分泌过程，构建了一个酿酒酵母蛋白质分泌模型pcSecYeast。pcSecYeast 该模型不仅包含基本的质量守恒和代谢通量约束，还引入了蛋白质合成与代谢的耦合约束。该耦合约束将代谢模块提供的能量和底物与蛋白质合成所需的酶系和底物连接起来，从而反映出细胞在不同条件下资源的分配。此外，pcSecYeast 引入了针对重组蛋白与天然分泌蛋白竞争分泌途径的约束，能够模拟蛋白质分泌途径在重组蛋白过表达下的响应。pcSecYeast 是第 1 个能够全局性模拟蛋白质合成和分泌过程的模型。该模型不仅涵盖了从代谢到蛋白表达分泌的完整过程，还能够预测不同重组蛋白生产过程中的改造靶点，为认知蛋白合成和分泌机制以及优化重组蛋白生产提供了有力的计算工具。

2 蛋白表达的影响因素分析及人工智能预测

蛋白质表达受多种因素的影响，涵盖了从基因转录、蛋白翻译到翻译后修饰的各个步骤。每一个环节中的变化都可能显著影响最终的蛋白质数量和功能。上一节重点介绍了细胞转录翻译的过程及相关的各种酶、蛋白因子、tRNA、核糖体等的影响。基于机理的 CRN 模型可以实现对整个蛋白表达过程的系统性数学模拟，但实际很难精确测量细胞中所有蛋白表达影响因素，因此目前 CRN 模型还主要应用于蛋白表达过程的理论分析及指导无细胞表达系统特别是组分确定的 PURE 表达系统的优化。在实际的蛋白表达优化研究中，研究人员更常采用优化启动子等调控序列及密码子优化等手段实现蛋白表达的优化。而蛋白编码区和非编码区序列的特征对蛋白表达的影响非常复杂，目前尚无可靠的机理模型能够准确预测序列改变对蛋白表达的影响。因此基于海量序列数据文库构建数据驱动的人工智能模型是更可行的途径。图 3 列出了影响蛋白表达的主要序列特征(包括编码区序列和调控区序列)。下面对这些特征可能对蛋白表达过程产生的影响进行分析，并介绍相关的人工智能模型预测研究进展。这些因素是提高蛋白表达量的主要控制变量，对实现细胞蛋白表达的准确预测和控制非常重要。

2.1 调控区序列对蛋白表达影响分析和人工智能模型预测

基因编码区的上游和下游序列均可能对蛋白表达产生定量的影响，主要包括启动子、核糖体结合位点(RBS)、转录因子结合位点等(图 3A)。其中位于基因上游的启动子是最重要的调控区域，控制着 RNA 聚合酶的结合和转录的启动。启动子强度决定了转录效率，进而影响蛋白质

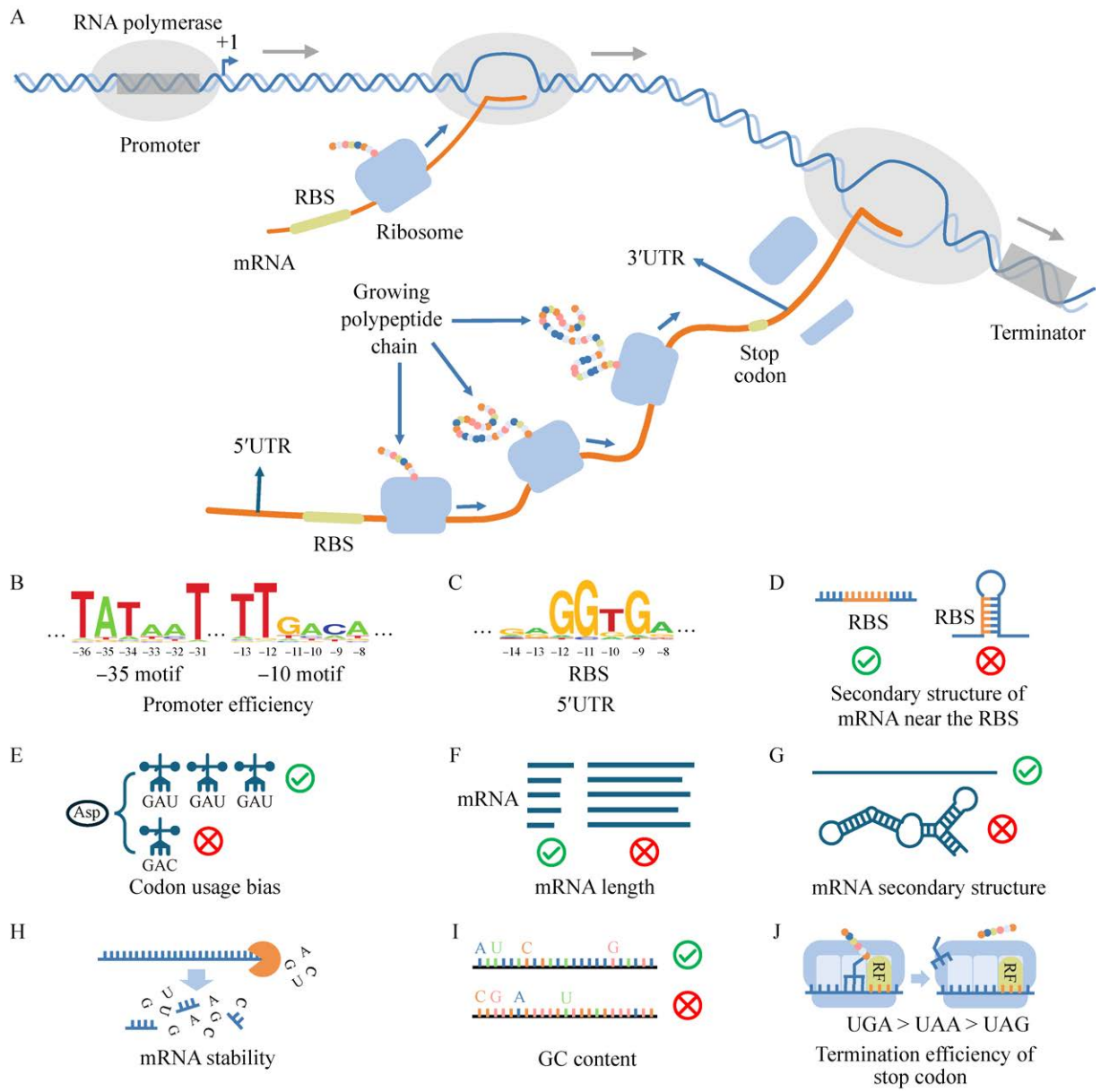

图 3　影响蛋白质表达的主要序列特征　A：原核生物蛋白质合成的简化示意图。其中多个 RNA 聚合酶(RNAP)分子同时转录单个基因，多个核糖体翻译单个单顺反子 mRNA。B–J：影响蛋白质表达的主要序列特征。绿色对号和红色叉号分别表示对蛋白表达有利或不利的特征。

