• AI 驱动底层技术 •

毛志涛 博士，中国科学院天津工业生物技术研究所副研究员，中国科学院青年创新促进会会员，中国生物信息学会(筹)生物数据资源专委会委员。主要从事大模型辅助的菌种知识库开发、多约束细胞模型构建、细胞设计方法和工具开发等研究。现主持国家重点研发计划子课题、国家自然科学基金、省部共建重点实验室开放课题等多个项目。在国际主流期刊 *Nucleic Acids Research*、*Metabolic Engineering* 等发表 SCI 论文 30 余篇，申请专利和软著 8 项。

面向生物制造的数据库、知识库与大模型

毛志涛[1,2*]，廖小平[1,2]，马红武[1,2]

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

毛志涛, 廖小平, 马红武. 面向生物制造的数据库、知识库与大模型[J]. 生物工程学报, 2025, 41(3): 901-916.
MAO Zhitao, LIAO Xiaoping, MA Hongwu. Databases, knowledge bases, and large models for biomanufacturing[J]. Chinese Journal of Biotechnology, 2025, 41(3): 901-916.

摘 要：生物制造技术是一种融合生物学、化学和工程学的前沿制造方法，利用可再生生物质和生物体作为生产介质，通过发酵过程规模化生产目标产品。与传统石化路线相比，生物制造在减少 CO_2 排放、降低能耗和成本方面具有显著优势。随着系统生物学、合成生物学的发展和生物大数据的积累，人工智能、大模型和高性能计算等信息技术与生物技术的融合，生物制造正逐步进入数据驱动时代。本文综述了面向生物制造的数据库、知识库与大语言模型的最新研究进展，探讨了该领域的发展方向、难点以及新兴技术方法，为相关领域的科研工作提供了参考和启示。

关键词：生物制造；数据驱动；数据库；知识库；大模型

资助项目：国家重点研发计划(2022YFC2106000)；国家自然科学基金(32300529)；天津市合成生物技术创新能力提升行动项目(TSBICIP-PTJJ-012)
This work was supported by the National Key Research and Development Program of China (2022YFC2106000), the National Natural Science Foundation of China (32300529), and the Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (TSBICIP-PTJJ-012).
*Corresponding author. E-mail: mao_zt@tib.cas.cn
Received: 2024-08-13; Accepted: 2025-01-24; Published online: 2025-01-24

Databases, knowledge bases, and large models for biomanufacturing

MAO Zhitao[1,2*], LIAO Xiaoping[1,2], MA Hongwu[1,2]

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

Abstract: Biomanufacturing is an advanced manufacturing method that integrates biology, chemistry, and engineering. It utilizes renewable biomass and biological organisms as production media to scale up the production of target products through fermentation. Compared with petrochemical routes, biomanufacturing offers significant advantages in reducing CO_2 emissions, lowering energy consumption, and cutting costs. With the development of systems biology and synthetic biology and the accumulation of bioinformatics data, the integration of information technologies such as artificial intelligence, large models, and high-performance computing with biotechnology is propelling biomanufacturing into a data-driven era. This paper reviews the latest research progress on databases, knowledge bases, and large language models for biomanufacturing. It explores the development directions, challenges, and emerging technical methods in this field, aiming to provide guidance and inspiration for scientific research in related areas.

Keywords: biomanufacturing; data-driven; databases; knowledge bases; large models

生物制造是一种前沿的制造技术，融合了生物学、化学和工程学等多学科技术。它以可再生生物质为原料，利用生物体作为生产介质，通过菌种、细胞和酶的作用，规模化发酵生产目标产品[1]。这一技术因其清洁、高效和可再生的特点，正迅速崭露头角。以 1,3-丙二醇的生物制造为例，与传统的石化路线相比，CO_2 排放减少 63%，能耗降低 30%，原料成本下降 37%[2]。因此，生物制造在替代工业基础原材料的化石原料、取代高能耗高物耗高排放的工艺路线以及实现二氧化碳的大规模工业化利用方面具有重要潜力[3]。

生物制造正在能源、化工、医药和食品等领域重塑工业格局，被视为引领“第四次工业革命”的重要力量。随着合成生物学、基因编辑等底层技术的突破性进展，全球生物制造产业正呈现指数级增长态势。经济合作与发展组织(Organisation for Economic Co-operation and Development, OECD)预测，到 2030 年，约 35% 的化学品和其他工业产品将来自工业生物技术，辐射产值可达 30 万亿美元[3]。在生物经济中，生物制造产业的贡献率将占 39%，且每年可减少 10–25 亿 t 二氧化碳排放，带来显著的经济和环境效益[4]。当前生物制造技术创新呈现多维度突破，包括高效底盘细胞构建、代谢途径优化、生物-化学偶联反应等核心技术的持续迭代。罗氏、诺维信、陶氏-杜邦等国际巨头，以及大量新兴企业，正持续投入数百亿美元开展生物制造相关研究，快速推进药物分子、聚合材料、未来食品和关键化学品的研发，逐步形成高技术壁垒[5]。

近年来，随着系统生物学和合成生物学的飞速发展，以及生物大数据的日益丰富，信息技术领域的人工智能、深度学习和高性能计算等正迅速与生物技术深度融合。这种融合为生物系统的设计带来了全新的理论框架和方法学支持，推动生物制造进入了以数据驱动为特征的新阶段。本文将全面回顾生物制造中数据库、知识

库和大模型的最新研究成果，分析当前的发展趋势和技术挑战，介绍新兴的方法和工具，以期为国内科研工作者提供借鉴和思路。

1 面向生物制造的数据库

生物制造作为一种新兴的制造模式，其核心在于利用生物学原理，通过生物体或其组成部分进行高效、环境友好的工业生产。这一过程涉及复杂的生物网络、多样化的生物催化剂，以及精细的生物过程控制等。生物制造领域的快速发展也带来了信息管理上的挑战。随着研究的深入和技术的迭代，大量的数据和知识需要被系统地整理和存储。这就需要构建一个全面且系统的数据库，它不仅要能够容纳现有的研究成果，还要能够适应未来技术发展的需求，为研究人员提供一个可靠、易用的信息资源平台。

1.1 数据库在生物制造中的应用

在生物制造中，数据库主要用于收集、存储、管理和分析大量的生物数据。这些数据包括基因序列、蛋白质结构、代谢通路、细胞反应以及环境条件等信息(表 1)。通过建立和利用这些数据库，研究人员可以更高效地进行数据分析和实验设计，优化生物制造过程，提高产品产量。

表 1 生物制造相关的数据库

Table 1 Databases related to biomanufacturing

Type	Name	Description	Reference
Gene	GenBank	A widely used public gene sequence database containing gene sequences and annotations from various organisms, one of the largest gene repositories in the world	[6]
	Ensembl	Provides genome sequences, detailed gene annotations, and functional information for various species, supporting gene function analysis and comparative genomics studies	[7]
Protein	ProteomicsDB	Integrates protein data from various experimental techniques and research projects, including mass spectrometry data, protein expression levels, and protein modification information	[8]
	PRIDE	A proteomics data repository that collects and stores protein and peptide identification data from mass spectrometry experiments, supporting data submission, sharing, and re-analysis	[9]
	PaxDb	A database that integrates protein abundance information across multiple species, providing quantitative proteomics data covering different tissues and cell types, supporting comparative analysis of protein abundance	[10]
	PeptideAtlas	A proteomics database that integrates global mass spectrometry experiment data, providing detailed peptide identification and annotation information, supporting protein sequence coverage and expression level analysis	[11]
	PDB	A database that stores 3D structural data of proteins and other biological macromolecules, providing high-resolution molecular structure information	[12]
Metabolite	ChEBI	A database of biologically relevant chemical entities, offering detailed information on chemical structures, names, and annotations, useful for the classification and annotation of metabolites	[13]
	PubChem	An open chemical database containing information on millions of chemical substances, providing detailed chemical structure and functional annotations for metabolites	[14]
	MetaboLights	A metabolomics data repository that provides structural, annotation, and experimental data for metabolites, supporting metabolite identification and pathway analysis	[15]

