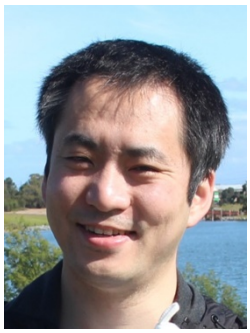
·人工细胞智能设计再造·

朱岩 中国科学院天津工业生物技术研究所研究员、博士生导师，中国科学院引才计划入选者。主要聚焦微生物系统生物学研究，整合多组学与数字细胞分析方法，“干”“湿”结合，深入探索微生物生长代谢与胁迫耐受的分子机制。在 *Advanced Science*、*Nature Communications*、*Cell Reports*、*Journal of Infection*、*Trends in Microbiology* 等国际 SCI 期刊上发表论文 96 篇，*h* 因子 35。担任 *International Journal of Antimicrobial Agents* 等期刊编委。



细胞生命过程数学刻画建模

朱岩[1,2*]，孙际宾[1,2,3]

1 中国科学院天津工业生物技术研究所系统生物学中心，天津 300308
2 低碳合成工程生物学全国重点实验室，天津 300308
3 国家合成生物技术创新中心，天津 300308

朱岩, 孙际宾. 细胞生命过程数学刻画建模[J]. 生物工程学报, 2025, 41(3): 1052-1078.
ZHU Yan, SUN Jibin. Mathematical modelling for cellular processes[J]. Chinese Journal of Biotechnology, 2025, 41(3): 1052-1078.

摘 要：生物制造是使用工程细胞实现化学品、医药、能源、材料产品的规模化生产，也是应对全球环境危机，实现“双碳”目标，推动经济社会绿色转型的一种新兴生产力。高效设计并构建工程细胞需要精准、全面的数字细胞模型。测序仪、质谱、光谱、微流控等前沿装备的迭代升级，数据科学、人工智能、自动化等高新技术集群的突破性进展，使得细胞组分动态变化的精确测量、海量生物数据的快速获取、细胞过程数理模型的精准构建成为可能。本文系统归纳了系统生物学中细胞建模的数理框架：首先从网络拓扑、随机过程、动力学方程等维度剖析了常见数学模型架构及其使用范围，继而分类评述了生长分裂、形态发生、DNA 复制、转录调控、生化代谢、信号转导、群体感应等单一细胞过程的建模策略，重点探讨了整合多个细胞过程构建全细胞模型的研究进展，最后讨论了数据不足、机制解析不充分、多维数据整合困难、计算复杂度指数增长等限制细胞生命过程数学刻画的若干关键挑战。本文汇总了系统生物学中细胞生命过程精确模拟的数理基础，增进了对细胞运作分子机制的理解，对未来工程生物的设计与构建具有重要意义。

关键词：生物制造；细胞生命过程；数学模型；系统生物学；全细胞模型

资助项目：国家重点研发计划(2023YFA0913903)；天津市合成生物技术创新能力提升行动(TSBICIP-CXRC-073, TSBICIP-PTJJ-012)
This work was supported by the National Key Research and Development Program of China (2023YFA0913903) and the Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (TSBICIP-CXRC-073, TSBICIP-PTJJ-012).
*Corresponding author. E-mail: zhuyan@tib.cas.cn
Received: 2025-01-22; Accepted: 2025-03-06; Published online: 2025-03-10

Mathematical modelling for cellular processes

ZHU Yan[1,2*], SUN Jibin[1,2,3]

1 Systems Biology Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China
2 National Key Laboratory of Engineering Biology for Low-carbon Manufacturing, Tianjin 300308, China
3 National Center of Technology Innovation for Synthetic Biology, Tianjin 300308, China

Abstract: Biomanufacturing harnesses engineered cells for the large-scale production of biochemicals, biopharmaceuticals, biofuels, and biomaterials, playing a vital role in mitigating global environmental crises, achieving carbon peaking and neutrality, and driving the green transformation of the economy and society. The effective design and construction of these engineered cells require precise and comprehensive computational models. Recent technological breakthroughs including high-throughput sequencing, mass spectrometry, spectroscopy, and microfluidic devices, coupled with advances in data science, artificial intelligence, and automation, have enabled the rapid acquisition of large-scale biological datasets, thereby facilitating a deeper understanding of cellular dynamics and the construction of mechanism-based models with enhanced accuracy. This review systematically summarises the mathematical frameworks employed in cellular modelling. It begins by evaluating prevalent mathematical paradigms, such as network topology analyses, stochastic processes, and kinetic equations, critically assessing their applicability across various contexts. The discussion then categorises modelling strategies for specific cellular processes, including cellular growth and division, morphogenesis, DNA replication, transcriptional regulation, metabolism, signal transduction, and quorum sensing. We also examine the recent progress in developing whole-cell models through the integration of diverse cellular processes. The review concludes by addressing key challenges such as data scarcity, unknown mechanisms, multi-dimensional data integration, and exponentially escalating computational complexity. Overall, this work consolidates the mathematical models for the precise simulation of cellular processes, thereby enhancing our understanding of the molecular mechanisms governing cellular functions and contributing to the future design and optimisation of engineered organisms.