Figure 3　Main sequence factors affecting protein expression. A: Simplified schematic view of procaryotic protein synthesis. Several RNA polymerase (RNAP) molecules simultaneously transcribe a single gene, and several ribosomes translate a single monocistronic mRNA. B–J: Main sequence factors affecting protein expression. Green check marks and red crosses indicate features that are beneficial or detrimental to protein expression respectively.

的表达水平(图 3B)。目前已有较多的采用人工智能模型预测启动子强度的研究工作，主要分为 2 类，一类首先基于 DNA 序列提取其物理化学特征参数，进而通过传统机器学习的方法如

支持向量机、随机森林、极端梯度提升算法(XGBoost)等预测启动子强度[65-67]。这类方法需要人工选取用于机器学习模型的序列特征，特征的选取对模型预测结果可能产生较大影响。优点是模型具有较好的可解释性并能确定对表达强度有重要影响的特征，从而便于指导优化启动子序列实现表达强度的准确控制。随着人工智能技术的发展，近年来基于深度学习和DNA语言模型的方法越来越受到重视，其跳过了繁琐的基于经验知识的特征提取步骤，直接由序列预测表达强度。特别是启动子序列区域常与一些转录因子结合位点重合或相连，难以确定明确的启动子边界及其特征，而直接由序列预测强度则可以更全面地考虑序列相关的各种因素影响，实现更准确地预测。早在2017年Umarov等[68]就提出一个CNNProm模型，通过卷积神经网络(convolutional neural network, CNN)定性识别不同生物中的启动子序列。Oubounyt等[69]进一步开发了深度学习模型DeePromoter，将CNN与双向长短期记忆模型(BiLSTM)结合，CNN提取序列的局部特征，BiLSTM捕获特征的长程依赖关系。并且创新地通过对启动子序列部分替换构造“硬负样本”，即在负样本中保留一定比例的启动子特征(如TATA-box)，让模型学习到更复杂的特征而非仅依赖特定位点的启动子序列特征。使用DeePromoter识别真核生物DNA序列中的启动子区域，特别是人类和小鼠的TATA和非TATA启动子，表现均优于CNNProm模型。Ma等[70]采用类似的卷积神经网络、双向长短期记忆模型和连接注意力机制相结合的方法开发了启动子计算识别工具DeeProPre，对真核生物启动子数据库(eukaryotic promoter database, EPD)的果蝇和小鼠启动子数据进行分析，预测准确率分别达到94.81%和98.62%。Wang等[71]将人工智能生成模型与预测模型相结合，通过生成对抗网络(generative adversarial network, GAN)模型从自然启动子中提取特征，并生成新的合成启动子序列；其训练数据集包含14 098个在*E. coli* K12 MG1655基因组中实验鉴定的启动子；GAN模型能够有效地捕捉关键的启动子序列特征，如k-mer频率、–10和–35基序及其间隔约束，生成的启动子在序列特征上更接近自然启动子；实验验证表明通过迭代优化后新生成的启动子中70.8%具有较高的活性。

近几年基于自注意力机制的变换器(Transformer)架构的深度学习模型极大地推动了人工智能研究的发展。Pipoli等[72]基于Transformer架构开发了一种通过处理基因启动子序列预测蛋白表达水平的新方法Transformer DeepLncLo。Transformer DeepLncLoc整合了Transformer架构和DeepLncLoc嵌入方法[73]，并使用word2vec算法规避稀疏矩阵问题，从而使模型能够更好地捕捉基因序列的语义特征。由Google开发的BERT模型继承了传统变换器的多头注意力机制，但仅包括编码器(encoder)部分并支持双向学习，非常适合语言文本及序列信息的处理[74]。DNABERT是一种借鉴BERT双向变换器结构提出的专门处理DNA序列数据的预训练模型，其通过前向和后向学习捕捉序列中每个核苷酸与其上下文的关系，从而更好地理解DNA序列的语义结构[75]。DNABERT将DNA序列分割成长度为k的连续核苷酸片段(k-mer)，每个k-mer作为一个“单词”输入模型。通过掩码语言模型进行预训练，将部分k-mer随机掩盖，模型根据上下文预测这些掩盖的k-mer，从而学习DNA序列中的潜在模式和特征。DNABERT具有高效的特征提取能力，因此被广泛应用于核酸序列相关的研究，如启动子识别、RNA结合蛋白预测、DNA甲基化位

点识别等。Le 等[76]使用预训练的 BERT 模型对 DNA 序列进行编码，开发了 BERT-Promoter 模型；其针对 Xiao 等[77]从 RegulonDB 数据库获取并标准化去重后得到的 3 382 个大肠杆菌启动子数据，应用 BERT 模型从中提取了 62 208 个特征；进而应用 Shapley 可加性解释(Shapley additive explanations, SHAP)分析，结合 Spearman 相关系数筛选出与预测精度关系最强的 653 个重要特征，以减少数据维度并消除冗余信息；并对选出的 BERT 特征使用多种机器学习算法评估模型性能。该方法将预训练模型用于特征提取，但预测仍采用传统机器学习的方法。而 Li 等[78]提出的基于 DNABERT 模型的集成预测器“msBERT-Promoter”则通过根据不同任务的微调和变换器的多头注意力机制实现多尺度特征提取，避免了复杂的特征提取步骤。其基于 DNABERT 模型，采用不同长度的 k-mer (3–6 个核苷酸)将序列分割为多个 token 获得多尺度特征。并通过软投票集成不同 k-mer 模型的预测结果以提高模型鲁棒性。该方法有效融合了局部和全局序列信息，在启动子识别和强度预测方面准确度均显著优于已有方法。

