特约评述

DOI: 10.12211/2096-8280.2024-090

机器学习驱动的基因组规模代谢模型构建与优化

吴柯[1,2]，罗家豪[1,2]，李斐然[1,2]
（[1]清华大学深圳国际研究生院，生物医药与健康工程研究院，广东 深圳 518055；[2]清华大学化学工程系，工业生物催化教育部重点实验室，北京 100084）

摘要：自1999年首个基因组规模代谢模型（genome-scale metabolic model，GEM）问世以来，GEM已成为解析生物代谢的重要工具。该模型包含代谢基因、代谢物和反应，并结合化学计量矩阵与约束优化，系统描述和模拟生物体内的代谢过程。此外，GEM能够整合热力学参数、动力学参数、多组学数据及多细胞过程，构建更精细且具有更强大预测能力的多约束多过程模型。然而，先验知识的局限成为其发展的瓶颈。机器学习技术凭借强大的数据处理和模式识别能力，为进一步扩展GEM提供了新思路。本综述系统总结了传统GEM及多约束多过程模型的构建流程，并着重探讨了机器学习在其中关键步骤中的应用前景，如基因功能注释、途径解析、空缺填补和生物学参数预测。机器学习技术作为新的驱动力，有望大幅度提升GEM的规模和质量，深化对生物代谢机制的理解，并推动实现数字孪生细胞。

关键词：基因组规模代谢模型；机器学习；合成生物学；代谢建模；多约束多过程模型

中图分类号：Q814.9　　**文献标志码：**A

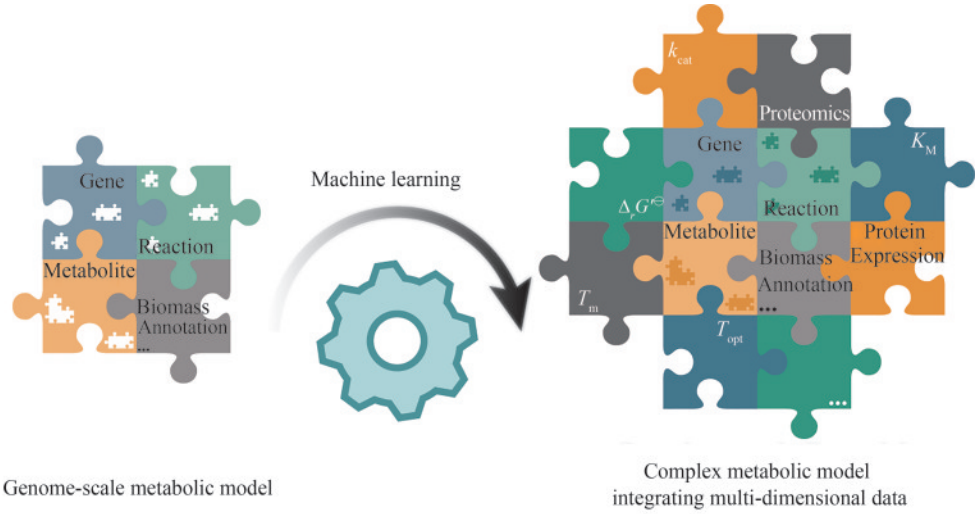
Applications of machine learning in the reconstruction and curation of genome-scale metabolic models

WU Ke[1,2], LUO Jiahao[1,2], LI Feiran[1,2]
([1]*Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, Guangdong, China*; [2]*Key Laboratory for Industrial Biocatalysis, Ministry of Education, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*)

Abstract: Since the publication of the first genome-scale metabolic model (GEM) in 1999, GEMs have become an essential tool for analyzing metabolism. The models integrate genes, metabolites, and reactions for combining stoichiometric matrices with constraint-based optimization to systematically describe and simulate metabolic processes in organisms. The development of automated pipelines for reconstructing GEMs has expanded their applicability to organisms from all kingdoms of life. Additionally, GEMs can integrate kinetic parameters, thermodynamic parameters, multi-omics data and multi-cellular processes to reconstruct more accurate models, thereby improving prediction

收稿日期：2024-12-02 修回日期：2025-02-12
基金项目：国家自然科学基金面上项目（22478223）
引用本文：吴柯，罗家豪，李斐然. 机器学习驱动的基因组规模代谢模型构建与优化[J]. 合成生物学，2025，6(3)：566-584
Citation：WU Ke，LUO Jiahao，LI Feiran. Applications of machine learning in the reconstruction and curation of genome-scale metabolic models[J]. Synthetic Biology Journal，2025，6(3)：566-584

accuracy. However, the reconstruction of GEMs remains heavily dependent on pre-existing knowledge, inherently limiting their scope to currently available information. This dependency restricts our ability to fully unravel the complexity and dynamic nature of metabolism. Recent advances in machine learning have demonstrated extraordinary capabilities for biological tasks such as protein structure prediction, disease identification and GEM reconstruction with functional annotation and large-scale data integration, showcasing its power in identifying patterns and uncovering hidden relationships within biological systems. Machine learning provides a promising pathway to overcome the limitations of GEMs by expanding their applicability to areas previously constrained by data availability and complexity. This review summarizes the traditional reconstruction methods of GEMs and their applications in integrating multi-dimensional data to build multi-constraint and multi-process models. The review also focuses on key applications of machine learning in gene function annotation, pathway analysis, gap-filling prediction in the reconstruction of GEMs. Additionally, the potential of machine learning in predicting kinetic, thermodynamic, and other key biochemical parameters in the reconstruction of multi-constraint and multi-process models is discussed. By combining GEMs with machine learning innovations, researchers can improve model accuracy, enhance scalability, and gain new insights into previously elusive metabolic mechanisms, bridging gaps in metabolic knowledge, and underscoring its importance as a cornerstone for future development in systems biology and biotechnology.



Keywords: genome-scale metabolic model; machine learning; synthetic biology; metabolic modeling; multi-constraint and multi-process model

代谢过程是细胞内所有生物化学反应的集合，涉及能量的转化、物质的合成与分解以及细胞功能的维持。理解这些代谢过程对于阐明生物体的功能机制、诊断疾病、设计药物、设计细胞工厂以及优化生物制造过程至关重要。然而，由于代谢过程的复杂性，传统的实验方法在解析整个代谢网络时面临诸多挑战。为了克服这些困难，代谢模型应运而生。基因组规模代谢模型（genome-scale metabolic model，GEM）是一种整合基因、代谢物及反应的系统性框架，用于全面描述生物体内的代谢网络及其化学计量关系[1]。该模型不仅能够借助优化算法预测代谢网络中的代谢流量[2]，还可结合热力学参数、动力学参数、多组学数据及多细胞过程构建满足多样化研究需求的多约束多过程模型[3-5]。自 1999 年发布首个流感嗜血杆菌的 GEM 以来，得益于大量完整基因组序列的测序和组装，目前已构建并发布了数千个物种的 GEM[6]。这些模型覆盖了细菌、酵母、植物和

动物等多种生物，成为研究生物代谢网络的重要工具，广泛应用于细胞表型解析、代谢工程策略开发以及治疗靶点识别等领域[7]。随着基因组测序技术和高通量组学技术的快速发展，GEM的性能和适用范围不断提升，这些进展极大促进了人们对生物体代谢机制的深入理解[6]。

然而，GEM依赖于现有科学知识的整合，其能力始终受限于当前已知信息。无论是代谢物、基因还是反应，即使是最先进的GEM，其知识覆盖范围仍存在不足，这在一定程度上限制了人们对生物代谢复杂性的全面认知。近年来，机器学习方法凭借其强大的模式识别与预测能力，已在大量学科得到广泛应用。尤其在生物学领域，机器学习技术已展现出显著优势，例如在蛋白质三维结构预测和抗体从头设计等领域，证明了其作为一种高效工具的潜力，能够基于已有知识深入探索未知空间[8-10]。此外，该技术还在与GEM构建和应用相关的酶注释[11]、途径挖掘[12-13]和细胞工厂设计[14-15]等方面展现出显著应用价值。本综述聚焦机器学习在GEM相关研究中的最新进展，重点探讨其在基因功能注释、代谢途径解析、代谢网络空缺填补，以及预测关键模型参数（如动力学、热力学和温度相关参数）中的应用。

1 基于先验知识的代谢模型

1.1 传统基因组规模代谢模型

传统GEM的构建过程通常包括两个关键模块：粗略模型生成和模型修正（图1）。在粗略模型生成阶段，主要任务是对基因组中的代谢基因注释功能以及代谢反应的组装；在模型修正阶段，主要任务是整合遗漏途径与填补代谢网络中的空缺，此外还需要执行验证GPR（gene-protein-reaction）关系、估算生物量组成、整合转运和交换反应等任务[16]。GEM的构建依赖于各种公共数据库，包括基因组数据库（如NCBI和KEGG[17]）、蛋白质数据库（如UniProt[18]、BRENDA[19]、Expasy[20]和SABIO-RK[21]）以及代谢反应数据库（如KEGG[17]、Rhea[22]、MetaCyc[23]和MetaNetX[24]）。此外，一些数据库（如BiGG[25]、BioModels[26]和KBase[27]）提供了已构建的GEM，可作为模型更新和新模型构建的模板。在模型构建完成后，必须经过一系列测试与评估，以确保其质量可靠性与形式规范性[28]。

由于GEM的构建涉及大量数据的整合，Thiele等[29]于2010年提出的GEM构建指南将GEM构建过程细化为96个步骤，但是这一过程烦琐且耗时。近年来新工具和自动化技术的发展显著加速了GEM构建过程。例如，COBRA Toolbox 3.0[30]和RAVEN 2.0[31]工具包提供了标准化平台，降低了模型构建和分析的学习成本。此外，ModelSEED[32]、CarveME[33]、Merlin[34]、gapseq[35]和AGORA2[36]等自动化工具可完成模型构建的大部分步骤，包括基因组注释、GPR关系生成、反应可逆性和酶定位区室预测等。这些方法已广泛应用于细菌、古细菌和真核生物的模型构建，其中AGORA2为人类肠道菌群中的7302个菌株构建了GEM，展示了其在肠道