Keywords: biomanufacturing; cellular process; mathematical modelling; systems biology; whole-cell model

生物制造作为前沿科技，正逐步展示其在推动可持续发展、绿色低碳经济以及产业链安全方面的巨大潜力[1]。运用合成生物学的学习-设计-构建-测试(learn-design-build-test, LDBT)方法论迭代升级工程细胞，可以高效转化一碳原料、可再生生物质资源与废弃物，规模化生产化学品、医药产品、食品饲料、生物燃料和生物材料[2-3]。高通量测序、高性能质谱、光谱、微流控等高新技术的迅猛发展，为精确测量细胞组分动态变化、快速获取海量生物数据提供了重要支撑[4-7]。数据科学、人工智能、自动化的迅猛发展，正在加速细胞生命过程的机制解析与工程生物的智能设计[8]。

数学模型是合成生物学和系统生物学的交

☏：010-64807509 ✉：cjb@im.ac.cn

汇点，也是工程细胞设计与构建的基础。传统生物学研究大多侧重于生物现象描述，缺乏定量分析与规律总结。模型使用数学语言刻画DNA复制、转录、翻译、大分子组装、代谢、膜形成、信号转导、生长分裂等细胞生命过程，定量分析复杂生物系统，深入探究生命活动基本规律，发现自组织、涌现、适应等复杂系统特征，预测工程细胞改造靶点，设计工程细胞构建策略，实现生物过程放大[9]。

构建机制模型描述细胞生命过程，遵循着简化、抽象、动态、可扩展、模块化、可验证、可重复等若干基本原则，但也存在着一些问题，包括：(1) 测量数据不足，无法充分捕捉生物系统的复杂性；(2) 细胞生命过程的分子机制不明确，依赖大量的假设和近似；(3) 孤立描述单一生物过程，忽略了细胞内多过程间的复杂相互作用；(4) 模型结构过于复杂，陷入复杂度灾难(curse of complexity)；(5) 包含大量经验参数，往往由体外实验获得或其他模式生物迁移而来，不完全准确；(6) 参数搜索空间巨大，导致维度灾难(curse of dimensionality)；(7) 缺乏统一构建和模拟标准，变量和参数注释各异，模拟流程不统一，无法横向比较结果，应用难度大；(8) 实验验证不够，由于资源和时间的限制，实验验证往往被简化或跳过；(9) 准确度评判不足，不同模型采用不同的准确度评判标准，增加了预测的主观性和不一致性。这些共性问题的存在，阻碍了对细胞过程的精准模拟和对生物学基本规律的认知，也限制了工程生物的高效精准设计。随着工程生物学和计算生物学的快速发展，这些问题越来越得到重视。高通量高精度组学技术的发展，将为构建精准模型提供海量数据支持[10-14]。全细胞模型(whole cell modelling)等多尺度模型(multiscale modelling)理论的发展，开创了整合模拟多个细胞过程的新路径[15-17]。机器学习(machine learning, ML)等人工智能算法的快速更新迭代，为模型构建以及基于模型的生物设计提供了全新的策略。

本文整理归纳了细胞生命过程数学建模的技术体系，详细阐述了单一细胞生命过程的数学模型，进一步总结了多过程模型以及全细胞模型的相关方法，系统评述了这些数学模型在解析细胞运作规律、工程细胞设计中的具体应用。此外，还总结了数学建模的当前挑战，探讨了可能的解决方案，以期推动该领域的研究进展和技术革新，加速合成生物学LDBT循环，实现高效可持续的绿色生物制造。

1 模型理论

系统生物学研究复杂生命过程，使用测序仪、质谱、光谱、微流控等高性能装备高通量解析细胞内核酸、蛋白质、脂质、小分子代谢物等组分的时空分布和动态变化规律，阐明组分间的互作关系以及随之涌现的生物系统功能。数学模型是解读组学大数据、刻画细胞生命过程的核心工具，也是设计工程生物的出发点[18]。

构建机制模型首先需要明确细胞过程机理，分析系统中的关键变量，结合基因组注释、文献和数据库提出必要的简化假设，选用合适的数学架构建模，进行参数估计和准确度评估(图1A)。细胞组分理化性质存在显著差异，细胞生命过程的动力学、尺度、机制也各不相同，所需数学模型各异。常见的机制模型架构包括网络模型、约束模型、概率模型与动力学模型等[19](图1B)。网络模型本质上是图，由节点和边组成，边上的权重可表示强度、概率、化学计量系数等。基于网络模型，可以分析度分布(degree distribution)、连接度(connectivity)、中心性(centrality)、基序(motif)、模块化(modulation)、路径(path)、网络效率(network efficiency)、介数中