除了启动子，mRNA 上也存在非编码的调控序列，其中位于起始密码子上游的 5′非翻译区(5′ untranslated region, 5′UTR)最为重要。5′UTR 区包含的 RBS 负责引导核糖体与 mRNA 正确结合，确保核糖体能够准确定位到起始密码子并启动翻译。在原核生物中，RBS 通常位于起始密码子上游约 3–10 个核苷酸的位置，通常为一个富含嘌呤的序列，称为 Shine-Dalgarno 序列(图 3C)；该序列与核糖体小亚基的 16S rRNA 上的反向互补序列配对，确保核糖体正确结合[79]。该序列的匹配程度直接影响翻译效率。同时 RBS 与起始密码子之间的距离对翻译效率也有很大影响，距离过远或过近都会干扰核糖体的正确定位，从而降低蛋白质表达水平。此外，RBS 附近的 mRNA 二级结构(如发夹结构)可能阻碍核糖体的结合，如果 RBS 被二级结构隐藏或折叠，核糖体难以与之结合，导致翻译效率下降(图 3D)。真核生物中的翻译启动机制与原核生物有很大差异，核糖体与 mRNA 的结合依赖于 5′帽结构和 Kozak 序列，而非 RBS 机制。针对原核生物 5′UTR 对蛋白表达水平的影响，Gilliot 等[80]开发了一种结合 CNN 和长短期记忆(long short-term memory, LSTM)神经网络的混合模型(CNN-LSTM)；研究人员利用迁移学习的微调技术，在小样本数据上有效调整该模型，以适应不同实验数据的上下文信息，从而高效地预测大肠杆菌中 5′UTR 序列对蛋白质表达的影响。Chu 等[81]在 CNN-LSTM 模型基础上开发了 5′UTR 语言模型 UTR-LM。UTR-LM 模型基于 Transformer 架构，通过自监督学习进行预训练，并结合二级结构、自由能等监督信息以增强模型的预测能力。在经过多个下游任务的微调后，该模型在多个蛋白表达相关的关键指标预测中表现出卓越的性能。

近年来，人工智能模型在调控区序列优化中的应用已从理论探索转向实际工程改造。例如，Zhang 等[82]开发的 DeepSEED 框架通过结合 GAN 模型和条件约束，成功设计出高活性的合成启动子序列；经 DeepSEED 优化后的启动子在大肠杆菌和哺乳动物细胞中的表达强度分别提升了 2.3 倍和 1.8 倍，显著解决了传统启动子设计依赖试错法、效率低下的问题；这一成果直接指导了工业酶和重组蛋白药物的高效生产，例如在表达 β-内酰胺酶时，优化后的启动子使产量达到野生型的 3 倍以上。此外，其开发的 GPro 工具包[83]通过 AI 生成-预测闭环系统，实现了跨物种启动子的快速定制。用户仅需输入目标表达强度和宿主类型，GPro 即可生

成适配序列，并在毕赤酵母中成功将脂肪酶表达量提高至原有水平的 4.2 倍，为复杂宿主系统的蛋白表达优化提供了标准化解决方案。针对 5′UTR 序列的优化，Pan 等[84]开发的 UTR-Insight 模型通过整合 Transformer 架构与实验验证数据，实现了对 mRNA 翻译起始效率的精准预测；在灵长类细胞中，该模型筛选出的 5′UTR 序列使新冠疫苗刺突蛋白的表达量提升 2.5 倍，同时降低了 mRNA 二级结构对核糖体结合的阻碍，解决了真核系统中翻译效率低的关键瓶颈。这些案例表明，AI 模型不仅能够解析序列特征与表达量的关联，更能直接指导实验设计，缩短优化周期。

2.2 编码区序列对蛋白表达影响分析和人工智能模型预测

与调控区序列相比，研究人员对蛋白编码区序列对蛋白表达的影响研究较少，主要集中在密码子使用偏好方面。生物中的同一氨基酸常有多个不同的密码子编码，但因不同物种中 tRNA 的丰度不同而偏好不同，某些密码子在特定物种中能被更有效地识别。编码区不常见密码子的使用可能降低翻译速度甚至导致翻译停滞[85]，从而影响蛋白质产量及其正确折叠(图 3E)。因此外源蛋白表达时常需要对密码子进行优化，即使用与宿主 tRNA 丰度匹配的密码子，以提高翻译效率。研究人员已经提出了一些定量指标来表征蛋白与宿主密码子的匹配程度，例如密码子适应指数(codon adaptation index, CAI)[86]、tRNA 适应指数(tRNA adaptation index, TAI)[87]、相对密码子偏好评分(relative codon bias score, RCBS)[88]等。其中 CAI 应用最广泛，其假设每种氨基酸只有一个最优密码子，并且使用该密码子的频率越高，蛋白质的表达水平越高。这种单一最优密码子的假设忽略了其他同义密码子的潜在贡献。因此 Zaytsev 等[89]提出了改进的指标密码子表达指数评分(codon expression index score, CEIS)和密码子生产率评分(codon productivity score, CPS)，其不将最优密码子作为唯一指标，而是分析每一个密码子的实际贡献，并引入正负效应的概念，即某些密码子可能会对蛋白质表达产生正面或负面影响。这种改进使模型能够更灵活地反映同义密码子的多样性和实际影响。该研究进一步提出了密码子对表达指数(codon pair expression index, CPEI)参数来分析相邻密码子对的影响，以提高预测准确性[89]。该参数对高表达基因预测效果更好，但对于表达水平较低的基因预测准确性较差，可能原因是低表达基因很可能受密码子以外的因素影响。