待续

续表 1

Type	Name	Description	Reference
Reaction	KEGG	A comprehensive bioinformatics resource containing information on genomes, metabolic pathways, and biological systems	[16]
	MetaCyc	A comprehensive database of metabolic pathways and enzyme-catalyzed reactions, covering metabolic information from multiple organisms, including data on known metabolic pathways, enzymes, compounds, and reactions	[17]
	BRENDA	A database that details enzyme and enzyme-catalyzed reaction information, providing data on enzyme classification, function, kinetic parameters, substrates, and products	[18]
	SMPDB	A database specializing in detailed information on small-molecule metabolic pathways, disease metabolism, drug action, and drug metabolism	[19]
	WikiPathways	An open-access database focused on integrating and displaying information on metabolic pathways, enzymes, compounds, and related reactions for various organisms	[20]
Biological netwok	MetaNetX	A platform for integrating and managing metabolic network and biochemical reaction information from multiple biological databases, supporting cross-species model reconstruction and pathway analysis	[21]
	BiGG Models	A comprehensive database providing metabolic network models for various organisms, supporting pathway analysis and systems biology research	[22]
	BioModels	A database of metabolic network models, encompassing mathematical models of multiple biological systems, aiming to promote systems biology research and model sharing and reuse	[23]
	STRING	A comprehensive database providing known and predicted protein-protein interaction information, covering multiple species	[24]
	STITCH	A database integrating information on compound-protein interactions, combining experimental data and computational predictions, aiming to facilitate drug discovery and biochemical research	[25]
	RegulonDB	A database focused on the gene regulatory network of *Escherichia coli*, providing information on genes, regulatory factors, and their interactions to support gene expression and metabolic regulation studies	[26]
	YEASTRACT+	An extended yeast gene regulatory database, integrating detailed information on transcription factors, binding sites, and their target genes, supporting transcriptional regulation and gene expression research	[27]
	CoryneRegNet	A database focused on the gene regulatory network of *Corynebacterium glutamicum*, providing detailed information on regulatory interactions	[28]
Fermentation process	SABIO-RK	A biochemical reaction kinetics database that collects kinetic parameters and related information on enzyme-catalyzed reactions from published scientific literature. It covers data across different organisms, tissues, cell types, and experimental conditions, making it highly valuable for modeling and simulating bioprocess fermentation	[29]
	BacDive	A metadatabase of bacterial and archaeal diversity, providing information on microbial strains such as taxonomy, physiological and biochemical characteristics, ecological environments, and culture conditions. It holds significant importance for strain selection and research in biological fermentation processes	[30]
	FermFooDb	A manually curated database of bioactive peptides derived from various foods, retaining comprehensive information about peptides and fermentation processes. It is suitable for bioactive peptide research and fermented food development	[31]

基因数据库是生物制造中最基础的数据库之一，包含了各种生物体的基因组序列信息。在生物制造中，基因数据库通过提供大量的基因组信息，使得研究人员可以识别与目标产物相关的关键基因，进行菌株的改造。例如，在氨基酸生产中，研究人员通过分析 NCBI 基因组数据库(GenBank)[6]中的相关基因序列，筛选出与产量提升相关的基因进行编辑，以提高目标产物的合成效率。这些数据库还可以用于菌株间的比较基因组学研究，帮助识别有利于代谢产物积累的遗传特性，进一步指导菌株优化[32]。

蛋白质数据库存储了大量的蛋白质结构和功能信息。在生物制造中，这些数据库可以帮助研究人员了解酶的结构和功能，从而设计和优化酶的催化性能。例如，通过 protein data bank (PDB)数据库中的三维结构信息，研究人员可以深入分析关键酶的活性位点，进行定点突变以提高反应效率[33]。在实际应用中，这类优化的酶已被用于生物燃料和药物前体的生产过程，提高了生物反应速率和产物选择性。

反应途径数据库记录了生物体内各种代谢反应及其相互关系。在构建高效代谢途径时，反应途径数据库能够提供详尽的代谢反应网络信息，帮助研究人员设计新的代谢途径以增加产物产量。例如，研究人员可以利用(Kyoto encyclopedia of genes and genomes, KEGG)数据库中的数据构建代谢网络模型[34]，预测不同工程改造对产物合成的影响，进而优化产品合成途径。此外，(Braunschweig Enzyme Database, BRENDA)数据库[18]提供了酶反应速率、底物特异性、抑制剂信息等，有助于研究人员针对具体代谢反应进行精准调整，优化生物制造过程中的酶催化效率。

各类组学数据库是生物制造领域的重要资源，它们为研究人员提供了从基因到蛋白质再到代谢物的全面数据，使得生物过程的精确调控和优化成为可能。例如，研究人员通过分析基因表达综合数据库(gene expression omnibus, GEO)[35]中的转录组数据，确定在不同发酵条件下高表达的基因，从而设计和优化生物反应器中的基因表达，提升目标产物的产量。类似地，蛋白丰度数据库(protein abundance database, PaxDb)[10]的数据有助于确定在不同实验条件下关键蛋白的表达水平，从而优化蛋白质的表达策略。代谢组数据库如 MetaboLights[15]可以提供代谢物丰度信息，帮助研究人员优化培养基成分和发酵参数，以提高代谢效率并减少副产物生成。

生物网络数据库为复杂的生物系统设计和分析提供了强有力的工具支持。通过这些数据库，研究人员可以有效地构建、分析和优化生物网络，从而改进基因线路设计、组学数据解析和代谢途径优化。例如，BiGG Models 数据库[22]提供了不同物种的高质量代谢网络模型，便于研究人员通过模拟和计算预测代谢流量，优化代谢途径的设计，提高目标产物的合成效率。STRING 数据库[24]通过提供蛋白质之间的相互作用信息，帮助研究人员构建和分析蛋白质互作网络。在基因线路设计中，研究人员可以利用 STRING 中的蛋白质互作数据，识别关键调控蛋白并优化其功能[36]，为生物制造过程中的优化和改造提供新思路。

生物过程数据库在生物发酵研究和应用中具有关键作用，整合了微生物发酵过程参数、代谢产物数据、生理生化数据等多种类型的数据，为研究人员提供系统化的数据支持。然而，目前缺乏专门面向生物过程的公开访问数据库。现有已发表的包含部分生物过程数据的数据库主要包括 SABIO-RK[29]、Explore Bacterial Diversity (BacDive)[30] 和 Fermented Food

Peptide Database (FermFooDb)[31]。SABIO-RK 汇集了丰富的生化反应动力学参数，可用于发酵过程的动力学建模和优化；BacDive 保留了有关生物活性肽和发酵过程的全面信息。这些数据库为研究人员深入理解发酵微生物的代谢机制、构建和优化代谢网络、识别关键酶和代谢途径提供了重要资源，有助于改进发酵工艺，提高目标产物的合成效率，推动生物制造领域的创新和发展。

总之，生物制造数据库是推动该领域研究和应用的关键资源，它们的发展和完善将为生物制造的创新和产业化提供强有力的数据支持。

1.2 生物制造相关数据库的发展现状

目前，生物制造领域已经建立了一些关键的数据库，这些数据库在数据存储、共享和分析方面发挥着重要作用。然而，现有的数据库在数据管理和应用方面仍存在一些问题和挑战。

(1) 数据分散。尽管已经有不少基础生物数据库存储了蛋白质和菌种等基础信息，但与生物制造相关的重要数据(如酶活性、稳定性和菌种的产物生成能力等)仍主要分散在文献中，甚至研究团队的内部文档中。例如，代谢工程相关的生物反应器条件和策略是极其宝贵的数据资源，但这些信息往往零散地存在于不同的文献中。手动从大量文献中提取数据不仅在一致性和准确性方面具有挑战性，还需要耗费大量时间，显著增加了研究成本和时间投入[37]。