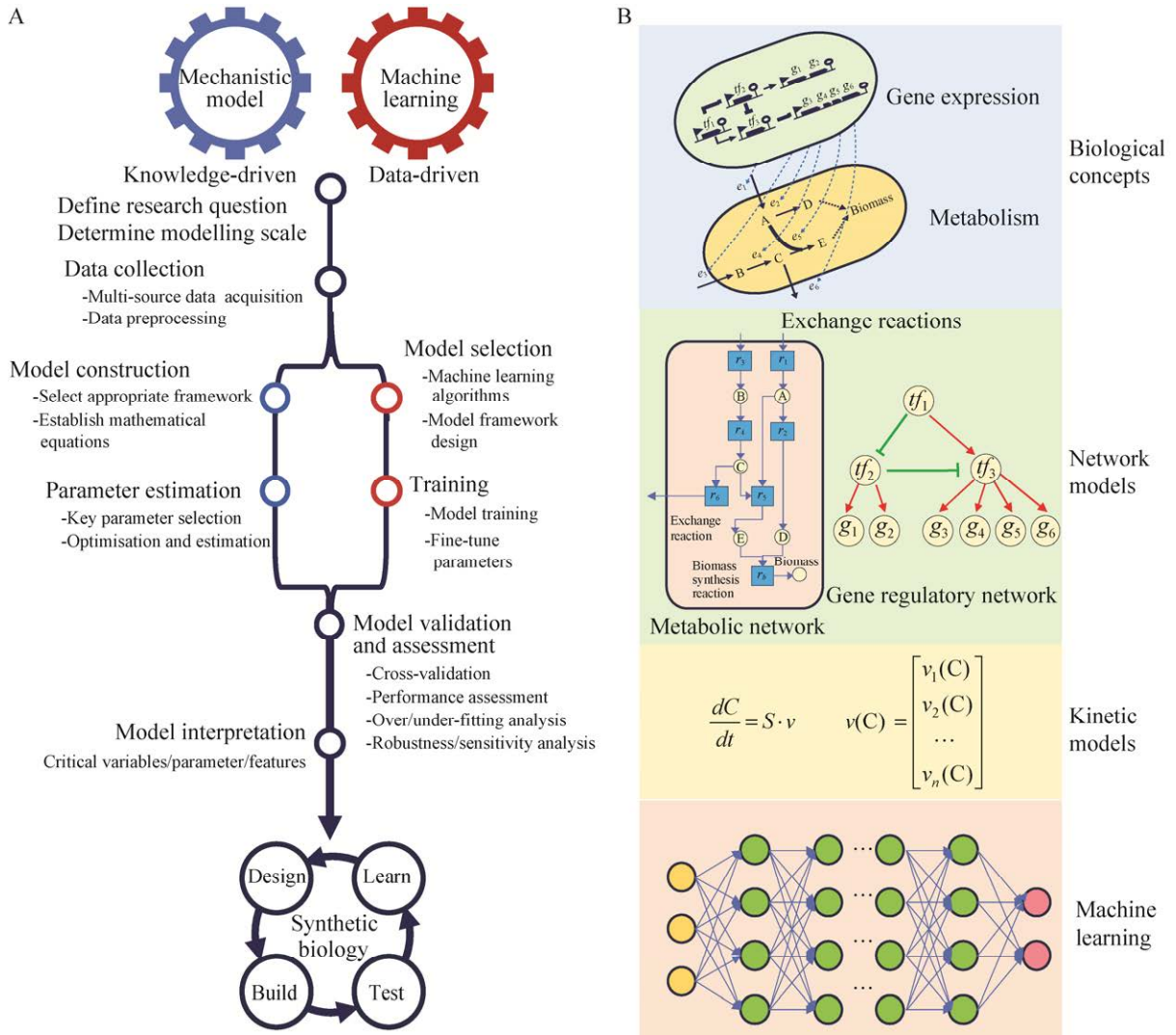


图 1　数学模型构建流程(A)以及主要数学模型类型(B)

Figure 1　The workflow of mathematical modelling of cellular processes (A) and major model types (B).

心性(betweenness centrality)、聚类系数(clustering coefficient)、鲁棒性(robustness)等特征。多数生物网络呈现无尺度(scale-free)特征，节点的度遵守幂律分布[power-law distribution, $P(k) \propto k^{-\lambda}$]，无尺度网络展现高鲁棒性、枢纽节点(hub)高连接度等特点[20]，也发展了基于图的富集算法来研究响应环境扰动的关键基因和酶[21]。Petri 网模型描述离散、异步和并行系统，由库所(place)、变迁(transition)、有向弧(arc)和令牌(token)组成。弧从库所指向变迁，或从变迁指向库所。如果一个变迁的每个输入库所都拥有数量足够的 token，该变迁即可发生。变迁发生后，输入库所的 token 被消耗，输出库所产生新的 token。Petri 网模型常用于研究生化代谢和基因调控的动态过程[22]。布尔网络(Boolean network)是用二进制节点和逻辑规则模拟复杂系统的数学模型，广泛应用于基因调控网络、信号转导网络的分析，通过动态模拟研究网络的吸引子、周期行为、混沌、鲁棒性、可控性等特征[23]。贝叶斯网络(Bayesian network)是一种概率图模

型，使用有向无环图表示随机变量之间的条件依赖关系，并通过条件概率表计算各种概率分布，经常用来推断基因调控关系[24]。

约束模型(constraint-based model)的数理基础涵盖线性代数、优化理论和凸分析。在线性规划(linear programming)中，目标是优化目标函数 c^Tx，同时满足约束 $Ax\leq b$，可行解空间通常是一个凸多面体[25]。单纯形法(simplex algorithm)和内点法(interior point method)是常见的求解算法。非线性规划拓展了函数类型，处理非线性目标函数 $f(x)$和非线性约束 $g_i(x)\leq 0$ 的优化问题，求解依赖于梯度法、拉格朗日乘数法以及KKT条件(Karush-Kuhn-Tucker conditions)。整数规划则增加了离散变量$x_i\in\mathbb{Z}$，通过分支定界法(branch and bound)等组合优化方法求解[26]。约束模型经常用于代谢网络分析和代谢途径设计。