除了密码子偏好性外，其他可能影响蛋白表达量的因素包括长度、mRNA 二级结构、稳定性、特定片段的核苷酸组合(例如 GC 含量)、终止密码子等。较长的 mRNA 序列通常会降低核糖体在整个翻译过程中保持活性的概率，导致蛋白质产量的减少(图 3F)。mRNA 的二级结构会影响核糖体的结合和其行进速度，从而降低蛋白质合成的效率(图 3G)。mRNA 的稳定性决定了单条 RNA 可以翻译成多肽链的个数(图 3H)。某些序列片段或局部的 GC 含量会影响核糖体的行进速度。过高的 GC 含量可能导致翻译停滞或延迟，而较低的 GC 含量则可以加快翻译效率(图 3I)。不同的终止密码子(UAA、UAG、UGA)在翻译终止效率上存在差异。使用高效的终止密码子可以确保翻译快速正确地结束，从而减少翻译延迟并提高蛋白质表达量(图3J)。此外，原核生物中转录和翻译同时进行，二者速率的不协调可能导致蛋白的错误折叠或聚集，影响蛋白表达。为了对编码序列特征对蛋白表达量的影响进行更深入系统的分析，Boël 等[90]从来自多种进化分支的 171 种不同生物体中选

取 6 348 个基因在大肠杆菌中进行表达，其中任意 2 条蛋白质序列的相似度均低于 60%，以减少序列冗余，确保序列多样性；研究人员通过广泛采样使得数据集涵盖了多种密码子的使用频率组合，从而更好地评估不同密码子频率对表达效率的影响；他们将表达水平从低到高分为 5 类，通过多变量逻辑回归模型分析各种因素对蛋白表达的影响；结果表明密码子比 mRNA 二级结构对蛋白表达有更大的影响，且编码序列中前 18 个核苷酸对表达量有重要影响；基于该模型他们提出了 2 种新的序列优化方法(针对 6 种氨基酸密码子优化的 6AA 和针对折叠优化的 31C-FO)并成功应用于 19 个蛋白的表达量优化。

蛋白序列的不同区域对蛋白表达影响程度不同，一般认为起始密码子附近的氨基酸序列对蛋白表达有较大影响。例如 Verma 等[91]通过在 *eGFP* 基因起始区插入随机核苷酸序列改变了编码第 3–5 个氨基酸的密码子，构建了一个包含超过 250 000 个报告基因的 *eGFP* 库。对相关数据进行的分析表明该区域对蛋白产量影响很大，而且其影响与 tRNA 丰度、翻译启动效率和 mRNA 整体结构等其他因素无关；这一区域被称为翻译延长减速带(translational elongation short ramp, TESR)：通过适当降低翻译延长速率避免分子撞车导致的蛋白合成提前终止。基于这一数据 Kim 等[92]提出了一个深度学习模型，通过卷积神经网络预测不同 TESR 序列对应的蛋白表达；应用该模型对大肠杆菌的 4 305 个编码序列进行分析，发现同源基因群集(clusters of orthologous groups, COGs)中 TESR 的保守性。Cambray 等[93]通过在 GFP 蛋白编码基因前端添加 96 个碱基生成了一个包含 244 000 个不同序列的基因库，针对 8 个已知的影响翻译效率的指标通过可控随机化的方法使生成序列覆盖较广的序列空间，并通过荧光激活细胞分选测序(fluorescence-activated cell sorting sequencing, FACS-sequencing)高通量分析方法确定不同序列对应的蛋白在大肠杆菌中的表达量，进而通过统计分析确定了影响蛋白表达的主要因素。蛋白大语言模型的发展也为序列特征的精准表征以及挖掘更多影响蛋白表达的序列因素提供了强有力的支持。Liu 等[94]基于 UniRef50 数据库构建了一个包含 59 142 917 个蛋白质序列的数据集，并开发了一个基于 Transformer 架构的 MP-TRANS 预训练模型；在此基础上，通过迁移学习技术微调生成了 88 个预测模型(MP-EXP)，用于预测 88 个物种的蛋白表达水平；这些模型的平均准确率达到 0.78，显著优于传统模型的表现；这一成果不仅展示了蛋白大语言模型在提升下游任务性能方面的巨大潜力，也为未来蛋白质表达预测和优化提供了新的思路和方法。

人工智能模型在编码区序列优化中的核心价值在于其能够系统性挖掘隐蔽的序列特征，并指导理性设计。Ding 等[95]开发的 MPEPE 模型通过深度学习从 6 348 个蛋白表达数据中提取出关键特征，成功预测了漆酶(13B22)和葡萄糖脱氢酶(FAD-AtGDH)的突变热点；实验结果显示，多点突变组合使两者的表达量分别提升了 3.49 倍和 7.86 倍，且酶活性未受显著影响，证明了 AI 模型在平衡表达量与功能完整性方面的优势；这一方法已被应用于工业酶制剂的规模化生产，例如在纤维素酶改造中，MPEPE 预测的突变使发酵液酶活达到 12.8 U/mL，较野生型提升 4.1 倍。针对密码子优化，Jain 等[96]开发的 ICOR 工具利用双向 LSTM 网络捕捉密码子上下文依赖性，突破了传统 CAI 指数仅考虑单密码子频率的局限；在表达人类胰岛素原的大肠杆菌系统中，经过 ICOR 密码子优化的

序列的可溶性蛋白产量提升了 2.2 倍，同时包涵体形成比例从 35%降至 12%，显著缓解了原核系统表达真核蛋白的折叠难题。此外，Nikolados 等[97]通过对比深度学习与多种传统机器学习方法的预测结果，证明了 CNN 能够更精准地预测 GC 含量与翻译停滞的关系；在膜蛋白 GPCR 的表达优化中，基于 CNN 模型的序列设计将全长蛋白产量从 0.5 mg/L 提升至 3.2 mg/L，为高难度蛋白的生产提供了新策略。

3 结论与展望

细胞的蛋白表达过程是一个受多种因素影响、多层次严密调控的复杂过程，其不仅与目标蛋白序列和相关调控序列有关，还与表达系统中的聚合酶、tRNA 丰度、氨酰 tRNA 活性等有关。这些胞内因素的实验测量难度较大，在不同宿主细胞中也存在较大差异，此外还有很多其他未知蛋白和因素也可能产生影响，同时蛋白质合成需要的原料和能量供应还与细胞全局代谢调控密切关联。这些都使得蛋白表达过程的建模预测要比蛋白结构功能预测更为复杂。化学反应网络等机理模型虽然可以描述转录翻译过程的定量动力学，其参数化过程(如 tRNA 浓度、聚合酶丰度)往往依赖于简化的质量作用定律假设和间接数据拟合。值得注意的是，GC 含量、mRNA 二级结构等宏观层级序列特征对蛋白质表达的影响难以在机理模型中有效整合，这进一步限制了其预测精度。