(2) 数据混乱。由于生物数据的生成缺乏统一标准，不同研究团队常常使用各自的实验流程和数据格式，导致数据整合困难。例如，在基因组尺度代谢网络模型(genome-scale metabolic model, GEM)构建时，不同课题组会使用各自的代谢物、基因和反应的命名方式和组织方式，导致基于特定 GEM 开发的分析流程无法迁移到其他 GEM，限制了模型数据的共享和重复利用[22]。

(3) 数据质量差。生物数据在取样、处理和测定过程中存在较大误差，且不同实验室重复实验结果的再现性较差。例如，在酶动力学研究中，即使在标准化的实验操作和一致的实验条件下，不同实验者对同一酶的动力学参数测量结果仍可能表现出显著差异。此外，即使同一位实验者在不同时间点重复相同的测定，所得的动力学参数也可能存在较大波动。

(4) 重要数据缺失。在菌种和酶改造研究中，成功往往伴随着大量的失败尝试。然而，这些负样本数据在论文发表过程中常常被忽略或丢弃，但对于人工智能模型来说，这些负样本数据同样具有重要价值，需予以保留。此外，许多数据集中还缺乏必要的元数据标记，这对数据的进一步分析和利用造成了障碍。例如，在 RNA sequencing (RNA-seq)研究领域，样本的元数据通常存储于 NCBI Biosample 数据库。然而，这些元数据往往信息量不足，原因在于提交者通常只提供最低要求的强制性信息[38]。此外，部分关键的元数据并未被纳入这些数据库，而是散见于相关研究的学术出版物中[38]。

总之，尽管生物制造领域的数据库在数据存储、共享和分析方面发挥了重要作用，但现存的问题和挑战仍需解决。通过改进数据标准化、提高数据质量、保存负样本数据以及加强跨学科合作，可以进一步提升数据库的有效性和应用价值，从而推动生物制造领域的持续发展。

1.3 未来生物制造相关数据库发展方向

随着生物制造技术的不断进步，构建一个全面、高效的数据库成为其发展的关键。未来的数据库不仅要实现生物制造过程中全链条数据的标准化，确保数据的准确性和可比性，还

要通过高度集成的系统整合多层次、多源数据。这包括酶元件的突变位点、基因回路设计、多组学数据以及工程参数等。由于图数据库在处理复杂关系和大规模数据集方面的优势，它将成为该数据库构建方向的核心。通过图数据库技术，研究人员将能够更深入地探索生物制造的潜力，实现数据驱动的创新和优化，为生物制造领域带来革命性的变化。

1.3.1 生物制造需要的全链条数据

生物制造过程中，酶或细胞工厂将原料转化为目标产物，这一过程不仅依赖于传统的组学信息，还涉及更多复杂的生物学特征和工程参数。例如，酶元件的突变位点信息对于理解酶的功能和工程改造极为关键；酶的结构分析和动力学表征可以帮助优化其催化效率和稳定性；而基因回路的设计和表征则是合成生物学中用于构建生物模块的基础。此外，发酵过程的工程化探索需要综合考虑多组学数据(如转录组学、蛋白质组学和代谢组学等)，这些数据不仅能帮助理解细胞在发酵过程中的行为，还能用于调控生产途径以提高产物的产量和质量。工程参数和发酵参数(如温度、pH 值和溶氧水平等)的精确控制和优化，是保证生物发酵过程稳定运行和高效产出的关键。为了有效利用这些复杂数据，必须进行数据标准化，以确保不同实验和实验室之间的数据具有可比性。

此外，数据集成是另一项挑战，需要建立强大的数据管理系统来整合各种生物学、化学和工程数据。最终，这些数据应当被设计成能够轻松地整合进设计-构建-测试-学习(design-build-test-learn, DBTL)循环，通过迭代学习不断优化生物制造过程。因此，传统的生物数据与专有生物制造数据的结合，为生物制造的各个方面提供了支持，极大地促进了生物制造技术的发展和应用(图 1)。

1.3.2 图数据库在生物制造领域中的应用

生物学研究中的数据量正以前所未有的速度增长，传统的关系型数据库已难以满足日益复杂的数据结构和分析需求。图数据库以其灵

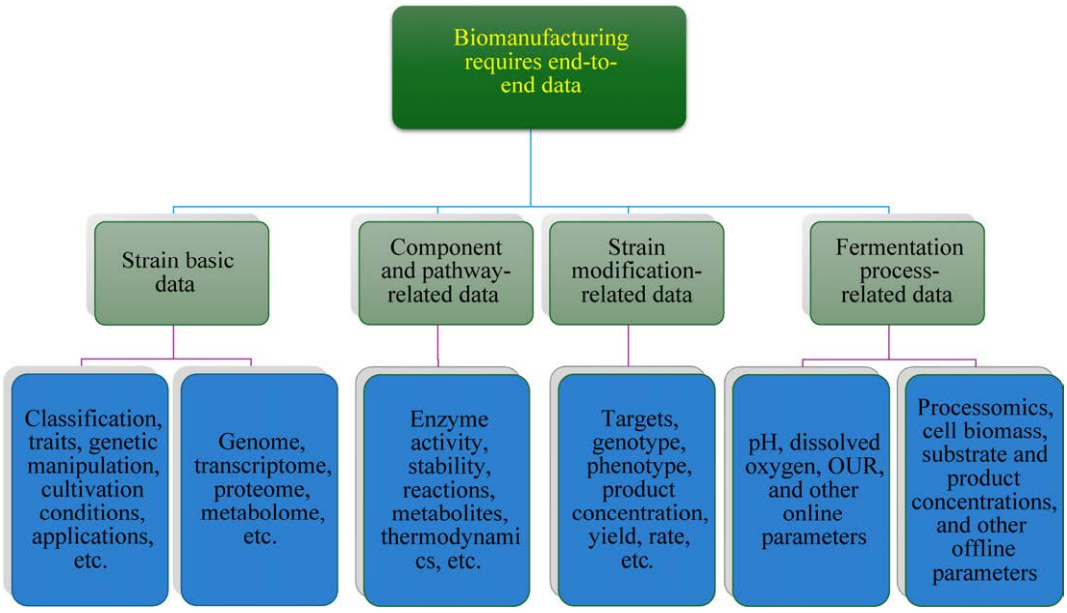

图 1 生物制造需要的全链条数据

Figure 1 End-to-end data required for biomanufacturing.

活的数据模型和高效的查询能力，成为解决这一问题的有效工具。例如，Reactome 由关系型数据库转变为 Neo4j 图数据库，大大提高了查询效率，平均查询时间减少了 93%，推动了生物网络数据的直观探索和分析[39]。此外，大肠杆菌调控知识图谱(*Escherichia coli* regulation miner, ERMer)也采用图数据库技术存储大肠杆菌中的各类调控数据，实现了多种复杂调控级联或模式的快速检索和可视化，推动了大肠杆菌中新调控模式和新菌种改造靶点的挖掘[40]。例如，对于 7 步以内的调控链，对存储于关系型数据库的调控数据使用虚拟表和使用 PostgreSQL 进行递归搜索，响应时间为 840 s，而图数据库只需 1.79 s[40]。此外，生物制造领域的数据来源多样，包括基因组数据、蛋白质组数据、代谢组数据等。图数据库能够整合这些异构数据，构建统一的知识图谱，支持跨数据源的综合分析。例如，BioGrakn 图数据库被用于整合疾病、基因和药物之间的关系数据，支持生物医学研究中的复杂查询和推理[41]。

总之，生物制造的全链条数据标准化和集成对于技术发展至关重要。这一过程涉及从酶元件的突变位点到基因回路设计的多层面数据，以及工程参数的精确控制。数据标准化确保了实验间的可比性，而数据集成则通过强大的管理系统整合了生物学、化学和工程数据，为 DBTL 循环的迭代优化提供了基础。图数据库的应用，以其高效的查询能力和灵活的数据模型，解决了传统数据库在处理复杂生物数据时的局限，推动了生物网络数据的直观探索和分析，预示着在生物制造领域中的广泛应用潜力。