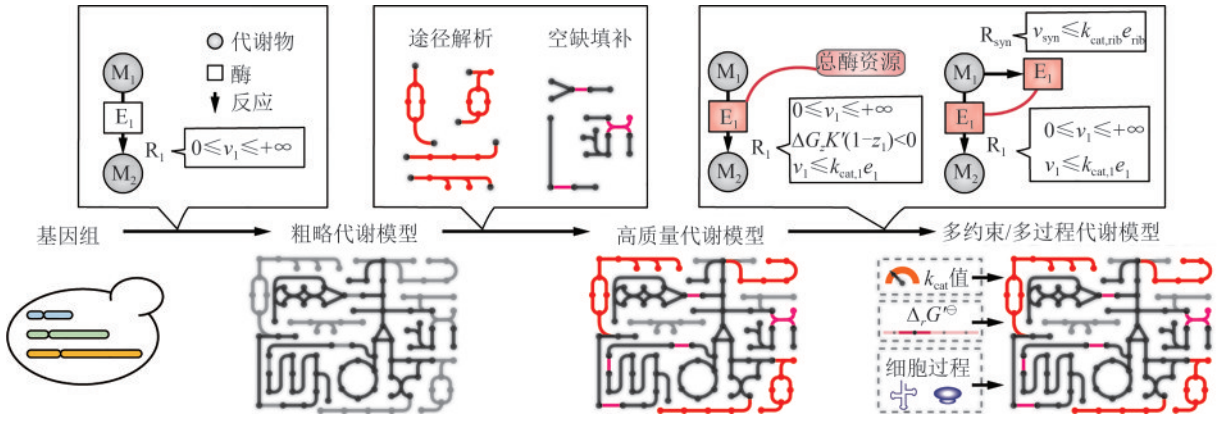


图1 GEM以及多约束多过程模型的构建过程

Fig. 1 Process for reconstructing GEMs and multi-constraint and multi-process models

微生物药物代谢预测中的潜力[36]。尽管自动化工具和数据库的快速发展已显著提高了模型构建的效率，但基因功能的精准注释，尤其是对未知蛋白功能的表征，仍对模型的全面性和准确性构成挑战。因此，如何进一步克服这些限制，持续改进和完善 GEM，依然是未来的关键任务。

1.2　多约束多过程模型

GEM 主要涵盖代谢途径的信息，通常不包含其他代谢层次的约束信息[37]。为克服这一局限，研究人员提出了更复杂的 GEM，例如多约束模型、多过程模型（图 1），以扩展代谢模型的应用范围[38]。多约束模型通常不会增加基因数目，而是通过增加约束条件，使反应通量分布更加合理。与此不同，多过程模型在 GEM 的基础上进一步扩展，加入其他细胞过程，如蛋白质表达和分泌等，并将这些过程耦合起来，因此模型中基因、反应和代谢物数目会显著增加。凭借更大的规模和更高的预测精度，这些模型在工业菌株的理性设计中显示出巨大潜力[39-40]。本文主要介绍不同类型多约束多过程模型及其构建所需的参数（图 2），具体的模型原理已在其他相关综述中进行了详细讨论[41]。

1.2.1　多约束模型

现有的多约束模型主要包括热力学约束、酶约束、酶热约束以及酶温度约束等。其中热力学约束可以避免代谢通量计算中的不可行循环[42]。热力学第二定律指出，正向反应的净通量对应于 Gibbs 自由能的负变化。热力学信息的整合可提升 GEM 的预测准确性[43]，同时为异源代谢途径的筛选与评估提供依据[44]。常用的热力学约束算法包括基于热力学的代谢通量分析（thermodynamics-based metabolic flux analysis，TMFA）[43]、最大-最小驱动力（max-min driving force，MDF）分析[45]以及 OptMDFpathway[46]。TMFA 在常规通量平衡分析（flux balance analysis，FBA）中引入热力学约束，将可逆反应拆分为正向和反向，利用 Gibbs 自由能数据来评估反应可行性[43]。MDF 方法用于在满足代谢物浓度约束的解空间内，寻找可使途径中热力学瓶颈反应的驱动力最高的浓度分布[47]。而 OptMDFpathway 框架进一步拓展了 MDF 方法的应用，利用混合整数线性规划（mixed-integer linear programming，MILP）在代谢网络中寻找热力学上可行的路径[46]。

酶约束模型（enzyme-constrained metabolic model，ecModel）是当前主流的多约束模型，其核心假设是细胞的蛋白质资源有限，需合理分配以确保各类生物过程的高效运转[48]。ecModel 通过在 GEM 中引入两大核心约束，进一步提高了代谢模型的性能。首先，反应通量受酶浓度限制，即每个代谢反应的速率受到参与该反应的酶浓度的限制。其次，细胞内总酶量有限，这意味着细胞必须在有限的蛋白质资源下合理分配，以确保不同的生物过程得以高效运行。这些约束使得模型能够更真实地反映细胞代谢过程，避免了传统 GEM 求解时出现的极端失真通量问题，从而在代

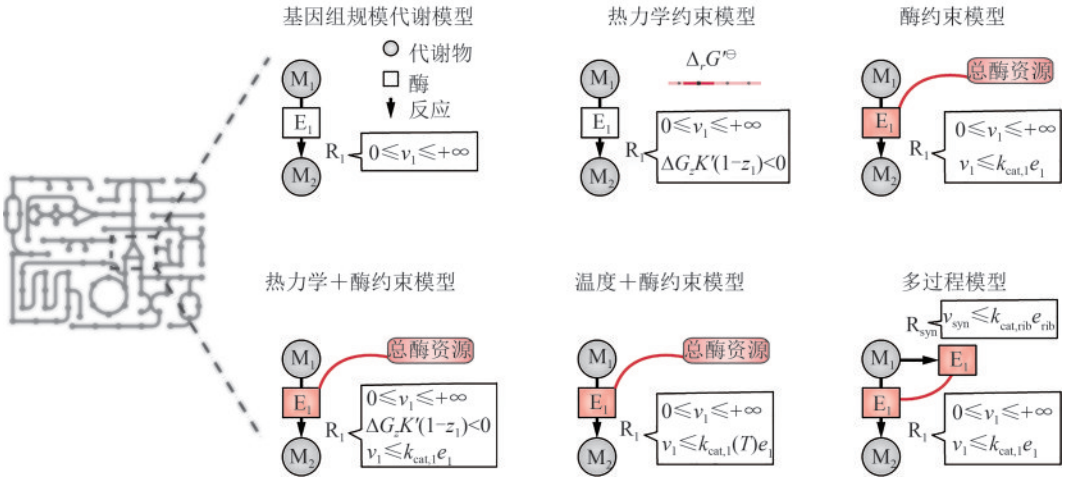


图 2　基因组规模代谢模型与多约束多过程模型的构建框架

Fig.2　Framework for constructing GEMs and multi-constraint and multi-process models

谢工程和生物制造领域中，提供更加精确的预测结果。ecModel的框架包括FBAwMC[49]、MOMENT[50]及其简化版sMOMENT[48]、ECMpy[51-52]和GECKO[53-55]等。细胞体积有限并对酶构成了空间限制[49]。在此基础上，MOMENT和sMOMENT进一步扩展了模型的精度和适用性。MOMENT引入了更复杂的优化目标，首次使用蛋白质质量分数约束酶浓度。sMOMENT作为MOMENT的简化版，减少了计算复杂度，使得大规模代谢网络的分析更加高效。此外，2017年发布以来，GECKO框架因其代码公开和更高的自动化程度，且能够整合蛋白质组数据，已经广泛用于酿酒酵母（*Saccharomyces cerevisiae*）等多种生物的ecModel构建[53-54]。ecModel的构建依赖于基因组模的酶活性参数。GECKO 3.0版本中引入了深度学习方法预测酶动力学参数，显著提升了模型的动力学参规模和预测精度[55]。GECKO框架已帮助数百种生物自动化构建ecModel，推动了精准代谢分析的发展，在代谢工程、疾病模型、药物开发等多个领域取得了显著进展[56]。除上述基础ecModel外，研究人员进一步结合其他约束条件扩展和优化模型，例如在酶约束基础上整合热力学约束[57]或温度约束[58]，适用于特定场景的代谢分析和表型预测。酶热约束模型依赖于酶活性参数和热力学参数，而酶与温度约束模型则通过整合蛋白质的最适温度（temperature optima，T_{opt}）和熔解温度（melting temperature，T_m）计算酶在不同温度下的酶活性参数，调整对模型的约束[58]。