数据驱动的人工智能方法为研究这一多因素非线性的复杂过程提供了新的思路，特别是基因合成成本的快速下降和基于测序的高通量蛋白表达分析技术的发展，推动了大规模蛋白质表达数据集的产生。本文系统总结了可用于蛋白表达分析的高通量数据集及相应的人工智能建模进展[71,76-77,82,90-93,95,97]。需要强调的是模型可靠性高度依赖于数据质量与覆盖度：高质量数据集需通过实验设计优化实现序列空间的无偏好采样。而研究关注的因素也影响着实验的序列空间采样偏好，例如 Verma 等[91]仅通过改变编码区第 3–5 个氨基酸的密码子就构建了一个包含 250 000 个报告基因的蛋白库，而大多数蛋白都包含数百个氨基酸，而且除了编码区外，调控区序列也有重要影响。因此实验设计中如何兼顾更多的因素并大幅减少数据的冗余度以避免长序列空间的组合爆炸是基于人工智能方法研究蛋白表达必须解决的问题。Nikolados 等[97]的研究表明，基于特定序列空间范围内的数据训练得到的模型，在预测不同序列空间范围的测试序列时会出现显著性能衰减。因此，实验条件的异质性(如宿主类型、培养条件、检测方法)导致多数据集整合面临严峻挑战，目前人工智能模型仍局限于特定宿主-条件体系的表达预测，缺乏普适性。

可见，机理模型与人工智能方法在建模维度上存在显著互补性。机理模型侧重描述底盘系统的分子动力学过程(如酶促反应、复合体组分的结合和解离相关动力学)，但将转录和翻译过程的动力学简化为与序列及其上下文无关的均质化参数；而人工智能模型虽然能捕获蛋白编码和调控序列特征与蛋白表达水平的关系，却无法量化底盘系统分子组分的动态影响。在此背景下，AI 模型与机理模型的结合可以弥补各自的缺点。例如，在确定机理模型动力学参数的过程中，如果已经纯化了某个蛋白酶并测量了其不同底物浓度下的催化反应速率则可以直接拟合得到其动力学参数如 k_{cat}、K_m 等，此时不需要人工智能方法。但很多情况下这种数据是缺失的，尤其是转录翻译过程中很难直接对序列延长过程进行动力学测量分析，因此在机理模型中常常简单地假定对所有核苷酸(氨

基酸)参数是相同的且与其序列位置无关。但实际上序列特征会对转录翻译效率产生很大影响，此时采用人工智能方法基于序列预测某一特定步骤的动力学参数变化会有所帮助。

AI模型与机理模型的结合为蛋白表达预测提供了新的思路，但目前缺乏成熟的结合方法。一个解决思路是采用类似于酶约束模型中通过AI 模型预测酶动力学参数的方法，通过 AI 模型预测转录翻译机理模型中每步序列延长过程的参数，使其不再是个固定值而是随着序列的变化而变化，从而可以将这个变化的参数值引入机理模型中。此外也可以在 AI 模型中将底盘细胞中关键因子的浓度也作为输入变量，从而同时考虑底盘中的因素和序列特征对蛋白表达的影响。例如，可以构建一个多模态深度学习模型，将目标蛋白的编码及调控序列转化为嵌入向量，同时将底盘系统中的关键因子(如 RNA 聚合酶、核糖体、tRNA 等)的浓度数据经过标准化后作为数值输入。两部分特征在中间层进行融合，再经过全连接层映射至表达水平预测。该方法既包含了序列特征，又反映了细胞环境，为 AI 与机理模型的协同优化提供了新思路。

在应用方面，AI 与机理模型的结合有望在生物制药和工业生物技术领域发挥重要作用。工业酶生产中，机理模型可以用于解析酶的合成过程，并结合 AI 模型优化其编码和表达调控序列，从而提高酶的产量和活性。在疫苗开发中，AI 模型可以基于抗原编码序列和表达调控序列的实验数据预测并优化不同疫苗抗原的表达水平，机理模型可用于模拟宿主细胞内的转录、翻译与蛋白折叠过程，解析影响抗原表达的关键因素，如核糖体占用率、分子伴侣的辅助折叠效应以及分泌机制的限制。结合 AI 和机理模型，不仅可以提高对抗原表达水平的预测精度，还能指导宿主优化策略，如调整培养条件或基因调控机制，以提升疫苗蛋白的表达效率。随着实验技术和计算方法的不断进步，AI 与机理模型的结合有望成为未来蛋白表达研究的重要方向，为蛋白质工程和合成生物学提供更强大的预测工具。

REFERENCES

[1] KEASLING JD. Manufacturing molecules through metabolic engineering[J]. Science, 2010, 330(6009): 1355-1358.

[2] ROSANO GL, CECCARELLI EA. Recombinant protein expression in *Escherichia coli*: advances and challenges[J]. Frontiers in Microbiology, 2014, 5: 172.

[3] ZHANG XP, AL-DOSSARY A, HUSSAIN M, SETLOW P, LI JH. Applications of *Bacillus subtilis* spores in biotechnology and advanced materials[J]. Applied and Environmental Microbiology, 2020, 86(17): e01096-20.

[4] AHMAD M, HIRZ M, PICHLER H, SCHWAB H. Protein expression in *Pichia pastoris*: recent achievements and perspectives for heterologous protein production[J]. Applied Microbiology and Biotechnology, 2014, 98(12): 5301-5317.

[5] YANG S, SONG LY, WANG J, ZHAO JZ, TANG HT, BAO XM. Engineering *Saccharomyces cerevisiae* for efficient production of recombinant proteins[J]. Engineering Microbiology, 2024, 4(1): 100122.

[6] KAIPA JM, KRASNOSELSKA G, OWENS RJ, van den HEUVEL J. Screening of membrane protein production by comparison of transient expression in insect and mammalian cells[J]. Biomolecules, 2023, 13(5): 817.

[7] LEE KH, KIM DM. Recent advances in development of cell-free protein synthesis systems for fast and efficient production of recombinant proteins[J]. FEMS Microbiology Letters, 2018, 365(17).