2 面向生物制造的知识库

随着生物制造技术的不断进步，研究人员面临着处理和分析大量复杂生物数据的挑战，这对数据管理和知识获取提出了新的要求。在这一背景下，“知识库”与“数据库”这 2 个术语虽然在文献和讨论中常被交替使用，但它们实际上代表了 2 种截然不同的概念，各自具有独特的功能和应用场景。首先，数据库是一种结构化的数据存储系统，其核心功能是高效地管理和查询大量数据。通过对数据的组织和索引，数据库确保信息的可访问性和操作性，尤其适合处理结构化数据，使用户能够快速检索所需信息。与数据库不同，知识库则更为复杂，除了包含结构化数据外，还整合了非结构化数据。知识库通过元数据和语义信息的使用，增强了对数据的理解和应用能力。相比数据库，知识库的功能更加广泛，除了数据存储和检索外，还支持信息整合、推理和知识发现等高级功能。

在生物制造领域，知识库的这些高级功能尤为重要。生物制造的研究过程需要对全链条数据进行标准化和集成，涉及多种数据格式和丰富的资源，这使得知识库的构建显得尤为合适。研究人员可以利用知识库来处理和分析生物数据，从而深入理解生物过程和机制。这种深入的理解不仅有助于推动生物制造技术的创新和进步，还为解决复杂问题提供了新的思路和解决方案。

因此，从数据库向知识库的过渡，不仅是技术手段的升级，更是对生物制造领域数据管理和知识获取方式的深刻变革。这一转变将极大提升研究人员在复杂生物数据环境中的工作效率和创新能力，为生物制造的未来发展铺平道路。

2.1 知识库在生物制造中的应用现状

在生物制造领域，多个知识库已经建立并被广泛使用，这些知识库在生物制造概念出现之前就已经存在，并且随着生物制造的发展，它们在这一领域发挥了重要作用(表 2)，包括反

表 2　生物制造相关的知识库

Table 2　Knowledge base related to biomanufacturing

Name	Description	Reference
Rhea	An authoritative knowledge base of biochemical reactions based on the ChEBI chemical ontology (biochemical entities). It provides detailed, manually curated information on enzymatic and non-enzymatic reactions, including reaction equations, reactants, products, and their equilibrium states	[42]
Reactome	Provides molecular details of signaling, transport, DNA replication, metabolism, and other cellular processes, presented as an ordered network of molecular transformations. It offers high-quality, downloadable individual reaction diagrams and an overview of all content. The stored information resources have high interoperability with different databases	[43]
Gene Ontology	A comprehensive knowledge base of gene ontology, covering the functions of genes and their products (proteins and non-coding RNAs). All GO resources are dynamic (ontology, annotations, GO-CAM, external ontology links, etc.) and updated monthly	[44]
UniProt	A comprehensive protein knowledge base that provides high-quality protein functional annotations, detailed sequence information, and cross-database links	[45]
iModulonDB	A knowledge base of prokaryotic transcriptional regulation computed using ICA from high-quality transcriptome datasets. Users can select an organism from the homepage and then search or browse the curated iModulons that comprise its transcriptome. Each iModulon and gene has its own interactive panel, with clickable, hoverable, and downloadable graphs and tables	[46]
KBase	The US Department of Energy's systems biology knowledge base supports the sharing, integration, and analysis of data on microbes, plants, and their communities. The knowledge base integrates public and user data on genomes, compounds, and reactions, linking these diverse data types with a range of analysis features via a web-based user interface	[47]
SwissLipids	A knowledge base on lipids and their biology, providing curated information on lipid structures and metabolism. This knowledge is used to generate a virtual library of feasible lipid structures, which are organized in a hierarchical classification linking mass spectrometry outputs with all possible lipid structures, metabolic reactions, and enzymes	[48]

应知识库 Rhea[42]和 Reactome[43]，基因知识库 Gene Ontology (GO)[44]，蛋白知识库 UniProt[45]，以及调控知识库 iModulonDB[46]等。这些知识库都有相同的共同点：数据组织上结合了非机构化数据，整合了元数据和语义信息。例如，UniProt 从 2011 年起变更为 UniProt Knowledgebase，除了蛋白序列信息外，还添加了与这些序列相关的功能、结构、亚细胞位置、相互作用等信息，提供了蛋白质序列和功能信息的统一视图，成为蛋白质知识的中心枢纽。在 2023 年发布的最新版中，UniProt 还添加了 325 250 条亚细胞定位图像[45]。

尽管这些知识库在数据整合和信息丰富性方面取得了显著进展，但在知识的推理和发现方面仍然存在一定的局限性。首先，Rhea 是一个手动注释的生物化学反应知识库，提供详细的反应机制信息和跨数据库整合的能力[42]。尽管数据准确性和可靠性高，但 Rhea 在实现自动化推理和新反应发现方面的能力有限，主要依赖于手动注释和已有知识的整合，这意味着 Rhea 不能独立进行复杂的知识推理或发现新反应途径。GO 提供标准化的基因产品命名法，涵盖功能、细胞位置和相关生物过程，促进了数据共享和比较[44]。尽管 GO 在基因功能注释和生物信息学分析中应用广泛，但其知识推理和新功能发现能力依赖于外部数据源和算法，无

法独立进行复杂的功能预测和关联分析。例如，GO 需要结合基因表达数据和其他生物信息学工具才能进行复杂的功能预测和关联分析，这使得其在发现和解释新基因功能时存在一定的局限性。其次，UniProt 作为蛋白质知识的中心枢纽，整合了蛋白质的序列、功能、结构、亚细胞位置和相互作用等信息，确保数据高质量和可靠性[45]。然而，UniProt 的推理能力主要依赖于用户查询和已有数据的关联分析，缺乏自主发现新知识的机制。最后，iModulonDB 是一个整合转录调控模块的数据库，提供基因调控网络的详细信息[46]。尽管数据丰富，用户界面友好，但其知识发现和推理能力依然有限，主要依赖于已有数据的关联和模式识别，它可以识别已知的调控模式和网络结构，但难以独立发现新的调控机制或预测未知的调控关系。

总之，虽然当前已发表的知识库在数据整合和信息提供方面取得了显著进展，但在自主知识推理和新知识发现方面仍有很大的提升空间。

2.2 生物制造知识库的创新趋势与展望

随着生物制造领域的不断进步，知识库的作用变得愈发重要。未来的知识库不仅需要整合数据存储、分析和推理功能，还需要适应更复杂和动态的生物制造环境。通过结合先进的人工智能技术，特别是大语言模型(large language models, LLMs)，知识库将能够更高效地处理海量数据，并从中推导出新的科学结论，为生物制造领域的创新提供强大支持。

2.2.1 深度学习与强化学习的应用

深度学习技术如卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)在蛋白质结构预测[49]、基因序列功能预测[50]等领域已经取得了显著成果。未来，这些技术将在更广泛的生物制造应用中发挥作用。通过识别生物数据中的潜在模式，深度学习技术将为生物制造的精准设计和优化提供有力支持。

强化学习(reinforcement learning, RL)也将在生物制造的过程优化中扮演关键角色。通过模拟和奖励机制，强化学习可以自动调整生物制造过程中的工程参数，从而实现最优控制，提升生产效率和质量[51-52]。此外，RL 可以通过虚拟实验的方式，探索不同的生物制造路径，这对于实际生产中的参数优化具有重要意义。

2.2.2 平台化与多工具集成的发展趋势

为了提升研究效率，未来的知识库将朝着平台化和多工具集成的方向发展。集成各种分析工具和算法的平台可以为研究人员提供统一的界面，使他们能够在同一环境下进行数据整合、分析和推理。例如，Cytoscape 作为一个集成多源数据和提供交互式分析环境的平台，显著降低了因数据传输和工具切换所需的时间成本，进而提升了研究效率[53]。通过这种无缝的数据整合，研究人员可以将注意力集中在科学探索上，而非深陷数据处理的繁琐细节，这极大地增强了他们的创新能力。