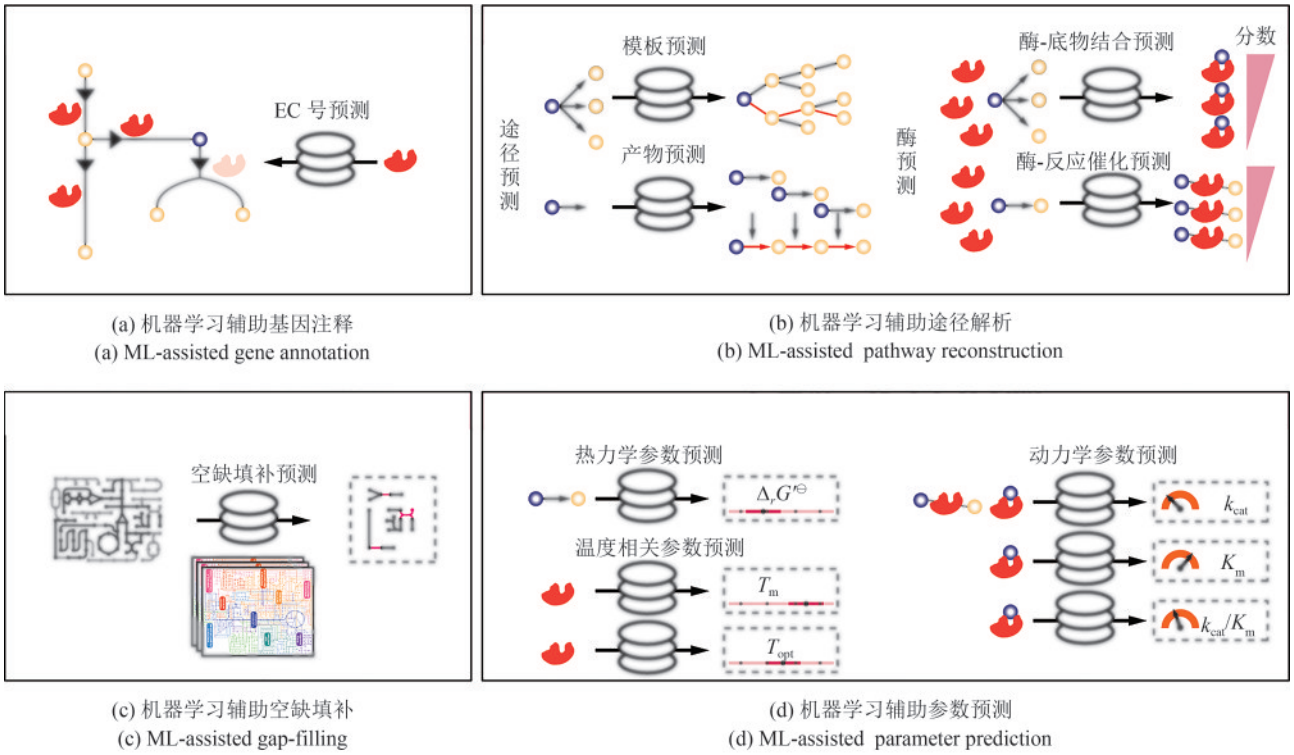
1.2.2 多过程模型

多过程模型通过在GEM中纳入更多细胞过程，整合多个层次的细胞代谢信息，考虑细胞过程之间的相互影响，更精确地反映细胞的真实状态。此类模型将代谢物（如氨基酸、ATP和GTP）作为大分子（如酶、mRNA、tRNA和核糖体）的合成原料，而大分子的合成过程也会限制代谢反应速率，从而描述胞内资源的约束和分配[41]。其中最常见的框架为代谢与表达（metabolism and expression，ME）模型和蛋白组约束模型（proteome constrained model，pcModel），两者的本质相同，主要是命名的差异。ME模型框架发展较早[59]，目前已针对热厌氧单胞菌（*Thermotoga maritima*）、大肠杆菌（*Escherichia coli*）、隆德克氏梭菌（*Clostridium ljungdahlii*）以及枯草芽孢杆菌（*Bacillus subtilis*）等生物开发了ME模型[60]。在基础模型框架外，研究者开发了FoldME[61]、AcidifyME[62]和OxidizeME[63]框架，分别扩展了热、酸和氧化应激细胞过程，并针对大肠杆菌开发了相应模型，描述了在不同胁迫下蛋白质的重新分配与保护机制。StressME框架进一步将3种细胞应激过程整合，针对大肠杆菌开发了相应模型，应用于模拟在多种环境胁迫下的生物学响应[64]。而pcModel发展稍晚，目前针对乳酸乳球菌（*Lactococcus lactis*）、酿酒酵母分别开发了pcLactis[65]和pcYeast[66]模型。在pcModel的基础上，针对酿酒酵母构建了pcSecYeast[67]以及CofactorYeast[68]，这两种模型框架分别整合了蛋白分泌过程和金属辅因子，扩展了多过程模型的细胞过程覆盖度和模拟适用范围。此外，ETFL（expression and thermodynamics flux framework）模型框架[69]，在蛋白质表达过程的基础上，整合了热力学约束，并采用MILP取代传统模型中使用的迭代线性规划（linear programming，LP），显著提升了模型的预测能力[69]。目前，ETFL框架已被用于构建大肠杆菌[69]和酿酒酵母[70]的模型。相较于已覆盖上百个物种的ecModel，多过程模型的物种覆盖度较低。这一局限主要源于其较高的复杂性和对参数的依赖。一方面，多过程模型需要整合更多的细胞过程，其所需的参数和过程信息尚未被充分表征；另一方面，目前仍缺乏成熟的自动化建模框架，使得构建此类模型的难度进一步增加。具体来说，ME模型和pcModel的参数输入主要包括酶活性参数、酶复合体的组成以及各蛋白质组分的计量系数。而ETFL模型在此基础上引入了热力学约束，因此还需输入热力学参数。

2 机器学习辅助传统基因组规模代谢模型构建

尽管GEM在过去取得了显著进展，但仍面临许多挑战，这些问题限制了GEM的准确性和全面

性。近年来，机器学习在基因注释、途径解析及空缺填补等方面发挥了重要作用，本文将从这些方面展开讨论（图 3），表 1 汇总了其中与机器学习相关的方法及方法介绍。



(a) 机器学习辅助基因注释
(a) ML-assisted gene annotation

(b) 机器学习辅助途径解析
(b) ML-assisted pathway reconstruction

(c) 机器学习辅助空缺填补
(c) ML-assisted gap-filling

(d) 机器学习辅助参数预测
(d) ML-assisted parameter prediction

图 3　机器学习辅助基因组规模代谢模型与多约束多过程模型构建

Fig. 3　Machine learning-aided reconstruction of GEMs and multi-constraint and multi-process models

表 1　机器学习辅助基因组规模代谢模型的方法扩展

Table 1　Machine learning assisted expansion of GEMs

模型应用	方法	模型框架	输入	输出	特点
辅助基因注释	DeepEC[71]	CNN	氨基酸序列	EC 编号	可区分酶与非酶，无法进行多功能注释
	CLEAN[11]	预训练蛋白质大语言模型（ESM-1b），对比学习	氨基酸序列	EC 编号	无法区分酶与非酶，可进行多功能注释，可用于数据极少的 EC 编号
	DeepECtransformer[72]	预训练蛋白质大语言模型（ProtBert），Transformer	氨基酸序列	EC 编号	可区分酶与非酶，可进行多功能注释，不可用于数据极少的 EC 编号
	ECRECer[73]	预训练蛋白质大语言模型（ESM-1b），GRU	氨基酸序列	EC 编号	可区分酶与非酶，可进行多功能注释
	ECPICK[74]	One-hot，CNN	氨基酸序列	EC 编号	无法区分酶与非酶，不可进行多功能注释，可输出预测 EC 编号的置信度
	EnzBert[75]	Transformer	氨基酸序列	EC 编号	无法区分酶与非酶，不可进行多功能注释，可推测关键残基
	EnzymeNet[76]	CNN，ResNet	氨基酸序列	EC 编号	可区分酶与非酶，无法进行多功能注释
	GraphEC[77]	预训练蛋白质大语言模型（ProtTrans），ESMFold	氨基酸序列和蛋白质三维结构	EC 编号	无法区分酶与非酶，可进行多功能注释，计算资源需求高

续表

模型应用	方法	模型框架	输入	输出	特点
辅助途径解析-反应预测	RetroPath RL[78]	基于蒙特卡洛树搜索的强化学习方法	化合物SMILES	合成途径	缓解组合爆炸问题，允许探索深度是只基于模板版本的两倍以上
	ASKCOS[79]	FNN	化合物SMILES	合成途径	缓解组合爆炸问题，适用于化学合成途径设计
	chemoenzymatic-ASKCOS[80]	DNN	化合物SMILES	合成途径	缓解组合爆炸问题，可进行化学合成与生物合成混合途径设计
	RetroBioCat[12]	DNN	化合物SMILES	合成途径	缓解组合爆炸问题，处理大分子化合物存在挑战
	Kreutter等开发的方法[81]	Transformer	化合物SMILES和酶功能描述	反应产物	预测单步反应，无法给出完整的反应，无法推荐酶功能
	Probst等[82]	Transformer	化合物SMILES和EC编号	反应产物	预测单步反应，无法给出完整的反应，无法推荐EC编号
	BioNavi-NP[83]	Transformer	天然产SMILES	合成途径	针对天然产物及类似物，可预测多步途径
	BioNavi[84]	Transformer	化合物SMILES	合成途径	可进行化学合成与生物合成混合途径设计
辅助途径解析-酶挖掘	ESP[85]	预训练蛋白质大语言模型（ESM-1b），GNN，XGBoost	氨基酸序列和分子指纹	酶-底物结合可能性	对于训练集中没出现代谢物的预测性能会有明显下降
	EnzRank[86]	分子指纹，CNN	氨基酸序列和分子指纹	酶-底物结合可能性	对于天然底物及其相似物具有良好的区分能力
	PU-EPP[87]	GNN，正样本和无标签学习	氨基酸序列和化合物SMILES	酶-底物结合可能性	鲁棒性强，可鉴定酶和底物的关键位点
	MEI	预训练蛋白质大语言模型（ESM-1b），GNN，DNN	氨基酸序列和化合物SMILES	酶-底物结合可能性	可利用专业数据集微调进行特定任务预测
	REME[88]	集成ESP、DLKcat/TurNuP、DeepET等模型	反应SMILES	酶列表	多维度评价与筛选有效提高了推荐酶列表的可信度
	SPEPP[89]	Word2Vec，Transformer	底物、反应物和酶	酶催化底物-产物反应的可能性	计算效率相较基于相似性的方法显著提高，可大规模使用，需要提供候选酶集合
辅助空缺填补	BoostGAPFILL[90]	基于矩阵分解的推荐系统技术	代谢模型	填补反应	融合了拓扑和约束方法，能够识别代谢网络中的潜在模式
	CHESHIRE[91]	GCN	代谢模型	填补反应	计算效率高，可解释性强，可能引入假阳性反应，缺少反应方向性信息
	DSHCNet[92]	GCN，MLP	代谢模型	填补反应	对反应数据依赖较强，在适应不同反应数据库中存在挑战
	DNNGIOR[93]	CNN	代谢模型	填补反应	预测性能受系统发育距离影响

注：CNN—卷积神经网络；FNN—前馈神经网络；DNN—深度神经网络；GNN—图神经网络；GCN—图卷积神经网络；MLP—多层感知机。

Note: CNN—Convolutional neural network; FNN—Feedforward neural network; DNN—Deep neural network; GNN—Graph neural network; GCN—Graph convolutional network; MLP—Multilayer perceptron.