[8] HARTMAN MCT, JOSEPHSON K, LIN CW, SZOSTAK JW. An expanded set of amino acid analogs for the ribosomal translation of unnatural peptides[J]. PLoS One, 2007, 2(10): e972.

[9] LU Y, WELSH JP, SWARTZ JR. Production and stabilization of the trimeric influenza hemagglutinin stem domain for potentially broadly protective influenza vaccines[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(1): 125-130.

[10] HOFFMANN B, LÖHR F, LAGUERRE A, BERNHARD F, DÖTSCH V. Protein labeling strategies for liquid-state NMR spectroscopy using cell-free synthesis[J]. Progress in Nuclear Magnetic Resonance Spectroscopy, 2018, 105: 1-22.

[11] ZHOU JW, HUANG L, LIAN JZ, SHENG JY, CAI J, XU ZN. Reconstruction of the UDP-N-acetylglucosamine biosynthetic pathway in cell-free system[J]. Biotechnology Letters, 2010, 32(10): 1481-1486.

[12] NUMATA K, MOTODA Y, WATANABE S, OSANAI T,

KIGAWA T. Co-expression of two polyhydroxyalkanoate synthase subunits from *Synechocystis* sp. PCC 6803 by cell-free synthesis and their specific activity for polymerization of 3-hydroxybutyryl-coenzyme A[J]. Biochemistry, 2015, 54(6): 1401-1407.
[13] BROOKWELL A, OZA JP, CASCHERA F. Biotechnology applications of cell-free expression systems[J]. Life, 2021, 11(12): 1367.
[14] GREGORIO NE, LEVINE MZ, OZA JP. A user's guide to cell-free protein synthesis[J]. Methods and Protocols, 2019, 2(1): 24.
[15] SHIMIZU Y, INOUE A, TOMARI Y, SUZUKI T, YOKOGAWA T, NISHIKAWA K, UEDA T. Cell-free translation reconstituted with purified components[J]. Nature Biotechnology, 2001, 19(8): 751-755.
[16] KURUMA Y, UEDA T. The PURE system for the cell-free synthesis of membrane proteins[J]. Nature Protocols, 2015, 10(9): 1328-1344.
[17] DOERR A, de REUS E, van NIES P, van der HAAR M, WEI K, KATTAN J, WAHL A, DANELON C. Modelling cell-free RNA and protein synthesis with minimal systems[J]. Physical Biology, 2019, 16(2): 025001.
[18] TAGUCHI H, NIWA T. Reconstituted cell-free translation systems for exploring protein folding and aggregation[J]. Journal of Molecular Biology, 2024, 436(19): 168726.
[19] GANESH RB, MAERKL SJ. Towards self-regeneration: exploring the limits of protein synthesis in the protein synthesis using recombinant elements (PURE) cell-free transcription-translation system[J]. ACS Synthetic Biology, 2024, 13(8): 2555-2566.
[20] CUI Y, CHEN XJ, WANG Z, LU Y. Cell-free PURE system: evolution and achievements[J]. BioDesign Research, 2022, 2022: 9847014.
[21] LERMAN JA, HYDUKE DR, LATIF H, PORTNOY VA, LEWIS NE, ORTH JD, SCHRIMPE-RUTLEDGE AC, SMITH RD, ADKINS JN, ZENGLER K, PALSSON BO. In silico method for modelling metabolism and gene product expression at genome scale[J]. Nature Communications, 2012, 3: 929.
[22] O'BRIEN EJ, LERMAN JA, CHANG RL, HYDUKE DR, PALSSON BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction[J]. Molecular Systems Biology, 2013, 9: 693.
[23] LIU JK, LLOYD C, AL-BASSAM MM, EBRAHIM A, KIM JN, OLSON C, AKSENOV A, DORRESTEIN P, ZENGLER K. Predicting proteome allocation, overflow metabolism, and metal requirements in a model acetogen[J]. PLoS Computational Biology, 2019, 15(3): e1006848.
[24] KARR JR, SANGHVI JC, MACKLIN DN, GUTSCHOW MV, JACOBS JM, BOLIVAL B Jr, ASSAD-GARCIA N, GLASS JI, COVERT MW. A whole-cell computational model predicts phenotype from genotype[J]. Cell, 2012, 150(2): 389-401.
[25] MACKLIN DN, AHN-HORST TA, CHOI H, RUGGERO NA, CARRERA J, MASON JC, SUN G, AGMON E, DeFELICE MM, MAAYAN I, LANE K, SPANGLER RK, GILLIES TE, PAULL ML, AKHTER S, BRAY SR, WEAVER DS, KESELER IM, KARP PD, MORRISON JH, COVERT MW. Simultaneous cross-evaluation of heterogeneous *E. coli* datasets *via* mechanistic simulation[J]. Science, 2020, 369(6502): eaav3751.
[26] STRICKER J, COOKSON S, BENNETT MR, MATHER WH, TSIMRING LS, HASTY J. A fast, robust and tunable synthetic gene oscillator[J]. Nature, 2008, 456(7221): 516-519.
[27] GARDNER TS, CANTOR CR, COLLINS JJ. Construction of a genetic toggle switch in *Escherichia coli*[J]. Nature, 2000, 403(6767): 339-342.
[28] MOON TS, LOU CB, TAMSIR A, STANTON BC, VOIGT CA. Genetic programs constructed from layered logic gates in single cells[J]. Nature, 2012, 491(7423): 249-253.
[29] BASU S, GERCHMAN Y, COLLINS CH, ARNOLD FH, WEISS R. A synthetic multicellular system for programmed pattern formation[J]. Nature, 2005, 434(7037): 1130-1134.
[30] GOODWIN BC. Oscillatory behavior in enzymatic control processes[J]. Advances in Enzyme Regulation, 1965, 3: 425-437.
[31] FEINBERG M. Complex balancing in general kinetic systems[J]. Archive for Rational Mechanics and Analysis, 1972, 49(3): 187-194.
[32] HORN F, JACKSON R. General mass action kinetics[J]. Archive for Rational Mechanics and Analysis, 1972, 47(2): 81-116.
[33] KRAMBECK FJ. The mathematical structure of chemical kinetics in homogeneous single-phase systems[J]. Archive for Rational Mechanics and Analysis, 1970, 38(5): 317-347.
[34] KLIPP E, NORDLANDER B, KRÜGER R, GENNEMARK P, HOHMANN S. Integrative model of the response of yeast to osmotic shock[J]. Nature Biotechnology, 2005, 23(8): 975-982.
[35] ELOWITZ MB, LEIBLER S. A synthetic oscillatory network of transcriptional regulators[J]. Nature, 2000, 403(6767): 335-338.
[36] GUPTA A, KHAMMASH M. Sensitivity analysis for stochastic chemical reaction networks with multiple time-scales[J]. Electronic Journal of Probability, 2014, 19: 1-53.
[37] RAO CV, ARKIN AP. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm[J]. The Journal of Chemical Physics, 2003, 118(11): 4999-5010.
[38] BARKAI N, LEIBLER S. Robustness in simple biochemical networks[J]. Nature, 1997, 387(6636): 913-917.
[39] THATTAI M, van OUDENAARDEN A. Intrinsic noise in gene regulatory networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(15): 8614-8619.
[40] ZOURIDIS H, HATZIMANIKATIS V. A model for protein translation: polysome self-organization leads to maximum protein synthesis rates[J]. Biophysical Journal, 2007, 92(3): 717-730.
[41] MEHRA A, HATZIMANIKATIS V. An algorithmic framework for genome-wide modeling and analysis of translation networks[J]. Biophysical Journal, 2006, 90(4): 1136-1146.