概率模型描述了生物系统的不确定性和随机现象。基本的概率分布如离散分布[二项分布(binomial distribution)、泊松分布(Poisson distribution)]和连续分布[如正态分布(normal distribution)]，通过概率密度函数和累积分布函数来描述变量的取值及概率。贝叶斯定理(Bayes' theorem)是概率模型中的重要工具，根据先验概率 $P(A)$、$P(B)$和新的证据(似然)$P(B|A)$来估计后验概率 $P(A|B)=P(A)P(B|A)/P(B)$。马尔可夫链(Markov chain)用状态转移概率 $P(X_{n+1}=x_{n+1}|X_n=x_n)$描述系统状态的依赖性。高斯混合模型(Gaussian mixture model)通过加权组合 $P(X)=\sum_{k=1}^{K}\pi_k\mathcal{N}(X|\mu_k,\Sigma_k)$处理多模态数据。隐马尔可夫模型(hidden Markov model, HMM)在观察序列的基础上推断隐藏状态。贝叶斯网络(Bayesian network)则用有向无环图表示变量间的条件依赖，联合概率 $P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|\text{Parents}(X_i))$。时间序列模型，如自回归积分移动平均模型(autoregressive integrated moving average model, ARIMA)，通过 $X_t=c+\sum_{i=1}^{p}\phi_iX_{t-i}+\sum_{i=1}^{q}\phi_i\varepsilon_{t-i}+\varepsilon_t$ 描述时间相关数据。这些模型通常通过最大似然估计(maximum likelihood estimation)或贝叶斯推断(Bayesian inference)方法来估算参数，提供了处理随机性和不确定性问题的强大数学框架[27]。

动力学模型通过微分方程、差分方程、随机过程方程等来表示系统中状态变量的变化规律，有常微分方程组(ordinary differential equations, ODEs)、微分-代数方程组(differential algebraic equations, DAEs)、偏微分方程组(partial differential equations, PDEs)以及随机微分方程组(stochastic differential equations, SDEs)等不同类型。求解时使用数值积分方法如欧拉法(Euler method)或龙格-库塔法(Runge-Kutta method)处理ODEs，使用有限差分法或有限元法来求解PDEs。动力学模型可以计算时序演化、预测未来状态、分析系统稳态、揭示混沌行为、研究初始条件敏感性、模拟系统对外部输入或控制策略的响应[28]。其中反应-扩散模型(reaction-diffusion system)描述了化学物质在空间中的扩散和反应过程，由偏微分方程 $\frac{\partial u}{\partial t}=D\nabla^2u+f(u,v)$ 表示，其中 u 和 v 是反应物浓度，D 是扩散系数，$f(u,v)$是反应项[29]。黏弹性模型(viscoelasticity model)描述了材料在外力作用下表现出的既有弹性(瞬时形变)又有黏性(时间依赖性形变)的行为，研究应力$\sigma(t)$和应变$\varepsilon(t)$的关系。在Maxwell模型 $\sigma(t)+\eta\frac{d\sigma(t)}{dt}=E\frac{d\varepsilon(t)}{dt}$ 中，E 是弹性模量，η 是黏性系数。黏弹性模型广泛用于描述细胞膜的机械响应和细胞分裂[30]。

由于生物系统的复杂性，机器学习(machine learning, ML)近年来得到了越来越多的应用[31]。ML可分为有监督学习(supervised learning)、

无监督学习(unsupervised learning)、半监督学习(semi-supervised learning)、强化学习(reinforcement learning)和迁移学习(transfer learning)等多种类型。有监督学习需要标签，可分为离散的分类问题(classification)和连续的回归问题(regression)。传统的 ML 方法有线性回归、分类，支持向量机(support vector machine, SVM)，随机森林(random forest, RF)，*k*-最近邻法(*k*-nearest neighbours, KNN)，神经网络(neural network, NN)。当前方法包括深度学习(deep learning, DL)，例如深度神经网络(deep neural network, DNN)、卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)、图神经网络(graph neural network, GNN)等。

细胞模型通常包含多个动力学参数，这些参数无法完全通过实验测定，在某些情况下，部分参数只能通过数值优化方法估算，通过最小化模型预测与实验数据之间的差异来找到最优参数集(图 1A)。常用方法包括最小二乘法(least squares)和最大似然估计(maximum likelihood estimation)，分别通过最小化平方误差或最大化实验数据在给定参数下出现的概率来确定参数值。参数估计经常面临多局部最优解的问题，因此需采用遗传算法(genetic algorithms)、蒙特卡罗方法(Monte Carlo methods)等全局优化策略搜索。这些方法通过模拟进化或随机采样的过程，探索参数空间中的不同区域，帮助找到最优解[32]。