2.1 机器学习辅助基因功能注释

GEM的构建依赖于基因功能注释数据，许多未知功能的基因限制了GEM中基因和反应的规模。传统的基因功能预测方法通常依赖于在大型注释数据库中识别相似或同源序列（如BLASTp和HMMER）。在基因功能注释中，GO[94]和EC编号被广泛应用。然而，由于GO无法直接关联具体的代谢反应，无法直接应用于代谢建模，因此本节重点讨论了基于EC编号的基因功能预测方法[图3(a)]。这一类方法通过训练机器学习或者深度学习模型预测目标蛋白的EC编号，相关的方法

有 EzyPred[95]、SVM-prot[96]、DEEPre[97]、DETECT v2[98]和ECPred[99]。然而，这些早期方法面临着数据的局限。例如，DETECT v2[98]和ECPred[99]使用的数据集分别仅涵盖786和858种EC编号。ExPASy目前已经收录超过8000个EC编号，以上方法难以胜任基因组规模的基因功能注释任务，因此这些方法未在表1中进行总结[20]。

随着数据的日益丰富，DeepEC应运而生。DeepEC利用了来自UniProt的数据集，其中包含4669种EC编号，采用多任务模型架构，兼顾酶与非酶的分类预测以及EC编号的预测，在未出现在训练数据中的通过文献搜集的201个酶序列集合上预测精度为0.92，召回率为0.45[71]。人工智能技术的不断发展，包括蛋白质大语言模型ESM系列模型[100]和蛋白质三维结构预测模型[8, 10, 101]的出现，也使得基因功能注释模型的性能得到了进一步的提升。CLEAN采用蛋白质的预训练模型ESM-1b生成嵌入向量表示，并引入对比学习框架，通过计算查询序列与每个EC编号嵌入向量的欧氏距离预测，有效缓解了EC编号分布不均的问题，能够高质量地注释研究较少的酶、纠正错误标记的酶以及识别具有两个或更多EC号的功能混杂酶。该方法成功注释了36种此前未注释的卤化酶，并通过体外实验验证了其高准确性[11]。随后发表的DeepECtransformer，作为DeepEC的升级版，引入蛋白质大语言模型ProtBert以强化蛋白质表征，并通过移除数据集中少于100条序列的EC编号条目缓解了极端类别数据稀缺的问题，成功预测并表征了大肠杆菌中3种酶的功能（YgfF、YciO和YjdM）[72]。此后发表的多个模型（例如基于分层次双核多任务学习架构的ECRECer[73]、基于证实深度学习的ECPICK[74]、基于Transformer架构的EnzBert[75]，以及基于残差神经网络的EnzymeNet[76]）的性能也都优于同源比对方法或早期的机器学习方法，但这些方法在发布时没有与CLEAN以及DeepECtransformer进行基准测试。随着蛋白质结构预测能力的显著提升，GraphEC首次在EC编号预测任务中整合了蛋白质三维结构信息，通过ESMFold[101]预测蛋白质结构，并结合ProtTrans生成的序列嵌入提取蛋白质特征。GraphEC通过引入酶的活性中心进行EC编号预测，在两个独立测试集上取得了现有方法中最高的召回率和AUC[77]。但是，由于GraphEC数据集中没有非酶标签，因此与CLEAN存在相似的缺陷，可能为非酶的蛋白质序列分配EC编号，在使用此类方法辅助GEM构建时，需要区分假阳性结果。利用各类EC编号预测模型对生物基因进行预测，并通过在KEGG等生化反应数据库中查询EC编号对应的反应，即可实现基因功能注释，对GEM传统构建流程中的基因功能注释环节起到完善或替代作用。

2.2 机器学习辅助途径解析

构建高质量GEM的另一个瓶颈是其代谢网络中许多代谢物的合成或降解途径尚未被完全解析。这些代谢物难以与现有的代谢网络连接，限制了GEM在代谢物和反应规模上的扩展[102]。这一问题主要源于两个方面：首先，酶底物的混杂性问题，即一类酶能够催化多种在化学结构上与其标准底物相似的反应物[103-104]；其次，存在尚未被揭示的酶反应机制。

酶的混杂性长期以来未引起充分重视[105]。研究显示，大肠杆菌中约37%的酶对与其主要底物结构相似的底物表现出混杂活性[106]。为扩展反应空间，基于模板的逆合成方法被广泛应用于代谢途径预测。这类方法依赖从生化反应数据库中构建的反应模板库，通过匹配目标产物与适当模板，推测可能的反应物并生成相关反应路径[图3(b)]。已发表的基于模板的逆合成方法包括PathPred[107]、RetroPath 2.0[108]、novoPathFinder[109]、RetroPath RL[78]、AiZynthFinder[110]、ASKCOS[79]和chemoenzymatic-ASKCOS[80]，在代谢途径挖掘中展现出重要的应用价值。然而，这些方法常面临组合爆炸的问题[78]。为应对此问题，在部分方法中引入了机器学习方法，优化反应模板的选择，有效限制组合爆炸范围，并显著提升路径预测的效率与准确性[111]。例如ASKCOS使用了前馈神经网络（feedforward neural network，FNN）预测与目标分子最相关的反应模板，减少假阳性反应数目和计算成本，确保预测途径的可行性，以最大化实验成功的可能性，但是ASKCOS只包含化学合成的模板[79]。chemoenzymatic-ASKCOS额外整合了酶反应的模板实现混合合成路径设计[80]。

RetroBioCat使用基于深度神经网络（deep neural networks，DNN）训练的SCScore[112]方法，通过对每个分子的复杂度进行评分，指导逆合成搜索选择更简单的起始分子[12]。RetroPath RL使用基于蒙特卡洛树搜索的强化学习方法来选择最佳途径，探索深度比RetroPath 2.0提升1倍以上[78]。

为了突破对已知反应模板的依赖，探索未知的反应机制，一些研究开发了无模板的反应预测方法，这些方法直接预测反应物或生成物[图3（b）]。一个典型的例子是Kreutter等[81]将USPTO数据集作为一般化学知识的来源，并将其迁移至从文献中收集的数千条酶催化反应数据，训练了一个基于反应文本表示的序列到序列预测模型，该模型通过将输入的反应物SMILES转换为生成物SMILES，实现对反应产物的预测，但是这个方法只考虑了正向预测的情况。Probst等[82]在SMILES作为输入的基础上，整合了反应EC编号的输入来提升模型性能，还实现了逆合成的拓展。Zheng等[83]基于Transformer架构开发的BioNavi-NP模型在有机反应数据和酶反应数据上训练，成功预测了多种天然产物的合成路径。Zeng等[84]基于BioNavi-NP提出了改进版的BioNavi，通过在深度学习模型中引入多任务学习和反应模板，以更为直观和可解释的方式设计混合合成路径。无模板的方法为探索未知反应机制提供了新的路径，有效推动代谢网络和酶催化反应的扩展。这些方法都为挖掘未知反应提供了有效工具，可以进一步拓展反应空间，提升模型的覆盖范围与准确性，为构建完善的GEM奠定基础。

途径解析过程中，新预测的生化反应尚未分配EC编号，无法通过现有EC编号预测模型进行酶注释，这使得将这些反应与特定基因对应成为挑战，进而影响了GEM中GPR关系的准确性[113]。这一任务不仅对于天然代谢途径的酶挖掘至关重要，对于开发新型非天然代谢途径同样具有重要意义[111]。目前，新反应的酶注释通常基于序列同源性及反应相似性等特征。例如，Selenzyme[113]、EC-BLAST[114]、RxnSim[115]以及SelenzymeRF[116]都是通过计算目标反应与已知酶注释的反应之间的相似性识别可能的酶催化反应[117]，其中，Selenzyme通过基于参与反应的所有化合物的完整结构来计算反应相似性，而EC-BLAST、RxnSim则通过反应的分子指纹来计算。SelenzymeRF作为Selenzyme的更新版本，通过引入sim_RF算法，并结合RXNMapper[118]实现反应中的原子映射标注进一步优化了推荐酶的能力[116]。然而，需要指出的是，序列相似性与酶的功能相似性并不总是严格对应，因而仅依赖序列相似性有时不足以准确预测酶的催化能力。

在这一背景下，涌现了许多酶-底物结合预测和酶-反应催化预测的深度学习模型来辅助酶挖掘[图3（b）]。传统的酶-底物结合预测方法依赖分子动力学模拟，尽管其精确性较高，但计算量庞大且耗时。然而，随着合成生物学和计算生物学的快速进展，酶反应的实验数据逐渐积累，机器学习方法逐步成为一种高效的替代方案。ESP模型[85]采用ESM-1b和图神经网络（graph neural network，GNN）分别表征蛋白质序列和小分子，再结合这两种特征表示，在约18 000对经过实验验证的酶-底物数据集上训练梯度提升决策树模型，用于酶-底物结合预测。另一个例子是EnzRanK[86]，其依赖卷积神经网络（convolutional neural network，CNN）获取酶序列和底物的特征，并生成结合概率评分。基于此评分，模型能够排序候选酶，识别出在新型底物上可能具有活性的酶。在酶-底物结合预测任务中，正样本可以从反应数据库中轻易获取，但是负样本的获取具有不确定性。一方面，直接将酶与非已知底物组合会导致正负样本数据失衡；另一方面，随机抽取的酶-底物对可能由于底物混杂性而实际属于正样本。这会导致模型的预测出现偏差。因此，PU-EPP模型中提出了一种结合正样本和无标签学习的策略，以最大限度减少不准确负样本的影响。PU-EPP成功鉴定了15种对赭曲霉毒素A和玉米赤霉烯酮具有特异性的降解酶[87]。最近，Qian等[119]使用ESM-1b表征蛋白质序列、GNN表征小分子，结合DNN训练了MEI模型，在从文献中手动收集的两个测试集上表现优于ESP。并且在MEI的基础上，利用专业数据集分别针对CYD抑制剂与底物预测任务与塑料降解酶预测任务进行了微调，均能达到特定领域预测模型的先进水平。这些方法的结合有效提高了酶-底物结合预测的准确性和筛选效率，可以帮助确定酶在代谢途

径中的功能，为GEM构建提供更准确的GPR关系。与此同时，Hu团队[89]开发了深度学习模型SPEPP，可以预测酶-底物-产物三元组发生反应的可能性。其优势在于克服了传统方法中反应相似性计算的局限，更加侧重于底物与产物之间的关系，拓宽了潜在酶催化反应的发现范围。此外，Shi等[88]开发了一个综合平台REME来辅助酶挖掘和评估。该平台通过整合原子映射、原子类型变化以及基于分子指纹的反应相似性计算，实现与非天然反应相似的已知反应的快速排序并获取候选酶。REME还参考了酶-底物结合模型和动力学参数预测模型的预测结果，有助于在筛选和评估过程中迅速识别出候选酶。

2.3 机器学习辅助模型空缺填补

除了基因功能注释和途径解析，GEM修正过程中的空缺填补环节也受益于机器学习的快速发展［图3(c)］。GEM是基于目标生物体中所有已知的代谢反应和基因构建的。然而，由于对生物体的认识不完善，这些网络重建中常常存在空缺（gap）[120]。