[42] STÖGBAUER T, WINDHAGER L, ZIMMER R, RÄDLER JO. Experiment and mathematical modeling of gene expression dynamics in a cell-free system[J]. Integrative Biology, 2012, 4(5): 494-501.
[43] TUZA ZA, SIEGAL-GASKINS D, KIM J, SZEDERKÉNYI G. Analysis-based parameter estimation of an *in vitro* transcription-translation system[C]. 2015 European Control Conference (ECC). July 15-17, 2015, Linz, Austria. IEEE, 2015: 1560-1566.
[44] MARSHALL R, NOIREAUX V. Quantitative modeling of transcription and translation of an all-*E. coli* cell-free system[J]. Scientific Reports, 2019, 9(1): 11980.
[45] MATSUBAYASHI H, UEDA T. Purified cell-free systems as standard parts for synthetic biology[J]. Current Opinion in Chemical Biology, 2014, 22: 158-162.
[46] MAVELLI F, MARANGONI R, STANO P. A simple protein synthesis model for the PURE system operation[J]. Bulletin of Mathematical Biology, 2015, 77(6): 1185-1212.
[47] MATSUURA T, TANIMURA N, HOSODA K, YOMO T, SHIMIZU Y. Reaction dynamics analysis of a reconstituted *Escherichia coli* protein translation system by computational modeling[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(8): E1336-E1344.
[48] MATSUURA T, HOSODA K, SHIMIZU Y. Robustness of a reconstituted *Escherichia coli* protein translation system analyzed by computational modeling[J]. ACS Synthetic Biology, 2018, 7(8): 1964-1972.
[49] JURADO Z, PANDEY A, MURRAY RM. A chemical reaction network model of PURE[J]. bioRxiv, 2023: 1-16.
[50] RICHARD P, MANLEY JL. Transcription termination by nuclear RNA polymerases[J]. Genes & Development, 2009, 23(11): 1247-1269.
[51] MOORE MJ, PROUDFOOT NJ. Pre-mRNA processing reaches back to transcription and ahead to translation[J]. Cell, 2009, 136(4): 688-700.
[52] KOZAK M. Initiation of translation in prokaryotes and eukaryotes[J]. Gene, 1999, 234(2): 187-208.
[53] NOLLER HF, YUSUPOV MM, YUSUPOVA GZ, BAUCOM A, CATE JHD. Translocation of tRNA during protein synthesis[J]. FEBS Letters, 2002, 514(1): 11-16.
[54] RODNINA MV, WINTERMEYER W. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms[J]. Annual Review of Biochemistry, 2001, 70: 415-435.
[55] KISSELEV LL, BUCKINGHAM RH. Translational termination comes of age[J]. Trends in Biochemical Sciences, 2000, 25(11): 561-566.
[56] POOLE W, PANDEY A, SHUR A, TUZA ZA, MURRAY RM. BioCRNpyler: Compiling chemical reaction networks from biomolecular parts in diverse contexts[J]. PLoS Computational Biology, 2022, 18(4): e1009987.
[57] PANDEY A, POOLE W, SWAMINATHAN A, HSIAO V, MURRAY RM. Fast and flexible simulation and parameter estimation for synthetic biology using bioscrape[J]. Journal of Open Source Software, 2023, 8(83): 5057.
[58] DOBSON CM. Experimental investigation of protein folding and misfolding[J]. Methods, 2004, 34(1): 4-14.
[59] ENGLANDER SW, MAYNE L, KRISHNA MMG. Protein folding and misfolding: mechanism and principles[J]. Quarterly Reviews of Biophysics, 2007, 40(4): 287-326.
[60] YE HQ, HAN Y, LI P, SU ZD, HUANG YQ. The role of post-translational modifications on the structure and function of tau protein[J]. Journal of Molecular Neuroscience, 2022, 72(8): 1557-1571.
[61] OWJI H, NEZAFAT N, NEGAHDARIPOUR M, HAJIEBRAHIMI A, GHASEMI Y. A comprehensive review of signal peptides: structure, roles, and applications[J]. European Journal of Cell Biology, 2018, 97(6): 422-441.
[62] AKOPIAN D, SHEN K, ZHANG X, SHAN SO. Signal recognition particle: an essential protein-targeting machine[J]. Annual Review of Biochemistry, 2013, 82: 693-721.
[63] BONIFACINO JS, GLICK BS. The mechanisms of vesicle budding and fusion[J]. Cell, 2004, 116(2): 153-166.
[64] LI FR, CHEN Y, QI Q, WANG YY, YUAN L, HUANG MT, ELSEMMAN IE, FEIZI A, KERKHOVEN EJ, NIELSEN J. Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints[J]. Nature Communications, 2022, 13(1): 2969.
[65] 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.
[66] 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.
[67] WANG Y, TAI SW, ZHANG SQ, SHENG N, XIE XP. PromGER: promoter prediction based on graph embedding and ensemble learning for eukaryotic sequence[J]. Genes, 2023, 14(7): 1441.
[68] UMAROV RK, SOLOVYEV VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks[J]. PLoS One, 2017, 12(2): e0171410.
[69] OUBOUNYT M, LOUADI Z, TAYARA H, CHONG KT. DeePromoter: robust promoter predictor using deep learning[J]. Frontiers in Genetics, 2019, 10: 286.
[70] MA ZW, ZHAO JP, TIAN J, ZHENG CH. DeeProPre: a promoter predictor based on deep learning[J]. Computational Biology and Chemistry, 2022, 101: 107770.
[71] 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.
[72] PIPOLI V, CAPPELLI M, PALLADINI A, PELUSO C, LOVINO M, FICARRA E. Predicting gene expression levels from DNA sequences and post-transcriptional information with transformers[J]. Computer Methods