为了深入理解生物系统，需要计算分析动力学特征，如相平面(phase plane)、吸引子(attractor)、排斥子(repeller)、奇点(singularity)、稳态(steady state)、振荡(oscillation)、随机性(stochasticity)、分岔(bifurcation)、全局稳定性(global stability)、敏感性(sensitivity)和鲁棒性(robustness)等[33]。相平面分析是通过绘制系统状态变量之间的相轨迹图来观察系统的动力学行为。对于一个具有 2 个状态变量 x_1 和 x_2 的系统，其相平面可以表示为 $\frac{dx_1}{dt}=f(x_1,x_2)$ 和 $\frac{dx_2}{dt}=g(x_1,x_2)$ 的解轨迹。通过分析相轨迹，可以识别系统的平衡点和周期解(periodic solution)，从而理解其长期行为。吸引子表征系统长期演化的收敛状态，数学上定义为系统方程 $\frac{d\mathrm{x}}{dt}=\mathrm{f}(\mathrm{x}), x\in\mathbb{R}^n$，当 t 趋于无穷大时解轨迹 x(t)的极限集合。奇点即平衡点 x*，是满足 f(x*)=0 的特殊状态，如鞍点(saddle point)、焦点(focus point)等。稳态分析通过求解 f(x)=0 确定平衡点 x*，其全局稳定性可通过构造李雅普诺夫函数 V(x)并验证 $\frac{dV}{dt}=\nabla V\cdot\mathrm{f}(\mathrm{x})\le 0,\forall \mathrm{x}\ne \mathrm{x}^*$。若条件成立，则 x* 全局渐进稳定。分岔研究系统参数变化引发的动力学突变，其特征为平衡点或周期解的稳定性发生临界转变，导致系统行为模式转变。数学上，分岔可以通过分析系统的雅可比矩阵 $\mathrm{J}=\left.\frac{\partial \mathrm{f}}{\partial \mathrm{x}}\right|_{\mathrm{x}^*}$ 的特征值随参数变化的行为来确定，也可以通过向量场(vector field)来确定吸引子、排斥子和鞍点，并通过分析变量对参数的敏感性，研究分岔现象[33]。敏感度分析和鲁棒性分析则进一步探讨系统在参数变化和外部扰动下的响应能力。敏感度分析通过计算偏导数 $\frac{\partial x}{\partial p}$ 来评估系统对参数 p 变化的响应，而鲁棒性分析通过模拟系统在不同初始条件或参数下的行为来评估系统的稳定性和适应性。这些动力学特征分析方法共同为理解和预测细胞生命过程中的复杂动态行为提供了强有力的数学工具。通过相互结合，它们能够揭示系统在多种条件下的稳态、过渡行为和长期稳定性，为进一步的生物学实验和理论研究提供重要的指导。

模型数值模拟既有 MATLAB、Mathematica 等通用性工具，也有 COBRA toolbox、COPASI 等专用软件[34-35]。准确性评估的方法多种多样，每种方法都有其独特的数学表达和应用场景。常见的分类模型评估指标包括准确率、精确率、召回率、F_1-score 和 AUC-ROC 曲线，而均方误差则主要用于回归任务。还可以使用交叉验证(cross-validation)方法评估模型的预测准确度，特别是在防止过拟合方面效果显著。交叉验证通过将数据集划分为训练集和验证集，进行多次训练和验证模型，从而得到更稳定的预测准确度估计。通用的可视化工具包括 Cytoscape、BioUML 等[36]。模型通用的标准语言包括系统生物学标记语言(systems biology markup language, SBML)、系统生物学图形表示(systems biology graphical notation, SBGN)、模拟实验描述标记语言(simulation experiment description markup language, SED-ML)。模型遵循 FAIR (findable, accessible, interoperable, and reusable)原则[37]。模型数据库包括 BioModels、Physiome Model Repository、JWS online、Metabolic Atlas、BIGG 等。

2 单一细胞生命过程的数学建模

细胞由核酸、蛋白质、脂质、多糖、代谢物等多种组分构成。这些组分的合成、转运、组装和降解以及组分之间的复杂相互作用，形成了 DNA 复制、转录、翻译、大分子装配、膜形成、细胞分裂等一系列细胞过程。传统工程细胞设计通常局限于细胞的生化代谢过程，但研究表明细胞生长、膜形成、转录调控等非代谢过程同样显著影响工程细胞的产量、产率、转化率等生产属性和生长、耐受、稳定等工业属性。因此，为了更准确地实现工程细胞设计，需要构建涵盖生化代谢在内的关键细胞功能的数学模型。