为解决这一问题，Thiele等[121]开发了fastGapFill算法，这是第一个能够高效检测并填补GEM网络缺口（gap-filling）的工具。该算法通过从通用生化反应数据库中筛选最小反应集合，将其整合到GEM中，使原本的死端反应可以具有通量。此外，Kumar等[122]开发了GapFill算法，此方法以GEM中无法生成的代谢物为目标，改变反应方向或从MetaCyc数据库添加反应，将这些代谢物与现有底物连接起来。上述基于网络拓扑结构的空缺填补算法依赖于候选反应数据库，并且最小化候选反应集合的优化方式缺乏相应理论依据。随着机器学习的发展，也出现了新的空缺填补方法。Oyetunde等[90]提出的BoostGAPFILL利用基于矩阵分解的机器学习和整数最小二乘优化方法填补代谢网络中的缺失反应，为空缺填补开辟了新的途径。Chen等[91]于2022年提出了CHESHIRE（CHEbyshev Spectral HyperlInk pREdictor），该算法通过引入超图和CNN的概念，基于代谢网络的拓扑特征预测GEM中的缺失反应，仅需输入GEM，即可生成候选反应的置信评分，有助于快速识别GEM中的关键缺失反应并增强表型预测。Huang等[92]开发的DSHCNet进一步考虑代谢反应的异质性，并在表示GEM的超图的顶点表示中明确区分底物和产物，基于超图结构实现GEM的空缺填补。此外，针对宏基因组组装过程中基因组不完整性导致的GEM空缺问题，Boer等[93]提出了一种名为DNNGIOR的深度神经网络引导的反应集合推测方法。该方法能够基于不完整的反应集合预测潜在缺失的反应，补全GEM，但是DNNGIOR的预测性能受到代谢反应在细菌群体中的出现频率以及查询基因组与训练基因组之间系统发育距离的显著影响。

3 机器学习辅助多约束多过程模型构建

多约束多过程模型通过整合多个层次的细胞过程及其相互作用，进一步提升了对复杂生物系统的描述能力。这类模型通常更加复杂，所需的参数显著增加，包括酶动力学、热力学等多种数据。然而，当前实验技术解析的生物学参数有限，难以全面覆盖模型所需数据，成为了多约束多过程模型快速发展的瓶颈。为应对这一问题，许多基于机器学习的研究致力于这些参数的预测，从而扩展代谢模型的规模和质量，提高模拟精度，使模拟结果更贴近细胞的真实状态［图3(d)］[16]。表2汇总了其中与机器学习相关的方法及方法介绍。

3.1 动力学参数预测

酶动力学参数在解析细胞代谢机制、蛋白质组分配及生理多样性中具有关键作用。其中，k_{cat}（催化常数）定义了酶促反应的最大化学转化速率，K_m（米氏常数）反映了酶达到最大催化速率一半时所需的底物浓度，k_{cat}/K_m则衡量酶的整体催化效率。在酿酒酵母的ecModel模型中，仅有约5%的酶促反应能够在BRENDA和SABIO-RK数据库中找到完全匹配的k_{cat}[142]。为支持ecModel的进一步扩展，研究者开发了多种动力学参数预测方法，以扩大其覆盖范围。现有k_{cat}预测方法包括Heckmann等[123]开发的方法、DLKcat[124]、TurNuP[125]、DLTKcat[126]和DeepEnzyme[127]，K_m

预测方法包括Kroll等[128]开发的方法、GraphKM[129]和MLAGO[130]，以及兼顾了k_{cat}、K_m预测的MPEK[131]，兼顾了k_{cat}、K_m和k_{cat}/K_m预测的UniKP[132]和EITLEM-Kinetics[133]。

表2 机器学习辅助多约束多过程模型获取参数方法

Table 2 Machine learning assisted obtaining of parameters for multi-constraint and multi-process models

参数类型	方法	模型框架	输入	输出	特点
动力学参数	Heckmann等开发的方法[123]	随机森林，MLP	GEM/蛋白质结构/EC号首位/pH等	k_{cat}	可预测体内k_{cat}，适用于大肠杆菌
	DLKcat[124]	GNN，CNN	氨基酸序列和化合物SMILES	k_{cat}	预测k_{cat}（R^2=0.44），内置于GECKO 3.0
	TurNuP[125]	预训练蛋白质大语言模型（ESM-1b），XGBoost	氨基酸序列和反应指纹	k_{cat}	预测k_{cat}（R^2=0.44），无法区分多底物反应中不同底物的k_{cat}
	DLTKcat[126]	GNN，CNN	氨基酸序列，化合物SMILES和温度	k_{cat}	预测不同温度下的k_{cat}（R^2=0.66）
	DeepEnzyme[127]	GCN	氨基酸序列，化合物SMILES和蛋白质三维结构	k_{cat}	预测k_{cat}（R^2=0.58），预测突变型需要具备蛋白质结构预测能力
	Kroll等开发的方法[128]	预训练蛋白质大语言模型（ESM-1b）	氨基酸序列和分子指纹	K_m	预测K_m（R^2=0.53）
	GraphKM[129]	预训练蛋白质大语言模型（ESM-2），GNN	氨基酸序列和化合物SMILES	K_m	预测K_m（R^2=0.62），模型的预测性能受限于训练数据集的规模和质量
	MLAGO[130]	随机森林	EC编号，KEGG ID和物种编号	K_m	预测K_m（R^2=0.53），泛化能力受限于EC编号、KEGG ID、物种编号信息
	MPEK[131]	预训练蛋白质大语言模型（ProtT5），预训练小分子大语言模型（Mole-BERT），多任务学习	氨基酸序列和化合物SMILES	k_{cat}和K_m	支持同时预测k_{cat}（R^2=0.64）和K_m（R^2=0.60）
	UniKP[132]	预训练蛋白质大语言模型（UniRef50），预训练小分子大语言模型（SMILES Transformer），极度随机树	氨基酸序列和化合物SMILES	k_{cat}、K_m和k_{cat}/K_m	支持分别预测k_{cat}（R^2=0.67）、K_m（R^2=0.60）和k_{cat}/K_m（R^2=0.56），鲁棒性强，支持温度和pH输入
	EITLEM-Kinetics[133]	预训练蛋白质大语言模型（ESM-1v），迁移学习	氨基酸序列和化合物SMILES	k_{cat}、K_m和k_{cat}/K_m	支持分别预测k_{cat}（R^2=0.72），K_m（R^2=0.69）和k_{cat}/K_m（R^2=0.68），蛋白突变体的k_{cat}预测性能优异
热力学参数	dGPredictor[134]	线性回归模型	分子指纹	$\Delta_r G'^{\ominus}$	不适用于异构化反应与涉及金属或聚合物结构的反应
	Alazmi等开发的方法[135]	线性回归模型	分子指纹	$\Delta_r G'^{\ominus}$	特征提取方式通用性强，可用于非天然反应
温度相关参数	DeepSTABp[136]	预训练蛋白质大语言模型（ProtTrans），MLP	氨基酸序列、生物体生长温度和实验条件	T_m	预测能力对点突变不敏感
	DeepTM[137]	GCN	氨基酸序列	T_m	特征提取复杂，训练数据未考虑其他对蛋白熔解温度影响的因素

续表

参数类型	方法	模型框架	输入	输出	特点
温度相关参数	Tome[138]	SVR，随机森林	氨基酸序列，OGT	T_{opt}	训练集高于85 °C的T_{opt}值占比不足5%，限制了Tome对高温稳定性酶的预测能力
	TOMER[139]	集成学习	氨基酸序列，OGT	T_{opt}	重采样缓解了数据分布不平衡，对高于85 ℃的T_{opt}值预测性能显著提升
	DeepET[140]	ResNet，迁移学习	氨基酸序列	T_m和T_{opt}	类似于Tome，数据分布不均衡会限制其对极端温度蛋白质的预测性能
	Preoptem[141]	One-hot，CNN	氨基酸序列	T_{opt}	Pearson相关系数r=0.58，适用于嗜热蛋白

注：MLP—多层感知机；GNN—图神经网络；CNN—卷积神经网络；GCN—图卷积神经网络；SVR—支持向量回归；OGT—最适生长温度。

Note: MLP—Multilayer perceptron; GNN—Graph neural network; CNN—Convolutional neural network; GCN—Graph convolutional network; SVR—Support vector regression; OGT—Optimal growth temperature.

一些方法已成功应用于GEM模型优化与合成生物学领域。例如，DLKcat是首个专注于大规模预测k_{cat}的深度学习工具，成功预测了343种酵母和真菌的k_{cat}，并辅助构建了相应的ecModel[124]。作为k_{cat}预测的关键模块，DLKcat已被整合到酶约束模型自动化构建流程GECKO 3.0[55]和ECMpy 2.0[52]中。TurNuP[125]进一步推动了k_{cat}值预测的发展，特别是在应对与训练集序列相似性较低的酶时表现出卓越的鲁棒性，并在基于ecModel的蛋白质组预测中展现出更高的准确性。此外，DLTKcat支持在指定温度下预测酶的k_{cat}，而UniKP进一步扩展了功能，能够在指定温度和pH条件下预测k_{cat}，满足实际应用中对环境参数的需求。UniKP和EITLEM-Kinetics在预测酶突变体的k_{cat}方面同样表现优异，其中UniKP通过k_{cat}和k_{cat}/K_m预测指导蛋白质突变，获得了高活性的酪氨酸解氨酶（TAL）。相比之下，K_m的预测方法尚未开发出适合的下游应用。除酶约束模型之外，动力学模型中也需要k_{cat}和K_m参数，随着酶活性预测模型的发展，这些预测方法也会在未来发挥出重要作用。

3.2 热力学参数预测

热力学分析的核心在于精准预测反应的标准Gibbs自由能变化（$\Delta_rG'^{\ominus}$）[143]。截至2019年，酶促反应热力学数据库（TECRDB）中仅收录了约600个酶促反应的实验测量热力学数据。为弥补实验数据的不足，研究者开发了基于基团贡献（group contribution，GC）的方法[144-146]，并将其应用于代谢模型的构建[57]。然而，基团贡献方法存在一些内在局限性，包括：①专家定义的基团覆盖范围有限，导致某些代谢物无法被分解，进而无法估算其$\Delta_rG'^{\ominus}$；②对于无基团变化的反应（如异构反应），$\Delta_rG'^{\ominus}$被赋值为零，而实验数据表明其实际值非零[134]。

Alazmi等[135]提出了一种基于化学指纹特征的机器学习算法——指纹贡献（fingerprint contribution，FC）方法，用于预测生化反应的$\Delta_rG'^{\ominus}$。该方法以二维化学指纹表示化合物特征，主要分为两个步骤：首先，从大量二维指纹特征中系统筛选相关特征；其次，利用正则化回归方法构建最终的线性预测模型。此外，Wang等[134]开发了dGPredictor工具，该工具能够考虑代谢物结构中的立体化学信息，显著提升反应的覆盖率。dGPredictor还可预测新反应的$\Delta_rG'^{\ominus}$，并可以集成至代谢途径从头设计工具，以避免在设计过程中引入方向性不可行的反应步骤。

3.3 温度相关参数预测

蛋白质作为一种执行大多数催化功能且对温度变化高度敏感的生物大分子在细胞中起着至关

重要的作用[147]。然而，整合温度的GEM建模仍面临挑战，主要原因包括代谢的复杂性以及缺乏足够带有温度信息的蛋白质属性数据[58]。为了更好地模拟和预测温度对细胞代谢的影响，尤其是蛋白质在不同温度下的行为，目前已开发多种方法来预测蛋白质与温度相关的参数，包括T_m和T_{opt}。