and Programs in Biomedicine, 2022, 225: 107035.
[73] ZENG M, WU YF, LU CQ, ZHANG FH, WU FX, LI M. DeepLncLoc: a deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding[J]. Briefings in Bioinformatics, 2022, 23(1): bbab360.
[74] DEVLIN J, CHANG MW, LEE K, TOUTANOVA K, ASSOC COMPUTAT L. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), 2019: 4171-4186.
[75] JI YR, ZHOU ZH, LIU H, DAVULURI RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome[J]. Bioinformatics, 2021, 37(15): 2112-2120.
[76] LE NQK, HO QT, NGUYEN VN, CHANG JS. BERT-Promoter: an improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection[J]. Computational Biology and Chemistry, 2022, 99: 107732.
[77] XIAO X, XU ZC, QIU WR, WANG P, GE HT, CHOU KC. iPSW(2L)-PseKNC: a two-layer predictor for identifying promoters and their strength by hybrid features *via* pseudo K-tuple nucleotide composition[J]. Genomics, 2019, 111(6): 1785-1793.
[78] LI YZ, WEI XM, YANG QL, XIONG A, LI XF, ZOU Q, CUI FF, ZHANG ZL. msBERT-Promoter: a multi-scale ensemble predictor based on BERT pre-trained model for the two-stage prediction of DNA promoters and their strengths[J]. BMC Biology, 2024, 22(1): 126.
[79] SHINE J, DALGARNO L. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites[J]. Proceedings of the National Academy of Sciences of the United States of America, 1974, 71(4): 1342-1346.
[80] GILLIOT PA, GOROCHOWSKI TE. Transfer learning for cross-context prediction of protein expression from 5′UTR sequence[J]. Nucleic Acids Research, 2024, 52(13): e58.
[81] CHU YY, YU D, LI YP, HUANG KX, SHEN Y, CONG L, ZHANG J, WANG MD. A 5′UTR language model for decoding untranslated regions of mRNA and function predictions[J]. Nature Machine Intelligence, 2024, 6(4): 449-460.
[82] 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.
[83] 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.
[84] PAN SC, WANG HY, ZHANG H, TANG Z, XU LQ, YAN ZX, HU Y. UTR-Insight: integrating deep learning for efficient 5′UTR discovery and design[J]. BMC Genomics, 2025, 26(1): 107.
[85] KOMAR AA. A code within a code: how codons fine-tune protein folding in the cell[J]. Biochemistry Biokhimiia, 2021, 86(8): 976-991.
[86] SHARP PM, LI WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications[J]. Nucleic Acids Research, 1987, 15(3): 1281-1295.
[87] dos REIS M, SAVVA R, WERNISCH L. Solving the riddle of codon usage preferences: a test for translational selection[J]. Nucleic Acids Research, 2004, 32(17): 5036-5044.
[88] DAS S, ROYMONDAL U, SAHOO S. Analyzing gene expression from relative codon usage bias in yeast genome: a statistical significance and biological relevance[J]. Gene, 2009, 443(1/2): 121-131.
[89] ZAYTSEV K, BOGATYREVA N, FEDOROV A. Link between individual codon frequencies and protein expression: going beyond codon adaptation index[J]. International Journal of Molecular Sciences, 2024, 25(21): 11622.
[90] BOËL G, LETSO R, NEELY H, PRICE WN, WONG KH, SU M, LUFF J, VALECHA M, EVERETT JK, ACTON TB, XIAO R, MONTELIONE GT, AALBERTS DP, HUNT JF. Codon influence on protein expression in *E. coli* correlates with mRNA levels[J]. Nature, 2016, 529(7586): 358-363.
[91] VERMA M, CHOI J, COTTRELL KA, LAVAGNINO Z, THOMAS EN, PAVLOVIC-DJURANOVIC S, SZCZESNY P, PISTON DW, ZAHER HS, PUGLISI JD, DJURANOVIC S. A short translational ramp determines the efficiency of protein synthesis[J]. Nature Communications, 2019, 10(1): 5774.
[92] KIM DJ, KIM J, LEE DH, LEE J, WOO HM. DeepTESR: a deep learning framework to predict the degree of translational elongation short ramp for gene expression control[J]. ACS Synthetic Biology, 2022, 11(5): 1719-1726.
[93] CAMBRAY G, GUIMARAES JC, ARKIN AP. Evaluation of 244, 000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*[J]. Nature Biotechnology, 2018, 36(10): 1005-1015.
[94] LIU TY, ZHANG YY, LI YJ, XU GS, GAO H, WANG PT, TU T, LUO HY, WU NF, YAO B, LIU B, GUAN FF, HUANG HQ, TIAN J. Effective gene expression prediction and optimization from protein sequences[J]. Advanced Science, 2025: e2407664.
[95] DING ZD, GUAN FF, XU GS, WANG YC, YAN YR, ZHANG W, WU NF, YAO B, HUANG HQ, TULLER T, TIAN J. MPEPE, a predictive approach to improve protein expression in *E. coli* based on deep learning[J]. Computational and Structural Biotechnology Journal, 2022, 20: 1142-1153.
[96] JAIN R, JAIN A, MAURO E, LESHANE K, DENSMORE D. ICOR: improving codon optimization with recurrent neural networks[J]. BMC Bioinformatics, 2023, 24(1): 132.
[97] 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.