2.2.3 与大语言模型的融合

人工智能在生物制造中的应用已经展现出巨大潜力，未来这一趋势将进一步加速。LLMs，如 GPT 系列，不仅能够理解和生成自然语言，还能够处理和分析生物学文献、专利数据以及实验记录等非结构化数据。LLM 可以自动化地阅读和分析大量的科研文献，提炼出生物反应器条件和代谢工程改造策略[37]，并将其整合到现有的知识库中，使研究人员能够快速获取相关信息，从而加速研究进程。

总之，人工智能，尤其是大语言模型，在生物制造领域的应用将继续推动该领域的创新和发展。随着技术的不断进步，知识库将变得

更加智能和自主，能够在更大程度上辅助研究人员进行科学发现和技术开发。通过不断增强知识库的推理能力和新知识发现能力，生物制造领域将迎来更多突破性进展，推动整个行业的持续创新。

3 面向生物制造的大模型

LLMs 的引入标志着生物制造领域数据库和知识库发展迈入了一个全新阶段。LLMs 不仅在处理和分析海量生物数据方面展现出卓越的能力，还具有强大的推理和预测功能。这些模型的应用能够与现有知识库深度结合，实现更为精准的预测和复杂生物过程的模拟，从而显著加速新知识的发现与生物制造技术的革新。

通过对更大规模、更复杂的数据集进行处理，LLMs 能够高度自动化地进行知识推理和新知识的发现。这不仅提升了知识库的智能化程度，还为生物制造领域带来了前所未有的研究和应用潜力。随着 LLMs 技术的不断成熟，其在生物制造中的应用将进一步扩展，为未来生物制造的创新奠定坚实的基础。

3.1 基于大模型的生物制造知识获取

随着合成生物学、基因组学和生物信息学的迅速发展，生物制造领域中的数据和知识量不断增加。这些信息不仅分散在科学文献中，还包括实验室数据、专利文件和技术报告等多种形式。如何高效地从这些多源数据中提取、整合和应用知识，成为提升生物制造效率和推动创新的关键问题。基于 LLMs 的生物制造知识获取正是解决这一问题的重要途径。

近年来，基于 LLMs 的自然语言处理技术在各个领域展示了其强大的信息处理能力。这些模型能够理解和生成自然语言文本，从大规模数据中提取模式、关系和趋势，为复杂问题的解决提供了新的途径[54]。首先，LLMs 在将非标准化文本转化为标准化数据的转化中起着关键作用。手动从大量文章中提取数据不仅耗时耗力，而且由于不同研究中的数据格式不统一，容易出现人为错误和数据质量不一致等问题[55]。自然语言处理工具(如 GPT-4)的发展可以加速提取与复杂菌株工程和生物反应器条件下微生物性能相关的信息，展示了生成式人工智能在更高效地收集数据和快速整理信息用于合成生物学研究方面的潜力，从而加快生物制造开发的 DBTL 循环[37]。例如，Xiao 等[37]使用 GPT-4 从有关解脂耶氏酵母的文章中提取了生物过程条件、代谢途径和基因工程方法等知识，首次将 GPT 与知识工程和机器学习相结合，用于预测微生物细胞工厂。

此外，这些利用 LLMs 从分散的文献和实验数据中提取的结构化信息可以生成用于数据库和知识库的标准化数据，从而弥补了传统数据库的不足[56-57]。LLMs 还能够自动化更新和扩展数据库和知识库，使其能够及时反映领域内的最新研究成果和技术发展[58]。通过这种方式，研究人员可以更方便地获取和利用已有的知识，加速研究进程和创新。

总之，随着技术的不断进步和应用场景的拓展，LLMs 有望成为生物制造领域的重要智能化工具，为推动科学研究和工程实践带来新的可能性和机遇。

3.2 基于大模型的生物制造知识生成

尽管 LLMs 展示了人类般的语言理解、推理和生成能力，但其在复杂领域的信息检索和知识生成任务中仍面临诸多挑战。例如，已有研究表明，LLMs 在文本到结构化查询语言(structured query language, SQL)任务中的性能比人类专家低约 40%[59]，这凸显了在高效生成和应用领域知识时的困难。为了克服这些挑

战，研究人员提出了各种优化方法。首先，Fine-tune 技术在通用预训练模型的基础上，通过对特定领域数据进行微调，使模型能够更好地适应特定任务和问题，该方法已在医学和法学等领域大模型得到应用[60-62]，并显著提升了模型回答的准确性。然而，微调的 LLMs 可能由于产生幻觉或使用过时知识而生成不准确的响应[63]。这也是微调策略在生物制造领域效果较差的原因之一。生物制造领域面临两大挑战。首先，缺乏高质量的微调数据集。生物制造领域的数据往往分散且不标准化，这使得获取和整理足够的高质量数据集变得困难。其次，生物制造的快速发展意味着用于微调的数据可能很快变得过时，无法完全满足领域不断发展的需求。这种快速变化使得模型难以保持最新和准确，限制了微调技术在生物制造知识生成中的应用效果。

为了解决这些挑战，Meta AI 研究人员引入了检索增强生成(retrieval-augmented generation, RAG)[64]技术，它无缝集成了信息检索与文本生成。RAG 通过从维基百科或文献等来源检索相关文档，并将其与输入提示串联起来，再利用文本生成器生成最终输出。这种方法确保了 LLMs 能够实时访问最新信息，而无需频繁重新训练，从而生成更加准确和可靠的知识[65-66]。此外，知识蒸馏和适配器模块的应用，可以在不完全重新训练的情况下，细化和调整预训练模型以适应特定任务[67-68]。这些技术使得 LLMs 能够更好地处理复杂数据并生成有价值的知识。

总之，基于 Fine-tune 和 RAG 等技术的生物制造大模型开发，不仅能够应对生物制造领域复杂性和特异性的挑战，还能够推动技术创新和生产效率的提升。这些技术方向的不断发展和应用，将为生物制造领域带来更广阔的发展和应用前景。

4 展望

随着生物制造技术的不断进步和人工智能技术的快速发展，数据库、知识库和大模型在生物制造领域将扮演越来越关键的角色。这些工具不仅为生物制造的设计、研发、生产和市场推广提供了强大支持，还在科学研究和工业应用中发挥重要作用。

未来的生物制造数据库和知识库将更加全面、智能和高效。通过积累涵盖基因组、蛋白质结构、代谢通路、生产过程监测等多维数据，这些数据库和知识库能够为生物制造的工艺优化、代谢工程设计和新型菌种构建提供深度支撑。在实际应用中，数据的准确性和可靠性至关重要，因此数据库和知识库的维护需要引入自动化的更新机制和严格的质量控制标准。例如，在酶优化过程中，知识库可根据实时数据更新实验参数，帮助研究人员迅速找到最佳酶催化条件。同时，统一的数据标准和智能分析工具的整合也非常关键。通过机器学习算法，可以对生产过程中产生的大量数据进行高效分析，实现代谢流量的优化与瓶颈反应的识别，为工艺改进提供方向。此外，面对全球化的研究和产业需求，生物制造数据库和知识库还需考虑到数据的隐私和安全问题，并需要符合伦理要求的使用和共享规范。这些举措不仅能够提升数据的可信度和可用性，也有助于促进全球生物制造研究社区的合作与交流。

未来的生物制造中，基于大模型的技术将成为重要的驱动力。LLMs，如基于 RAG 技术的生物制造大模型，结合了信息检索、注意力机制和生成模型，能够智能化地处理复杂的生物数据和工艺信息，为知识获取、数据解析和模式识别提供了新的解决方案。在菌种构建和优化的实际应用中，大语言模型可以通过挖掘