早期的T_m预测方法通常采用one-hot编码、氨基酸组成等简单特征来表征蛋白质，然后使用传统机器学习模型预测T_m。随着深度学习技术的发展，研究逐渐采用了更复杂的模型。例如，Jung等[136]开发了基于Transformer的DeepSTABp算法，该算法结合了热蛋白组分析的实验条件、氨基酸序列和宿主的最适生长温度（optimal growth temperatures，OGT）来预测T_m。此外，Li等[137]提出了模型DeepTM，利用图卷积神经网络（graph convolutional neural network，GCN）和自注意力网络，通过序列信息直接预测蛋白质的T_m。该方法的蛋白质表征结合了宿主的OGT、进化信息、理化特性（包括疏水性、体积、极化率、等电点等）等特征。

此外，机器学习模型Tome能够基于蛋白质氨基酸序列预测T_{opt}[138]。预测的T_{opt}进一步应用于贝叶斯基因组规模代谢模型（Bayesian-GEM）中，以模拟温度对酿酒酵母细胞代谢的影响。研究揭示，当培养条件超过最适温度时，导致酵母生长受限的最关键限速酶是鲨烯环氧化酶（ERG1）。通过用耐热酵母菌株中的同源酶替代ERG1，获得了生长能力超过野生型的耐热菌株[58]。值得注意的是，在Tome的训练数据集中，T_{opt}小于85 ℃的样本占比超过95%，导致其在预测高温稳定酶时的能力不足。为了解决问题，Gado等[139]提出了TOMER模型。TOMER在Tome模型的数据集上进行了重新训练，并通过重采样和集成学习策略缓解了数据不平衡问题，使得在测试集中T_{opt}大于85 ℃的数据预测R^2分数从0.52提升到0.63。为了更好地捕捉蛋白质序列与温度之间的关系，Li等[140]利用包含300万种酶的OGT数据集训练了DeepET模型，并通过微调DeepET进而预测T_{opt}和T_m。DeepET的微调模型预测仅依赖蛋白质序列，消除了此前T_{opt}预测方法对OGT参数的依赖。类似的基于蛋白质序列的T_{opt}预测方法还包括Preoptem，该方法采用one-hot编码表示蛋白质序列，并结合CNN进行训练。Preoptem助力从海洋宏基因组中挖掘到一种新的嗜热几丁质酶，展示了其在实际应用中的潜力[141]。

4 挑战与展望

机器学习在GEM中的应用正逐步展现出巨大的潜力，尤其是在扩展GEM规模和质量、提升GEM细胞表型预测精度等方面。然而，尽管机器学习为GEM提供了更为强大的工具，当前常用的评估标准在应用到生物学领域时仍面临一定的局限性。常见的机器学习模型评估标准，如AUC、准确率和独立数据集的使用，通常受到数据分布的影响，容易导致高估模型性能和未能准确反映模型泛化能力的问题。这种问题在处理基因组规模的预测时尤为突出，过于依赖机器学习领域的评估标准可能忽视机器学习模型的普适性和多样性，特别是在零样本或少样本的预测场景下，可能导致对模型性能的误判。因此，未来研究需要开发更加完善的评估框架，以准确反映机器学习模型在GEM中的实际应用价值。

此外，当前机器学习方法的应用仍然局限于传统机理代谢模型的框架内。机器学习更多地扮演着辅助角色，例如辅助基因功能注释、途径解析和参数预测等，而未能突破传统机理代谢模型的限制。许多细胞过程，尤其是那些难以精确数学表达的过程，仍然是当前GEM发展的挑战。因此，要充分发挥机器学习的潜力，未来的研究需要开发机理模型与数据驱动方法的深度融合框架，推动更多细胞过程的数学解析，实现对更复杂生物学系统的精准描述。通过将机器学习的多种模块有机结合，GEM不仅能够整合多尺度的生物学约束，还能够推动白箱与黑箱模型的融合，从而提升GEM的规模和性能。随着机器学习算法的进步，AI驱动的GEM自动化建模将逐步成为现实，这将大大提高其精度、速度和可靠性，推动其进入AI时代，进一步推动数字孪生细胞的实现。

然而，数字孪生细胞的发展要求跨学科的深度合作。随着合成生物学和计算机科学的迅速发展，模型框架的优化需要生物学、计算机技术、

数据分析等多个领域的紧密协作。这种跨学科的知识融合不仅对技术和方法的进步至关重要，更对研究人员的综合能力提出了更高要求。未来的成功将依赖于这一多元化知识体系的构建和创新。总的来说，目前机器学习在GEM中的应用仍局限于现有的框架，未来的研究应突破这些限制，推动机理与数据驱动方法的协同发展，从而为数字孪生细胞和精准生物学研究带来新的突破。

参 考 文 献

[1] GONG Z J, CHEN J Y, JIAO X Y, et al. Genome-scale metabolic network models for industrial microorganisms metabolic engineering: Current advances and future prospects [J]. Biotechnology Advances, 2024, 72: 108319.

[2] ORTH J D, THIELE I, PALSSON B Ø. What is flux balance analysis?[J]. Nature Biotechnology, 2010, 28(3): 245-248.

[3] WAGNER A, WANG C, FESSLER J, et al. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity[J]. Cell, 2021, 184(16): 4168-4185. e21.

[4] KIM B J, KIM W J, KIM D I, et al. Applications of genome-scale metabolic network model in metabolic engineering[J]. Journal of Industrial Microbiology & Biotechnology, 2015, 42(3): 339-348.

[5] RYU J Y, KIM H U, LEE S Y. Reconstruction of genome-scale human metabolic models using omics data[J]. Integrative Biology, 2015, 7(8): 859-868.

[6] GU C D, KIM G B, KIM W J, et al. Current status and applications of genome-scale metabolic models[J]. Genome Biology, 2019, 20(1): 121.

[7] EDWARDS J S, PALSSON B O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype[J]. Journal of Biological Chemistry, 1999, 274(25): 17410-17416.

[8] ABRAMSON J, ADLER J, DUNGER J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold3[J]. Nature, 2024, 630(8016): 493-500.

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

[10] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[11] YU T H, CUI H Y, LI J C, et al. Enzyme function prediction using contrastive learning[J]. Science, 2023, 379(6639): 1358-1363.

[12] FINNIGAN W, HEPWORTH L J, FLITSCH S L, et al. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades[J]. Nature Catalysis, 2021, 4(2): 98-104.

[13] 曾涛, 巫瑞波. 数据驱动的酶反应预测与设计[J]. 合成生物学, 2023, 4(3): 535-550.
ZENG T, WU R B. Data-driven prediction and design for enzymatic reactions[J]. Synthetic Biology Journal, 2023, 4(3): 535-550.

[14] SABZEVARI M, SZEDMAK S, PENTTILÄ M, et al. Strain design optimization using reinforcement learning[J]. PLoS Computational Biology, 2022, 18(6): e1010177.

[15] 禹伟, 高教琪, 周雍进. 一碳生物转化合成有机酸的研究进展[J]. 合成生物学, 2024, 5(5): 1169-1188.
YU W, GAO J Q, ZHOU Y J. Bioconversion of one carbon feedstocks for producing organic acids[J]. Synthetic Biology Journal, 2024, 5(5): 1169-1188.

[16] KUNDU P, BEURA S, MONDAL S, et al. Machine learning for the advancement of genome-scale metabolic modeling[J]. Biotechnology Advances, 2024, 74: 108400.

[17] KANEHISA M, FURUMICHI M, SATO Y, et al. KEGG: biological systems database as a model of the real world[J]. Nucleic Acids Research, 2025, 53(D1): D672-D677.

[18] The UniProt Consortium. UniProt: the universal protein knowledgebase[J]. Nucleic Acids Research, 2018, 46(5): 2699.

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

[20] DUVAUD S, GABELLA C, LISACEK F, et al. Expasy, the Swiss bioinformatics resource portal, as designed by its users [J]. Nucleic Acids Research, 2021, 49(W1): W216-W227.

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

[22] BANSAL P, MORGAT A, AXELSEN K B, et al. Rhea, the reaction knowledgebase in 2022[J]. Nucleic Acids Research, 2022, 50(D1): D693-D700.

[23] KARP P D, BILLINGTON R, CASPI R, et al. The BioCyc collection of microbial genomes and metabolic pathways[J]. Briefings in Bioinformatics, 2019, 20(4): 1085-1093.

[24] MORETTI S, TRAN V D T, MEHL F, et al. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models[J]. Nucleic Acids Research, 2021, 49(D1): D570-D574.

[25] KING Z A, LU J, DRÄGER A, et al. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models [J]. Nucleic Acids Research, 2016, 44(D1): D515-D522.

[26] MALIK-SHERIFF R S, GLONT M, NGUYEN T V N, et al. BioModels-15 years of sharing computational models in life science[J]. Nucleic Acids Research, 2020, 48(D1): D407-D415.

[27] ARKIN A P, COTTINGHAM R W, HENRY C S, et al. KBase: the United States Department of Energy systems biology knowledgebase[J]. Nature Biotechnology, 2018, 36(7): 566-569.

[28] LIEVEN C, BEBER M E, OLIVIER B G, et al. Publisher Correction: MEMOTE for standardized genome-scale metabolic model testing[J]. Nature Biotechnology, 2020, 38(4): 504.

[29] THIELE I, PALSSON B Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction[J]. Nature Protocols, 2010, 5(1): 93-121.

[30] HEIRENDT L, ARRECKX S, PFAU T, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0[J]. Nature Protocols, 2019, 14(3): 639-702.

[31] WANG H, MARCIŠAUSKAS S, SÁNCHEZ B J, et al. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*[J]. PLoS Computational Biology, 2018, 14(10): e1006541.

[32] DEVOID S, OVERBEEK R, DEJONGH M, et al. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED[J]. Methods in Molecular Biology, 2013, 985: 17-45.

[33] MACHADO D, ANDREJEV S, TRAMONTANO M, et al. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities[J]. Nucleic Acids Research, 2018, 46(15): 7542-7553.

[34] CAPELA J, LAGOA D, RODRIGUES R, et al. Merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models[J]. Nucleic Acids Research, 2022, 50(11): 6052-6066.

[35] ZIMMERMANN J, KALETA C, WASCHINA S. Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models[J]. Genome Biology, 2021, 22(1): 81.