2.1 细胞生长分裂与 DNA 复制

细胞生长和分裂是细胞周期的核心过程。在细菌细胞指数生长模型中，生长速率与细胞数量 N 成正比，即 $\frac{dN}{dt}=\mu N$ (图 2A)。Logistic 模型 $\frac{dN}{dx}=\mu N\left(1-\frac{N}{K}\right)$ 描述对称 S 型生长，早期生长加速，中期达到最大，后期减速趋近环境承载力 K (图 2A)；Gompertz 模型 $\frac{dN}{dx}=\mu N\log\left(\frac{K}{N}\right)$ 描述不对称生长(图 2A)，早期生长速率快，随后速率逐渐减小；扩展的 Richards 模型 $\frac{dN}{dx}=\mu N\left[1-\left(\frac{N}{K}\right)^a\right]$，其中参数 a 控制曲线的形状；进一步考虑延滞期，则得到 Baranyi 模型 $\frac{dN}{dx}=\mu qN\left(1-\frac{N}{K}\right)$，其中 q 表示细胞对环境的适应度，使用 $\frac{dq}{dt}=\mu(1-q)$ 描述延滞期内细胞逐渐适应环境的过程；Huang 模型简化了延滞项 $q=\frac{t}{t+\lambda}$，其中 λ 为延滞期时间常数[38]。Monod 模型 $\mu=\mu_{\max}\frac{[S]}{K_s+[S]}$ 涵盖了底物饱和效应；高底物浓度时，可以使用 Haldane-Andrew 模型 $\mu=\mu_{\max}\frac{[S]}{K_s+[S]+[S]^2/K_i}$ 描述底物抑制，其中 K_i 是抑制常数，浓度高于此值时会开始抑制生长；Andrew 模型用于描述底物中毒性化合物产生的非竞争性抑制 $\mu=\mu_{\max}\frac{[S]}{(K_s+[S])(1+[S]/K_i)}$；Aiba-Edward 模型 $\mu=\mu_{\max}\frac{[S]}{K_s+[S]}e^{\left(-\frac{[P]}{K_i}\right)}$ 用于量化产物抑制效应[39]。反应-扩散模型则用于描述细胞生长的空间效应 $\frac{\partial N}{\partial t}=D_N\nabla^2N+\mu_NN\left(1-\frac{N}{N_{\max}}\right)$[40]。

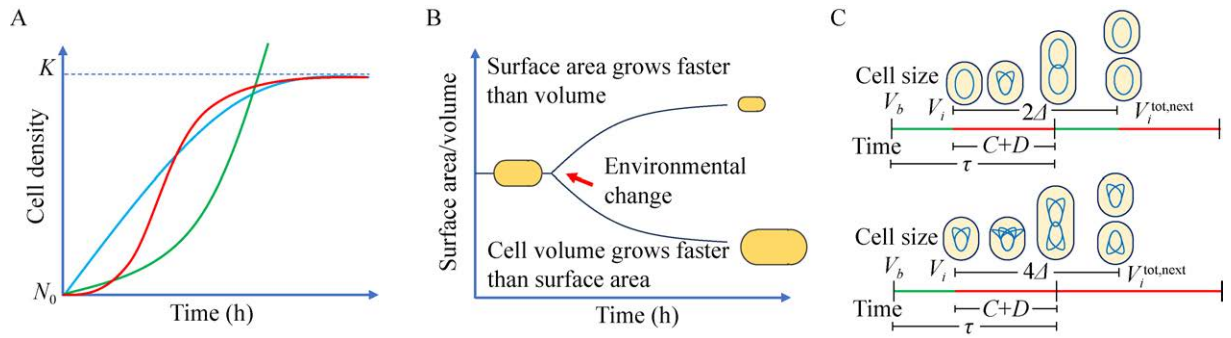


图 2　细胞生长和形态发生模型　A：细胞生长动力学模型。指数生长(绿色)、Monod 模型(蓝色)和 Logistic 模型(红色)，其中 N_0 和 K 分别表示接种量和生长上限。B：细胞表面积和体积生长的动力学模型。C：细胞分裂和 DNA 复制模型。

Figure 2　Mathematical models for cellular growth and morphogenesis. A: Kinetic models for cellular growth, including exponential growth model (green), Monod model (blue) and Logistic model (red). N_0 and K represent inoculum and capacity, respectively. B: Kinetic models for surface area and cell volume. C: Model for cytokinesis and DNA replication.

对于单个细胞来说，细胞生长过程中不断合成细胞壁和细胞膜，细胞的体积 V 和表面积 S 呈指数增长，是比生长速率的函数，即 $\frac{dV}{dt}=\alpha V(t)$, $\frac{dA}{dt}=\beta V(t)$，并且表面积/体积比呈现衰减过程，$\frac{SA}{V}=\frac{\beta}{\alpha}+\frac{c}{V_o}e^{-\alpha t}$，达到稳态时，$SA/V=\beta/\alpha$ 为定值，α 为细胞的体积增长速率，β 为细胞的表面物质合成速率(图 2B)[41]。SA/V 是关联细胞形状和大小的关键变量。保持细胞体积不变，不同形状的细胞则有不同的 SA/V；保持形状不变，体积的增加必然对应 SA/V 的减少。不同环境下会影响 SA/V，例如营养匮乏时，细胞生长较慢，SA/V 增大，细胞变小；反之营养充足时，细胞生长快，体积快速增加，表面积的增长跟不上体积的增长速度，导致 SA/V 逐渐减小，细胞变大；再如使用 β-内酰胺类抗生素抑制细胞壁合成，也会导致 SA/V 减小，细胞变大。