大量文献和实验数据，推荐最优的基因编辑策略或代谢途径调整方案，显著缩短研发周期[32]。而在蛋白质结构预测方面，AlphaFold3 等大模型取得了重要突破，通过更高的预测精度，为生物制造中的蛋白质功能改造和优化提供了宝贵的结构信息[69]。此外，类似于 RoseTTAFold[49] 和 DeepAccNet[70]等模型在提高蛋白质结构预测精度和稳定性评估方面也显示了卓越的性能，为蛋白质设计和改造提供了坚实的技术支撑。未来的研究应集中在提升各类大模型的数据质量和解释性，以应对数据质量不均、模型解释性不足和计算资源需求高等挑战。通过改进模型的训练算法、优化模型结构和加强计算能力，可以进一步提升模型的预测能力和实用性，从而推动生物制造领域的技术前沿。

综上所述，未来的生物制造将依赖于先进的数据库、知识库和大模型技术，这些技术将成为促进生物制造科学研究、技术创新和工业应用的重要工具和平台。通过持续的技术创新和跨学科合作，有望实现更智能化、自动化的生物制造过程，为人类社会的可持续发展和健康做出更大的贡献。

作者贡献声明

毛志涛：方案设计、经费支持、初稿写作、稿件润色修改；廖小平：监督指导、稿件润色修改；马红武：监督指导、稿件润色修改。

作者利益冲突公开声明

作者声明没有任何可能会影响本文所报告工作的已知经济利益或个人关系。

REFERENCES

[1] 侯隽. 生物制造将成为新增长引擎[J]. 中国经济周刊, 2024(6): 38-40.
HOU J. Bio-manufacturing will become a new growth engine[J]. China Economic Weekly, 2024(6): 38-40. (in Chinese).

[2] 张媛媛, 曾艳, 王钦宏. 合成生物制造进展[J]. 合成生物学, 2021, 2(2): 145-160.
ZHANG YY, ZENG Y, WANG QH. Advances in synthetic biomanufacturing[J]. Synthetic Biology Journal, 2021, 2(2): 145-160 (in Chinese).

[3] 程强. 生物制造：一场新的产业革命[N]. 中国石化报, 2023-09-06(5).
CHENG, Q. Biomanufacturing: a new industrial revolution[N]. China Petrochemical News, 2023-09-06(5) (in Chinese).

[4] 马延和. 生物制造产业是生物经济重点发展方向[J]. 中国生物工程杂志, 2022, 42(5): 4-5.
Ma YH. Biomanufacturing is a key area to accelerate the development of bioeconomy[J]. China Biotechnology, 2022, 42(5): 4-5 (in Chinese).

[5] 高振, 段珺, 黄英明, 种国双. 中国生物制造产业与科技现状及对策建议[J]. 科学管理研究, 2019, 37(5): 68-75.
GAO Z, DUAN J, HUANG YM, ZHONG GS. Research on current situation of bio manufacturing industry and science and technology in China[J]. Scientific Management Research, 2019, 37(5): 68-75 (in Chinese).

[6] SAYERS EW, CAVANAUGH M, CLARK K, PRUITT KD, SHERRY ST, YANKIE L, KARSCH-MIZRACHI I. GenBank 2024 update[J]. Nucleic Acids Research, 2024, 52(D1): D134-D137.

[7] HARRISON PW, AMODE MR, AUSTINE-ORIMOLOYE O, AZOV AG, BARBA M, BARNES I, BECKER A, BENNETT R, BERRY A, BHAI J, BHURJI SK, BODDU S, BRANCO LINS PR, BROOKS L, RAMARAJU SB, CAMPBELL LI, MARTINEZ MC, CHARKHCHI M, CHOUGULE K, COCKBURN A, et al. Ensembl 2024[J]. Nucleic Acids Research, 2024, 52(D1): D891-D899.

[8] SAMARAS P, SCHMIDT T, FREJNO M, GESSULAT S, REINECKE M, JARZAB A, ZECHA J, MERGNER J, GIANSANTI P, EHRLICH HC, AICHE S, RANK J, KIENEGGER H, KRCMAR H, KUSTER B, WILHELM M. ProteomicsDB: a multi-omics and multi-organism resource for life science research[J]. Nucleic Acids Research, 2019: gkz974.

[9] BHARGAVA S, JANKOWSKI J. The PRIDE database resources in 2023[J]. Nephrology, Dialysis, Transplantation, 2023, 39(1): 4-6.

[10] HUANG QY, SZKLARCZYK D, WANG MC, SIMONOVIC M, von MERING C. Pa$_x$Db 5.0: curated protein quantification data suggests adaptive proteome changes in yeasts[J]. Molecular & Cellular Proteomics, 2023, 22(10): 100640.

[11] DESIERE F, DEUTSCH EW, KING NL, NESVIZHSKII AI, MALLICK P, ENG J, CHEN S, EDDES J, LOEVENICH SN, AEBERSOLD R. The PeptideAtlas project[J]. Nucleic Acids Research, 2006, 34(database issue): D655-D658.

[12] BURLEY SK, BHIKADIYA C, BI CX, BITTRICH S, CHAO H, CHEN L, CRAIG PA, CRICHLOW GV, DALENBERG K, DUARTE JM, DUTTA S, FAYAZI M, FENG ZK, FLATT JW, GANESAN S, GHOSH S, GOODSELL DS, GREEN RK, GURANOVIC V, HENRY J, et al. RCSB protein data bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of