[36] HEINKEN A, HERTEL J, ACHARYA G, et al. Genome-scale metabolic reconstruction of 7302 human microorganisms for personalized medicine[J]. Nature Biotechnology, 2023, 41(9): 1320-1331.

[37] FANG X, LLOYD C J, PALSSON B O. Reconstructing organisms *in silico*: genome-scale models and their emerging applications[J]. Nature Reviews Microbiology, 2020, 18(12): 731-743.

[38] LU H Z, XIAO L C, LIAO W B, et al. Cell factory design with advanced metabolic modelling empowered by artificial intelligence[J]. Metabolic Engineering, 2024, 85: 61-72.

[39] CARRASCO MURIEL J, LONG C, SONNENSCHEIN N. Simultaneous application of enzyme and thermodynamic constraints to metabolic models using an updated Python implementation of GECKO[J]. Microbiology Spectrum, 2023, 11(6): e0170523.

[40] BI X Y, CHENG Y, XU X H, et al. etiBsu1209: a comprehensive multiscale metabolic model for *Bacillus subtilis*[J]. Biotechnology and Bioengineering, 2023, 120(6): 1623-1639.

[41] SCHROEDER W L, SUTHERS P F, WILLIS T C, et al. Current state, challenges, and opportunities in genome-scale resource allocation models: a mathematical perspective[J]. Metabolites, 2024, 14(7): 365.

[42] HENRY C S, JANKOWSKI M D, BROADBELT L J, et al. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism[J]. Biophysical Journal, 2006, 90(4): 1453-1461.

[43] HENRY C S, BROADBELT L J, HATZIMANIKATIS V. Thermodynamics-based metabolic flux analysis[J]. Biophysical Journal, 2007, 92(5): 1792-1805.

[44] DASH S, OLSON D G, JOSHUA CHAN S H, et al. Thermodynamic analysis of the pathway for ethanol production from cellobiose in *Clostridium thermocellum*[J]. Metabolic Engineering, 2019, 55: 161-169.

[45] NOOR E, BAR-EVEN A, FLAMHOLZ A, et al. Pathway thermodynamics highlights kinetic obstacles in central metabolism[J]. PLoS Computational Biology, 2014, 10(2): e1003483.

[46] HÄDICKE O, VON KAMP A, AYDOGAN T, et al. OptMDFpathway: identification of metabolic pathways with maximal thermodynamic driving force and its application for analyzing the endogenous CO_2 fixation potential of *Escherichia coli*[J]. PLoS Computational Biology, 2018, 14(9): e1006492.

[47] FLAMHOLZ A, NOOR E, BAR-EVEN A, et al. Glycolytic strategy as a tradeoff between energy yield and protein cost[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(24): 10039-10044.

[48] BEKIARIS P S, KLAMT S. Automatic construction of metabolic models with enzyme constraints[J]. BMC Bioinformatics, 2020, 21(1): 19.

[49] BEG Q K, VAZQUEZ A, ERNST J, et al. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(31): 12663-12668.

[50] ADADI R, VOLKMER B, MILO R, et al. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters[J]. PLoS Computational Biology, 2012, 8(7): e1002575.

[51] MAO Z T, ZHAO X, YANG X, et al. ECMpy, a simplified workflow for constructing enzymatic constrained metabolic network model[J]. Biomolecules, 2022, 12(1): 65.

[52] MAO Z T, NIU J H, ZHAO J X, et al. ECMpy 2.0: a Python package for automated construction and analysis of enzyme-constrained models[J]. Synthetic and Systems Biotechnology, 2024, 9(3): 494-502.

[53] SÁNCHEZ B J, ZHANG C, NILSSON A, et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints[J]. Molecular Systems Biology, 2017, 13(8): 935.

[54] DOMENZAIN I, SÁNCHEZ B, ANTON M, et al. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0[J]. Nature Communications, 2022, 13(1): 3766.

[55] CHEN Y, GUSTAFSSON J, TAFUR RANGEL A, et al. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0[J]. Nature Protocols, 2024, 19(3): 629-667.

[56] KERKHOVEN E J. Advances in constraint-based models: methods for improved predictive power based on resource allocation constraints[J]. Current Opinion in Microbiology, 2022, 68: 102168.

[57] YANG X, MAO Z T, ZHAO X, et al. Integrating thermodynamic and enzymatic constraints into genome-scale metabolic models[J]. Metabolic Engineering, 2021, 67: 133-144.

[58] LI G, HU Y T, ZRIMEC J, et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism [J]. Nature Communications, 2021, 12(1): 190.

[59] O'BRIEN E J, LERMAN J A, CHANG R L, et al. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction[J]. Molecular Systems Biology, 2013, 9: 693.

[60] ALSIYABI A, CHOWDHURY N B, LONG D N, et al. Enhancing *in silico* strain design predictions through next generation metabolic modeling approaches[J]. Biotechnology Advances, 2022, 54: 107806.

[61] CHEN K, GAO Y, MIH N, et al. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(43): 11548-11553.

[62] DU B, YANG L, LLOYD C J, et al. Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in *Escherichia coli*[J]. PLoS Computational Biology, 2019, 15(12): e1007525.

[63] YANG L, MIH N, ANAND A, et al. Cellular responses to reactive oxygen species are predicted from molecular mechanisms[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 14368-14373.

[64] ZHAO J, CHEN K, PALSSON B O, et al. StressME: unified computing framework of *Escherichia coli* metabolism, gene expression, and stress responses[J]. PLoS Computational Biology, 2024, 20(2): e1011865.

[65] CHEN Y, VAN PELT-KLEINJAN E, VAN OLST B, et al. Proteome constraints reveal targets for improving microbial fitness in nutrient-rich environments[J]. Molecular Systems Biology, 2021, 17(4): e10093.

[66] ELSEMMAN I E, RODRIGUEZ PRADO A, GRIGAITIS P, et al. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies[J]. Nature Communications, 2022, 13(1): 801.

[67] LI F R, CHEN Y, QI Q, et al. Improving recombinant protein production by yeast through genome-scale modeling using proteome constraints[J]. Nature Communications, 2022, 13(1): 2969.

[68] CHEN Y, LI F R, MAO J W, et al. Yeast optimizes metal utilization based on metabolic network and enzyme kinetics[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(12): e2020154118.

[69] SALVY P, HATZIMANIKATIS V. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models[J]. Nature Communications, 2020, 11(1): 30.

[70] OFTADEH O, SALVY P, MASID M, et al. A genome-scale metabolic model of *Saccharomyces cerevisiae* that integrates expression constraints and reaction thermodynamics[J]. Nature Communications, 2021, 12(1): 4790.

[71] RYU J Y, KIM H U, LEE S Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(28): 13996-14001.

[72] KIM G B, KIM J Y, LEE J A, et al. Functional annotation of enzyme-encoding genes using deep learning with transformer layers[J]. Nature Communications, 2023, 14(1): 7370.

[73] SHI Z K, DENG R, YUAN Q Q, et al. Enzyme commission number prediction and benchmarking with hierarchical dual-core multitask learning framework[J]. Research, 2023, 6: 0153.

[74] HAN S R, PARK M, KOSARAJU S, et al. Evidential deep learning for trustworthy prediction of enzyme commission number[J]. Briefings in Bioinformatics, 2023, 25(1): bbad401.

[75] BUTON N, COSTE F, LE CUNFF Y. Predicting enzymatic function of protein sequences with attention[J]. Bioinformatics, 2023, 39(10): btad620.

[76] WATANABE N, YAMAMOTO M, MURATA M, et al. EnzymeNet: residual neural networks model for Enzyme Commission number prediction[J]. Bioinformatics Advances, 2023, 3(1): vbad173.

[77] SONG Y D, YUAN Q M, CHEN S, et al. Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures[J]. Nature Communications, 2024, 15(1): 8180.

[78] KOCH M, DUIGOU T, FAULON J L. Reinforcement learning for bioretrosynthesis[J]. ACS Synthetic Biology, 2020, 9(1): 157-168.

[79] COLEY C W, THOMAS D A, LUMMISS J A M, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning[J]. Science, 2019, 365(6453): eaax1566.

[80] LEVIN I, LIU M J, VOIGT C A, et al. Merging enzymatic and synthetic chemistry with computational synthesis planning[J]. Nature Communications, 2022, 13(1): 7747.

[81] KREUTTER D, SCHWALLER P, REYMOND J L. Predicting enzymatic reactions with a molecular transformer[J]. Chemical Science, 2021, 12(25): 8648-8659.

[82] PROBST D, MANICA M, NANA TEUKAM Y G, et al. Biocatalysed synthesis planning using data-driven learning[J]. Nature Communications, 2022, 13(1): 964.

[83] ZHENG S J, ZENG T, LI C T, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP[J]. Nature Communications, 2022, 13(1): 3342.

[84] ZENG T, JIN Z H, ZHENG S J, et al. Developing BioNavi for hybrid retrosynthesis planning[J]. JACS Au, 2024, 4(7): 2492-2502.

[85] KROLL A, RANJAN S, ENGQVIST M K M, et al. A general model to predict small molecule substrates of enzymes based on machine and deep learning[J]. Nature Communications, 2023, 14(1): 2787.

[86] UPADHYAY V, BOORLA V S, MARANAS C D. Rank-ordering of known enzymes as starting points for re-engineering novel substrate activity using a convolutional neural network[J]. Metabolic Engineering, 2023, 78: 171-182.

[87] ZHANG D C, XING H D, LIU D L, et al. Discovery of toxin-degrading enzymes with positive unlabeled deep learning[J]. ACS Catalysis, 2024, 14(5): 3336-3348.

[88] SHI Z K, WANG D H, LI Y, et al. REME: an integrated platform for reaction enzyme mining and evaluation[J]. Nucleic Acids Research, 2024, 52(W1): W299-W305.

[89] XING H D, CAI P L, LIU D L, et al. High-throughput prediction of enzyme promiscuity based on substrate-product pairs[J]. Briefings in Bioinformatics, 2024, 25(2): bbae089.

[90] OYETUNDE T, ZHANG M H, CHEN Y X, et al. BoostGAPFILL: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods[J]. Bioinformatics, 2017, 33(4): 608-611.

[91] CHEN C, LIAO C, LIU Y Y. Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning [J]. Nature Communications, 2023, 14(1): 2375.

[92] HUANG W H, YANG F, ZHANG Q, et al. A dual-scale fused hypergraph convolution-based hyperedge prediction model for predicting missing reactions in genome-scale metabolic networks[J]. Briefings in Bioinformatics, 2024, 25(5): bbae383.