在一个细胞周期 τ 中，DNA 复制、细胞分裂和细胞生长之间相互协调。Adder 模型认为，无论初始大小如何，细胞在每次分裂前增加恒定体积 Δ；Sizer 模型假设细胞达到一个临界体积，分裂才会开始。细菌细胞周期包括 B (细胞分裂后到 DNA 复制启动前的间隔)、C (DNA 复制)、D (DNA 复制完成到细胞分裂完成前的间隔)这 3 个阶段。当营养不足生长缓慢时，$C+D<\tau$；当营养丰富生长速度较快，$C+D>\tau$，形成 $2^{(C+D)/\tau}$个 DNA 复制起始位点[42]，基因组平均含量由 Cooper-Helmstetter 模型决定，即 $\bar{G}=\frac{\tau}{C\ln 2}\left[2^{(C+D)/\tau}-2^{C/\tau}\right]$[43](图 2C)。Schaechter-Maaløe-Kjeldgaard (SMK)生长定律表明细菌细胞大小与生长速率呈指数关系。Donachie 模型结合 Cooper-Hamstetter 模型与 SMK 生长定律，认为细菌在 DNA 复制起始时的细胞体积是相对恒定的，是环境条件的函数。然而，中国科学院深圳先进技术研究院刘陈立团队研究得到个体生长方程，得出 DNA 复制、细胞分裂时间与生长速率的关系 $e^{\lambda(C+D)}=e^{\alpha+\beta\lambda}$，其中 λ 为生长速率，α 与 β 均为常数，同时有 $\frac{1}{C+D}=\frac{\lambda}{\alpha+\beta\lambda}$，细胞质量 $\bar{m}=m_0\lambda(C+D)$[44]；复制启动时细胞质

量为 $m_i=\frac{m_0}{\ln 2}(\alpha+\beta\lambda)e^{-(\alpha+\beta\lambda)}$；细胞周期由 FtsK、FtsZ 等“分裂许可物”浓度控制，当浓度超过阈值时，触发 DNA 复制。对于 DNA 复制过程，可以使用分支链模型模拟 DNA 复制叉的延伸和复制过程。假设 DNA 复制从多个复制起点开始，每个起点以一定速度 v 向两侧延伸。使用偏微分方程描述复制叉位置的动态变化，$\frac{\partial C(x,t)}{\partial x}+\frac{\partial C(x,t)}{\partial t}=D\frac{\partial^2 C(x,t)}{\partial x^2}$。其中，$C(x,t)$表示时间 t 时刻在位置 x 处的复制状态，D 是扩散系数。DNA 突变是在复制过程中由于碱基错配等原因随机发生的，因此采用泊松分布描述基因组中随机突变的发生，即 $P(k)=\frac{\lambda^k e^{-\lambda}}{k^{\lambda}}$，其中，$\lambda$ 是平均突变数。碱基突变还可以通过 Markov chain 模型描述，解析突变在 DNA 序列中累积的过程，特别是碱基置换的过程。Kimura 的中性进化理论模型用于描述基因频率在自然选择和突变之间的平衡。突变的影响被认为是中性的，即大多数突变不影响生物的适应性。突变率 μ 和选择系数 s 之间的关系决定了等位基因频率的变化，$\frac{dp}{dt}=\mu(1-p)+\mu p$，其中，$p(t)$是 t 时刻某种等位基因的频率。用于研究 DNA 复制的精确性和突变如何在基因组中积累，有助于推动基因组稳定性、突变率和进化选择压力的相关研究。

细胞分裂方式多种多样，包括二分裂(binary fission)、减数分裂(meiosis)、芽殖(budding)、孢子形成(spore formation)等。二分裂是大多数细菌的分裂方式，其过程包括 DNA 复制、细胞生长、隔膜(septum)形成、细胞分裂及分离。横隔形成是细菌二分裂的关键步骤，细胞壁和细胞膜在中间部位逐渐向内生长，最终将细胞分成 2 个子细胞。描述横隔形成的数学模型，通常考虑细胞分裂过程中的动力学、几何形状变化以及生物化学反应。细胞形态发生常使用生物力学模型[45]。有研究使用水平集(level-set method)方法和黏弹性模型(viscoelastic model)成功模拟了原生生物盘基网柄菌细胞二分裂的过程；结果表明，细胞两极的牵引力或由肌球蛋白(myosin II)产生的收缩力可以启动分裂沟(furrow)的形成；肌动蛋白皮层张力(actin cortical tension)和细胞表面曲率产生的被动应力有助于分裂沟的最终收缩[46]。FtsZ 是一种类微管蛋白，在细菌分裂时形成 Z 环，Z 环是其他分裂蛋白(如 FtsA、FtsQ、FtsL、FtsB、FtsW、FtsI 和 FtsN)的支架，也可引起膜收缩。FtsZ 主要以单体形式存在，聚合形成原丝(protofilaments)，原丝进一步聚集成细胞分裂的 Z 环。FtsZ 聚合包括初始核化和链式延伸，聚合长度达到一定临界值，FtsZ 原丝形成环状结构(cycling)。使用 GTP 水解驱动 FtsZ 亚基的解聚和再聚合，这通常用一阶动力学方程表示，即 $\frac{d[P]}{dt}=k_{\mathrm{on}}[Z]-k_{\mathrm{off}}[P]$，其中[$P$]是 Z 环中的聚合态 FtsZ 浓度，$k_{\mathrm{on}}$ 和 k_{off} 分别是聚合和解聚速率常数。对于分裂蛋白 FtsX 的合成，有 $\frac{d[FtsX]}{dt}=k_{\mathrm{X}}[FtsX]-k_{\mathrm{d}}[FtsX]$；其中[$FtsX$]是分裂蛋白 X (FtsA, FtsK)的浓度，k_{X} 和 k_{d} 分别是其招募和解离的速率常数[47]。分裂复合体的主要功能是生成机械力以推动细胞膜向内收缩，最终完成细胞的物理分裂。这一机械力的生成主要由 FtsZ 蛋白的聚合-解聚动力学驱动，并通过分裂复合体内的其他蛋白调控。