proteins from artificial intelligence/machine learning[J]. Nucleic Acids Research, 2023, 51(D1): D488-D508.
[13] HASTINGS J, OWEN G, DEKKER A, ENNIS M, KALE N, MUTHUKRISHNAN V, TURNER S, SWAINSTON N, MENDES P, STEINBECK C. ChEBI in 2016: improved services and an expanding collection of metabolites[J]. Nucleic Acids Research, 2016, 44(D1): D1214-D1219.
[14] KIM S, CHEN J, CHENG TJ, GINDULYTE A, HE J, HE SQ, LI QL, SHOEMAKER BA, THIESSEN PA, YU B, ZASLAVSKY L, ZHANG J, BOLTON EE. PubChem 2023 update[J]. Nucleic Acids Research, 2023, 51(D1): D1373-D1380.
[15] YUREKTEN O, PAYNE T, TEJERA N, AMALADOSS FX, MARTIN C, WILLIAMS M, O'DONOVAN C. MetaboLights: open data repository for metabolomics[J]. Nucleic Acids Research, 2024, 52(D1): D640-D646.
[16] KANEHISA M, FURUMICHI M, SATO Y, KAWASHIMA M, ISHIGURO-WATANABE M. KEGG for taxonomy-based analysis of pathways and genomes[J]. Nucleic Acids Research, 2023, 51(D1): D587-D592.
[17] CASPI R, BILLINGTON R, KESELER IM, KOTHARI A, KRUMMENACKER M, MIDFORD PE, ONG WK, PALEY S, SUBHRAVETI P, KARP PD. The MetaCyc database of metabolic pathways and enzymes: a 2019 update[J]. Nucleic Acids Research, 2020, 48(D1): D445-D453.
[18] CHANG A, JESKE L, ULBRICH S, HOFMANN J, KOBLITZ J, SCHOMBURG I, NEUMANN-SCHAAL M, JAHN D, SCHOMBURG D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates[J]. Nucleic Acids Research, 2021, 49(D1): D498-D508.
[19] JEWISON T, SU YL, DISFANY FM, LIANG YJ, KNOX C, MACIEJEWSKI A, POELZER J, HUYNH J, ZHOU Y, ARNDT D, DJOUMBOU Y, LIU YF, DENG L, GUO AC, HAN B, PON A, WILSON M, RAFATNIA S, LIU P, WISHART DS. SMPDB 2.0: big improvements to the small molecule pathway database[J]. Nucleic Acids Research, 2014, 42(D1): D478-D484.
[20] AGRAWAL A, BALCI H, HANSPERS K, COORT SL, MARTENS M, SLENTER DN, EHRHART F, DIGLES D, WAAGMEESTER A, WASSINK I, ABBASSI-DALOII T, LOPES EN, IYER A, ACOSTA JM, WILLIGHAGEN LG, NISHIDA K, RIUTTA A, BASARIC H, EVELO CT, WILLIGHAGEN EL, KUTMON M, PICO AR. WikiPathways 2024: next generation pathway database[J]. Nucleic Acids Research, 2024, 52(D1): D679-D689.
[21] MORETTI S, TRAN VDT, MEHL F, IBBERSON M, PAGNI M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models[J]. Nucleic Acids Research, 2021, 49(D1): D570-D574.
[22] KING ZA, LU J, DRÄGER A, MILLER P, FEDEROWICZ S, LERMAN JA, EBRAHIM A, PALSSON BO, LEWIS NE. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models[J]. Nucleic Acids Research, 2016, 44(D1): D515-D522.
[23] MALIK-SHERIFF RS, GLONT M, NGUYEN TVN, TIWARI K, ROBERTS MG, XAVIER A, VU MT, MEN JH, MAIRE M, KANANATHAN S, FAIRBANKS EL, MEYER JP, ARANKALLE C, VARUSAI TM, KNIGHT-SCHRIJVER V, LI L, DUEÑAS-ROCA C, DASS G, KEATING SM, PARK YM, et al. BioModels: 15 years of sharing computational models in life science[J]. Nucleic Acids Research, 2019: gkz1055.
[24] SZKLARCZYK D, KIRSCH R, KOUTROULI M, NASTOU K, MEHRYARY F, HACHILIF R, GABLE AL, FANG T, DONCHEVA NT, PYYSALO S, BORK P, JENSEN LJ, von MERING C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest[J]. Nucleic Acids Research, 2023, 51(D1): D638-D646.
[25] SZKLARCZYK D, SANTOS A, von MERING C, JENSEN LJ, BORK P, KUHN M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data[J]. Nucleic Acids Research, 2016, 44(D1): D380-D384.
[26] SALGADO H, GAMA-CASTRO S, LARA P, MEJIA-ALMONTE C, ALARCÓN-CARRANZA G, LÓPEZ-ALMAZO AG, BETANCOURT-FIGUEROA F, PEÑA-LOREDO P, ALQUICIRA-HERNÁNDEZ S, LEDEZMA-TEJEIDA D, ARIZMENDI-ZAGAL L, MENDEZ-HERNANDEZ F, DIAZ-GOMEZ AK, OCHOA-PRAXEDIS E, MUÑIZ-RASCADO LJ, GARCÍA-SOTELO JS, FLORES-GALLEGOS FA, GÓMEZ L, BONAVIDES-MARTÍNEZ C, del MORAL-CHÁVEZ VM, et al. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *E. coli* K-12[J]. Nucleic Acids Research, 2024, 52(D1): D255-D264.
[27] TEIXEIRA MC, VIANA R, PALMA M, OLIVEIRA J, GALOCHA M, MOTA MN, COUCEIRO D, PEREIRA MG, ANTUNES M, COSTA IV, PAIS P, PARADA C, CHAOUIYA C, SÁ-CORREIA I, MONTEIRO PT. YEASTRACT+: a portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis[J]. Nucleic Acids Research, 2023, 51(D1): D785-D791.
[28] PARISE MTD, PARISE D, KATO RB, PAULING JK, TAUCH A, AZEVEDO VAC, BAUMBACH J. CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks[J]. Scientific Data, 2020, 7(1): 142.
[29] WITTIG U, REY M, WEIDEMANN A, KANIA R, MÜLLER W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics[J]. Nucleic Acids Research, 2018, 46(D1): D656-D660.
[30] REIMER LC, VETCININOVA A, CARBASSE JS, SÖHNGEN C, GLEIM D, EBELING C, OVERMANN J. BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis[J]. Nucleic Acids Research, 2019, 47(D1): D631-D636.
[31] CHAUDHARY A, BHALLA S, PATIYAL S, RAGHAVA GPS, SAHNI G. FermFooDb: a database of bioactive peptides derived from fermented foods[J]. Heliyon, 2021, 7(4): e06668.
[32] SUN ZH, HARRIS HMB, McCANN A, GUO CY, ARGIMÓN S, ZHANG WY, YANG XW, JEFFERY IB, COONEY JC, KAGAWA TF, LIU WJ, SONG YQ,

SALVETTI E, WROBEL A, RASINKANGAS P, PARKHILL J, REA MC, O'SULLIVAN O, RITARI J, DOUILLARD FP, et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera[J]. Nature Communications, 2015, 6: 8322.
[33] PETROVIĆ D, FRANK D, KAMERLIN SCL, HOFFMANN K, STRODEL B. Shuffling active site substate populations affects catalytic activity: the case of glucose oxidase[J]. ACS Catalysis, 2017, 7(9): 6188-6197.
[34] KARLSEN E, SCHULZ C, ALMAAS E. Automated generation of genome-scale metabolic draft reconstructions based on KEGG[J]. BMC Bioinformatics, 2018, 19(1): 467.
[35] CLOUGH E, BARRETT T, WILHITE SE, LEDOUX P, EVANGELISTA C, KIM IF, TOMASHEVSKY M, MARSHALL KA, PHILLIPPY KH, SHERMAN PM, LEE H, ZHANG NG, SEROVA N, WAGNER L, ZALUNIN V, KOCHERGIN A, SOBOLEVA A. NCBI GEO archive for gene expression and epigenomics data sets: 23-year update[J]. Nucleic Acids Research, 2024, 52(D1): D138-D144.
[36] SAXENA P, RAUNIYAR S, THAKUR P, SINGH RN, BOMGNI A, ALABA MO, TRIPATHI AK, GNIMPIEBA EZ, LUSHBOUGH C, SANI RK. Integration of text mining and biological network analysis: identification of essential genes in sulfate-reducing bacteria[J]. Frontiers in Microbiology, 2023, 14: 1086021.
[37] XIAO ZY, LI WY, MOON H, ROELL GW, CHEN YX, TANG YJ. Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology[J]. ACS Synthetic Biology, 2023, 12(10): 2973-2982.
[38] KUMAR P, HALAMA A, HAYAT S, BILLING AM, GUPTA M, YOUSRI NA, SMITH GM, SUHRE K. MetaRNA-seq: an interactive tool to browse and annotate metadata from RNA-seq studies[J]. BioMed Research International, 2015, 2015: 318064.
[39] FABREGAT A, KORNINGER F, VITERI G, SIDIROPOULOS K, MARIN-GARCIA P, PING PP, WU GM, STEIN L, D'EUSTACHIO P, HERMJAKOB H. Reactome graph database: efficient access to complex pathway data[J]. PLoS Computational Biology, 2018, 14(1): e1005968.
[40] MAO ZT, WANG RY, LI HR, HUANG YX, ZHANG Q, LIAO XP, MA HW. ERMer: a serverless platform for navigating, analyzing, and visualizing *Escherichia coli* regulatory landscape through graph database[J]. Nucleic Acids Research, 2022, 50(W1): W298-W304.
[41] MESSINA A, PRIBADI H, STICHBURY J, BUCCI M, KLARMAN S, URSO A. BioGrakn: a Knowledge Graph-based Semantic Database for Biomedical Sciences[M]//Complex, Intelligent, and Software Intensive Systems. Cham: Springer International Publishing, 2017: 299-309.
[42] BANSAL P, MORGAT A, AXELSEN KB, MUTHUKRISHNAN V, COUDERT E, AIMO L, HYKA-NOUSPIKEL N, GASTEIGER E, KERHORNOU A, NETO TB, POZZATO M, BLATTER MC, IGNATCHENKO A, REDASCHI N, BRIDGE A. Rhea, the reaction knowledgebase in 2022[J]. Nucleic Acids Research, 2022, 50(D1): D693-D700.
[43] MILACIC M, BEAVERS D, CONLEY P, GONG CQ, GILLESPIE M, GRISS J, HAW R, JASSAL B, MATTHEWS L, MAY B, PETRYSZAK R, RAGUENEAU E, ROTHFELS K, SEVILLA C, SHAMOVSKY V, STEPHAN R, TIWARI K, VARUSAI T, WEISER J, WRIGHT A, et al. The reactome pathway knowledgebase 2024[J]. Nucleic Acids Research, 2024, 52(D1): D672-D678.
[44] CONSORTIUM TGO, ALEKSANDER SA, BALHOFF J, CARBON S, CHERRY JM, DRABKIN HJ, EBERT D, FEUERMANN M, GAUDET P, HARRIS N L. The Gene Ontology knowledgebase in 2023[J]. Genetics 2023, 224(1): iyad031.
[45] CONSORTIUM U. UniProt: the universal protein knowledgebase in 2023[J]. Nucleic Acids Research, 2023, 51(D1): D523-D531.
[46] RYCHEL K, DECKER K, SASTRY AV, PHANEUF PV, POUDEL S, PALSSON BO. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning[J]. Nucleic Acids Research, 2021, 49(D1): D112-D120.
[47] ARKIN AP, COTTINGHAM RW, HENRY CS, HARRIS NL, STEVENS RL, MASLOV S, DEHAL P, WARE D, PEREZ F, CANON S, SNEDDON MW, HENDERSON ML, RIEHL WJ, MURPHY-OLSON D, CHAN SY, KAMIMURA RT, KUMARI S, DRAKE MM, BRETTIN TS, GLASS EM, et al. KBase: the United States department of energy systems biology knowledgebase[J]. Nature Biotechnology, 2018, 36(7): 566-569.
[48] AIMO L, LIECHTI R, HYKA-NOUSPIKEL N, NIKNEJAD A, GLEIZES A, GÖTZ L, KUZNETSOV D, DAVID FPA, GISOU van der GOOT F, RIEZMAN H, BOUGUELERET L, XENARIOS I, BRIDGE A. The SwissLipids knowledgebase for lipid biology[J]. Bioinformatics, 2015, 31(17): 2860-2866.
[49] BAEK M, DiMAIO F, ANISHCHENKO I, DAUPARAS J, OVCHINNIKOV S, GR L, WANG J, CONG Q, LN K, SCHAEFFER RD, MILLAN C, PARK H, ADAMS C, GLASSMAN CR, DeGIOVANNI A, PEREIRA JH, RODRIGUES AV, van DIJK A, EBRECHT AC, OPPERMAN DJ, et al. Accurate prediction of protein structures and interactions using a three-track neural network[J]. Science, 2021, 373(6557): 871-876.
[50] ZHANG X, XIAO WX, XIAO WJ. DeepHE: accurately predicting human essential genes based on deep learning[J]. PLoS Computational Biology, 2020, 16(9): e1008229.
[51] KOCH M, DUIGOU T, FAULON JL. Reinforcement learning for bioretrosynthesis[J]. ACS Synthetic Biology, 2020, 9(1): 157-168.
[52] ZHOU ZP, KEARNES S, LI L, ZARE RN, RILEY P. Optimization of molecules *via* deep reinforcement learning[J]. Scientific Reports, 2019, 9(1): 10752.
[53] SHANNON P, MARKIEL A, OZIER O, BALIGA NS, WANG JT, RAMAGE D, AMIN ND, SCHWIKOWSKI B, IDEKER T. Cytoscape: a software environment for integrated models of biomolecular interaction networks[J]. Genome Research, 2003, 13(11): 2498-2504.
[54] OpenAI. GPT-4 technical report[EB/OL]. [2024-03-15]. https://arxiv.org/abs/2303.08774.
[55] WINKLER JD, HALWEG-EDWARDS AL, GILL RT.