[93] BOER M D, MELKONIAN C, ZAFEIROPOULOS H, et al. Improving genome-scale metabolic models of incomplete genomes with deep learning[J]. iScience, 2024, 27(12): 111349.

[94] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium[J]. Nature Genetics, 2000, 25(1): 25-29.

[95] SHEN H B, CHOU K C. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses[J]. Biochemical and Biophysical Research Communications, 2007, 364(1): 53-59.

[96] LI Y H, XU J Y, TAO L, et al. SVM-prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity[J]. PLoS One, 2016, 11 (8): e0155290.

[97] LI Y, WANG S, UMAROV R, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning[J]. Bioinformatics, 2018, 34(5): 760-769.

[98] NURSIMULU N, XU L L, WASMUTH J D, et al. Improved enzyme annotation with EC-specific cutoffs using DETECT v2 [J]. Bioinformatics, 2018, 34(19): 3393-3395.

[99] DALKIRAN A, RIFAIOGLU A S, MARTIN M J, et al. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature[J]. BMC Bioinformatics, 2018, 19(1): 334.

[100] RIVES A, MEIER J, SERCU T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[101] LIN Z M, AKIN H, RAO R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model[J]. Science, 2023, 379(6637): 1123-1130.

[102] HADADI N, HAFNER J, SHAJKOFCI A, et al. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies[J]. ACS Synthetic Biology, 2016, 5(10): 1155-1166.

[103] KHERSONSKY O, TAWFIK D S. Enzyme promiscuity: a mechanistic and evolutionary perspective[J]. Annual Review of Biochemistry, 2010, 79: 471-505.

[104] PONTRELLI S, FRICKE R C B, TEOH S T, et al. Metabolic repair through emergence of new pathways in *Escherichia coli* [J]. Nature Chemical Biology, 2018, 14(11): 1005-1009.

[105] NI Z F, STINE A E, TYO K E J, et al. Curating a comprehensive set of enzymatic reaction rules for efficient novel biosynthetic pathway design[J]. Metabolic Engineering, 2021, 65: 79-87.

[106] NAM H, LEWIS N E, LERMAN J A, et al. Network context and selection in the evolution to enzyme specificity[J]. Science, 2012, 337(6098): 1101-1104.

[107] MORIYA Y, SHIGEMIZU D, HATTORI M, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server[J].

Nucleic Acids Research, 2010, 38(S2): W138-W143.

[108] DELÉPINE B, DUIGOU T, CARBONELL P, et al. RetroPath2.0: a retrosynthesis workflow for metabolic engineers[J]. Metabolic Engineering, 2018, 45: 158-170.

[109] DING S Z, TIAN Y, CAI P L, et al. novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model[J]. Nucleic Acids Research, 2020, 48(W1): W477-W487.

[110] GENHEDEN S, THAKKAR A, CHADIMOVÁ V, et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning[J]. Journal of Cheminformatics, 2020, 12(1): 70.

[111] GRICOURT G, MEYER P, DUIGOU T, et al. Artificial intelligence methods and models for retro-biosynthesis: a scoping review[J]. ACS Synthetic Biology, 2024, 13(8): 2276-2294.

[112] COLEY C W, ROGERS L, GREEN W H, et al. SCScore: synthetic complexity learned from a reaction corpus[J]. Journal of Chemical Information and Modeling, 2018, 58(2): 252-261.

[113] CARBONELL P, WONG J, SWAINSTON N, et al. Selenzyme: enzyme selection tool for pathway design[J]. Bioinformatics, 2018, 34(12): 2153-2154.

[114] RAHMAN S A, CUESTA S M, FURNHAM N, et al. EC-BLAST: a tool to automatically search and compare enzyme reactions[J]. Nature Methods, 2014, 11(2): 171-174.

[115] GIRI V, SIVAKUMAR T V, CHO K M, et al. RxnSim: a tool to compare biochemical reactions[J]. Bioinformatics, 2015, 31(22): 3712-3714.

[116] STONEY R A, HANKO E K R, CARBONELL P, et al. SelenzymeRF: updated enzyme suggestion software for unbalanced biochemical reactions[J]. Computational and Structural Biotechnology Journal, 2023, 21: 5868-5876.

[117] WILLETT P. Searching techniques for databases of two- and three-dimensional chemical structures[J]. Journal of Medicinal Chemistry, 2005, 48(13): 4183-4199.

[118] SCHWALLER P, HOOVER B, REYMOND J L, et al. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions[J]. Science Advances, 2021, 7 (15): eabe4166.

[119] QIAN W J, WANG X R, HUANG Y S, et al. Deep learning-driven insights into enzyme-substrate interaction discovery[J]. Journal of Chemical Information and Modeling, 2025, 65(1): 187-200.

[120] ORTH J D, PALSSON B Ø. Systematizing the generation of missing metabolic knowledge[J]. Biotechnology and Bioengineering, 2010, 107(3): 403-412.

[121] THIELE I, VLASSIS N, FLEMING R M T. fastGapFill: efficient gap filling in metabolic networks[J]. Bioinformatics, 2014, 30(17): 2529-2531.

[122] KUMAR V S, DASIKA M S, MARANAS C D. Optimization based automated curation of metabolic reconstructions[J]. BMC Bioinformatics, 2007, 8: 212.

[123] HECKMANN D, LLOYD C J, MIH N, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models[J]. Nature Communications, 2018, 9(1): 5252.

[124] LI F R, YUAN L, LU H Z, et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction[J]. Nature Catalysis, 2022, 5(8): 662-672.

[125] KROLL A, ROUSSET Y, HU X P, et al. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning[J]. Nature Communications, 2023, 14(1): 4139.

[126] QIU S Z, ZHAO S M, YANG A D. DLTKcat: deep learning-based prediction of temperature-dependent enzyme turnover rates[J]. Briefings in Bioinformatics, 2023, 25(1): bbad506.

[127] WANG T, XIANG G M, HE S W, et al. DeepEnzyme: a robust deep learning model for improved enzyme turnover number prediction by utilizing features of protein 3D-structures[J]. Briefings in Bioinformatics, 2024, 25(5): bbae409.

[128] KROLL A, ENGQVIST M K M, HECKMANN D, et al. Deep learning allows genome-scale prediction of Michaelis constants from structural features[J]. PLoS Biology, 2021, 19 (10): e3001402.

[129] HE X, YAN M. GraphKM: machine and deep learning for K_m prediction of wildtype and mutant enzymes[J]. BMC Bioinformatics, 2024, 25(1): 135.

[130] MAEDA K, HATAE A, SAKAI Y, et al. MLAGO: machine learning-aided global optimization for Michaelis constant estimation of kinetic modeling[J]. BMC Bioinformatics, 2022, 23(1): 455.

[131] WANG J J, YANG Z J, CHEN C, et al. MPEK: a multitask deep learning framework based on pretrained language models for enzymatic reaction kinetic parameters prediction[J]. Briefings in Bioinformatics, 2024, 25(5): bbae387.

[132] YU H, DENG H X, HE J H, et al. UniKP: a unified framework for the prediction of enzyme kinetic parameters[J]. Nature Communications, 2023, 14(1): 8211.

[133] SHEN X W, CUI Z H, LONG J Y, et al. EITLEM-Kinetics: a deep-learning framework for kinetic parameter prediction of mutant enzymes[J]. Chem Catalysis, 2024, 4(9): 101094.

[134] WANG L, UPADHYAY V, MARANAS C D. dGPredictor: automated fragmentation method for metabolic reaction free energy prediction and *de novo* pathway design[J]. PLoS Computational Biology, 2021, 17(9): e1009448.

[135] ALAZMI M, KUWAHARA H, SOUFAN O, et al. Systematic selection of chemical fingerprint features improves the Gibbs energy prediction of biochemical reactions[J]. Bioinformatics, 2019, 35(15): 2634-2643.

[136] JUNG F, FREY K, ZIMMER D, et al. DeepSTABp: a deep learning approach for the prediction of thermal protein stability[J]. International Journal of Molecular Sciences, 2023, 24(8): 7444.

[137] LI M Y, WANG H Z, YANG Z W, et al. DeepTM: a deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences[J]. Computational and Structural Biotechnology Journal, 2023, 21: 5544-5560.

[138] LI G, RABE K S, NIELSEN J, et al. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima[J]. ACS Synthetic Biology, 2019, 8(6): 1411-1420.

[139] GADO J E, BECKHAM G T, PAYNE C M. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning[J]. Journal of Chemical Information and Modeling, 2020, 60(8): 4098-4107.

[140] LI G, BURIC F, ZRIMEC J, et al. Learning deep representations of enzyme thermal adaptation[J]. Protein Science, 2022, 31(12): e4480.

[141] ZHANG Y, GUAN F F, XU G S, et al. A novel thermophilic chitinase directly mined from the marine metagenome using the deep learning tool Preoptem[J]. Bioresources and Bioprocessing, 2022, 9(1): 54.

[142] NILSSON A, NIELSEN J, PALSSON B O. Metabolic models of protein allocation call for the kinetome[J]. Cell Systems, 2017, 5(6): 538-541.

[143] SUTHERS P F, FOSTER C J, SARKAR D, et al. Recent advances in constraint and machine learning-based metabolic modeling by leveraging stoichiometric balances, thermodynamic feasibility and kinetic law formalisms[J]. Metabolic Engineering, 2021, 63: 13-33.

[144] JANKOWSKI M D, HENRY C S, BROADBELT L J, et al. Group contribution method for thermodynamic analysis of complex metabolic networks[J]. Biophysical Journal, 2008, 95(3): 1487-1499.

[145] NOOR E, HARALDSDÓTTIR H S, MILO R, et al. Consistent estimation of Gibbs energy using component contributions[J]. PLoS Computational Biology, 2013, 9(7): e1003098.

[146] BEBER M E, GOLLUB M G, MOZAFFARI D, et al. eQuilibrator 3.0: a database solution for thermodynamic constant estimation[J]. Nucleic Acids Research, 2022, 50(D1): D603-D609.

[147] LEUENBERGER P, GANSCHA S, KAHRAMAN A, et al. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability[J]. Science, 2017, 355(6327): eaai7825.

通讯作者：李斐然（1993—），女，助理教授，博士生导师。研究方向为基因组规模代谢模型开发、微生物细胞工厂设计、酶参数预测，致力于构建数字生命模型。

E-mail：feiranli@sz.tsinghua.edu.cn

第一作者：吴柯（2000—），男，博士研究生。研究方向为机器学习辅助基因组规模代谢模型开发。

E-mail：wk37@tju.edu.cn