The LASER database: formalizing design rules for metabolic engineering[J]. Metabolic Engineering Communications, 2015, 2: 30-38.
[56] PATINY L, GODIN G. Automatic extraction of FAIR data from publications using LLM[EB/OL]. ChemRxiv, 2023.
[57] LI N, ZAHRA S, BRITO M, FLYNN C, GÖRNERUP O, WOROU K, KURFALI M, MENG CJ, THIERY W, ZSCHEISCHLER J, MESSORI G, NIVRE J. Using LLMs to build a database of climate extreme impacts[C]. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change, 2024.
[58] PAN JZ, RAZNIEWSKI S, KALO JC, SINGHANIA S, CHEN JY, DIETZE S, JABEEN H, OMELIYANENKO J, ZHANG W, LISSANDRINI M, BISWAS R, de MELO G, BONIFATI A, VAKAJ E, DRAGONI M, GRAUX D. Large language models and knowledge graphs: opportunities and challenges[EB/OL]. 2023: 2308.06374.
[59] LI J, HUI B, QU G, YANG J, LI B, LI B, WANG B, QIN B, GENG R, HUO N. Can llm already serve as a database interface? A big bench for large-scale database grounded text-to-sqls[J]. Advances in Neural Information Processing Systems, 2024, 36.
[60] LI YX, LI ZH, ZHANG K, DAN RL, JIANG S, ZHANG Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge[J]. Cureus, 2023, 15(6): e40895.
[61] WANG HC, LIU C, XI NW, QIANG ZW, ZHAO SD, QIN B, LIU T. HuaTuo: tuning LLaMA model with Chinese medical knowledge[EB/OL]. [2023-04-14]. https://arxiv.org/abs/2304.06975.
[62] NGUYEN HT. A brief report on LawGPT 1.0: a virtual legal assistant based on GPT-3[EB/OL]. [2023-02-11]. https://arxiv.org/abs/2302.05729.
[63] YUE SB, CHEN W, WANG SY, LI BX, SHEN CC, LIU SJ, ZHOU YX, XIAO Y, YUN S, HUANG XJ, WEI ZY. DISC-LawLLM: fine-tuning large language models for intelligent legal services[EB/OL]. [2023-09-20]. https://arxiv.org/abs/2309.11325.
[64] LEWIS P, PEREZ E, PIKTUS A, PETRONI F, KARPUKHIN V, GOYAL N, KÜTTLER H, LEWIS M, YIH WT, ROCKTÄSCHEL T, RIEDEL S, KIELA D, LEWIS P, PEREZ E, PIKTUS A, PETRONI F, KARPUKHIN V, GOYAL N, KÜTTLER H, LEWIS M, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020.
[65] TANG YX, YANG Y. MultiHop-RAG: benchmarking retrieval-augmented generation for multi-hop queries [EB/OL]. [2024-01-27]. https://arxiv.org/abs/2401.15391.
[66] BORGEAUD S, MENSCH A, HOFFMANN J, Cai T, RUTHERFORD E, MILLICAN K, DRIESSCHE G, LESPIAU J, DAMOC B, CLARK A, CASAS D, GUY A, MENICK J, RING R, HENNIGAN T, HUANG S, MAGGIORE L, JONES C, CASSIRER A, BROCK A, et al. Improving Language Models by Retrieving from Trillions of Tokens[C]. In Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research; 2022.
[67] ZHANG K, TAO CY, SHEN T, XU C, GENG XB, JIAO BX, JIANG DX, ZHANG K, TAO CY, SHEN T, XU C, GENG XB, JIAO BX, JIANG DX. LED: lexicon-enlightened dense retriever for large-scale retrieval[C]. Proceedings of the ACM Web Conference 2023.30 April 2023, Austin, TX, USA. ACM, 2023: 3203-3213.
[68] PAL V, LASSANCE C, DÉJEAN H, CLINCHANT S. Parameter-efficient Sparse Retrievers and Rerankers Using Adapters[M]. Advances in Information Retrieval. Cham: Springer Nature Switzerland, 2023: 16-31.
[69] ABRAMSON J, ADLER J, DUNGER J, EVANS R, GREEN T, PRITZEL A, RONNEBERGER O, WILLMORE L, BALLARD AJ, BAMBRICK J, BODENSTEIN SW, EVANS DA, HUNG CC, O'NEILL M, REIMAN D, TUNYASUVUNAKOOL K, WU Z, ŽEMGULYTĖ A, ARVANITI E, BEATTIE C, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024, 630(8016): 493-500.
[70] HIRANUMA N, PARK H, BAEK M, ANISHCHENKO I, DAUPARAS J, BAKER D. Improved protein structure refinement guided by deep learning based accuracy estimation[J]. Nature Communications, 2021, 12(1): 1340.