$$\mu\frac{d^2R(t)}{dt^2}=-\gamma\frac{dR(t)}{dt}+k_{\mathrm{c}}N(t)\frac{dR(t)}{dt}+\sigma(t)$$

$$\sigma(t)+\tau\frac{d\sigma(t)}{dt}=\eta\frac{dR(t)}{dt}$$

$$S(t)=\alpha[FtsI]\frac{dR(t)}{dt}$$

其中：$R(t)$是细胞膜的半径；μ 是细胞膜的有效

质量密度；γ 是黏性阻尼系数，表示膜在运动过程中的阻力；$F(t)$是由分裂复合体生成的有效收缩力；$\sigma(t)$是膜上的应力；τ 是松弛时间，表征材料的弹性和黏性特性；η 是黏性系数；$S(t)$是细胞壁材料的合成速率；[*FtsI*]是合成酶的浓度；α 是与合成速率相关的比例系数。

真核细胞周期的调控网络模拟采用微分方程、随机方程和布尔网络的方法。通过构建芽殖酵母的细胞分裂检查点信号转导布尔网络模型，并通过动力学模拟，成功找到了全局吸引子，分别对应着分裂完成、分裂中等若干细胞稳态[48]。研究人员通过构建 ODEs 研究了酿酒酵母(*Saccharomyces cerevisiae*)中 APC/CAma1 和 Ndt80 等蛋白协同互作介导减数分裂前期向中期转变的分子机制[49]。革兰氏阳性菌枯草芽孢杆菌(*Bacillus subtilis*)的孢子形成受 Spo0A 磷酸化级联系统调控；研究通过建立 ODEs 模型结合单细胞实验发现，细胞生长速率下降通过双重机制调控该过程：一方面，减缓的稀释效应(即蛋白质合成速率不减，因细胞体积增长放缓导致蛋白质浓度被动升高)促使磷酸化级联关键蛋白累积；另一方面，延长的染色体复制周期改变了基因剂量平衡时间窗口；这种生长速率依赖的调控使 Spo0A 活性达到阈值时触发孢子形成，而无需直接感知特定代谢信号；该机制揭示了细菌通过物理生长参数整合多维度环境信息的独特策略，为理解微生物应激分化提供了新视角[50]。

2.2 转录调控

细胞需要精准动态地调控基因表达。顺式作用元件(如启动子、增强子)和反式作用元件(如 σ 因子、转录因子和 sRNA)与靶基因互作，形成了复杂的基因调控网络(gene regulatory network, GRN)。研究人员利用网络模型来映射基因调控的复杂路径，或通过动力学模型来模拟基因表达随时间变化的动态过程。从基因调控数据库如 GRAND、RegulonDB、DBTBS、Abasy Atlas、PRODPROC、coryneRegNet 等出发，结合高通量组学如 RNA 测序(RNA-seq)、染色质免疫沉淀测序(chromatin immunoprecipitation sequencing, ChIP-seq)和 DNA-pulldown 等实验，采用机制模型、信息论方法和机器学习技术来构建 GRN (图 3)。

数学模型方法包括布尔网络、贝叶斯网络和微分方程等。在概率布尔网络(probabilistic Boolean network, PBN)中，每个基因 i 由 m_i 个布尔函数调控，这些布尔函数以概率 p_{ij} 选择，总和为 1。贝叶斯网络计算条件概率推断基因调控关系。动态贝叶斯网络(dynamic Bayesian network, DBN)拓展了处理时序数据的能力。DBN 将网络分为初始网络和转移网络，分别结构寻优，其联合概率为

$$P\left(X_1^{(1)},X_2^{(1)},\cdots,X_n^{(1)},\cdots,X_1^{(T)},X_2^{(T)},\cdots,X_n^{(T)}\right)=\prod_{t=1}^{T}\prod_{i=1}^{n}P\left(X_i^{(t)}\mid X_1^{(t-1)},X_1^{(t-1)},\cdots,Pa\left(X_i^{(t)}\right)_1^{(t)}\right)$$

ODEs 描述基因调控网络中的复杂动态行为，基因 i 的表达为 $\frac{dx_i}{dt}=f\left(x_1,x_2,\cdots,x_n\right)-d_ix_i$。其中 d_i 为降解速率，f 为调控函数，常见形式为 $f(x_i)=\frac{\alpha x_i^n}{K^n+x_i^n}$，$\alpha$ 为最大表达速率，K 是半饱和系数，n 是 Hill 系数。SDEs 则考虑了基因表达的随机波动 $dx_i=f(x_1,x_2,\cdots,x_n)dt+g(x_1,x_2,\cdots,x_n)dW_t$，其中 $f(x)$漂移向量表示基因调控的确定性动力学部分，$g(x)$扩散矩阵表示随机影响的强度，dWt 为 Wiener 过程，表示随机表达。如果进一步考虑空间扩散的影响，将基因表达 x_i 写成时间 t 和空间位置 r 的函数，用 PDEs 模拟 $\frac{\partial x_i}{\partial t}=f\left(x_1,x_2,\cdots,x_n,\nabla x_1,\nabla x_2,\cdots,\nabla x_n,\mathrm{r}\right)-d_ix_i$，其中$\nabla x_i$为梯度算符，PDEs 求解较为困难，并需要合理的边界条件。

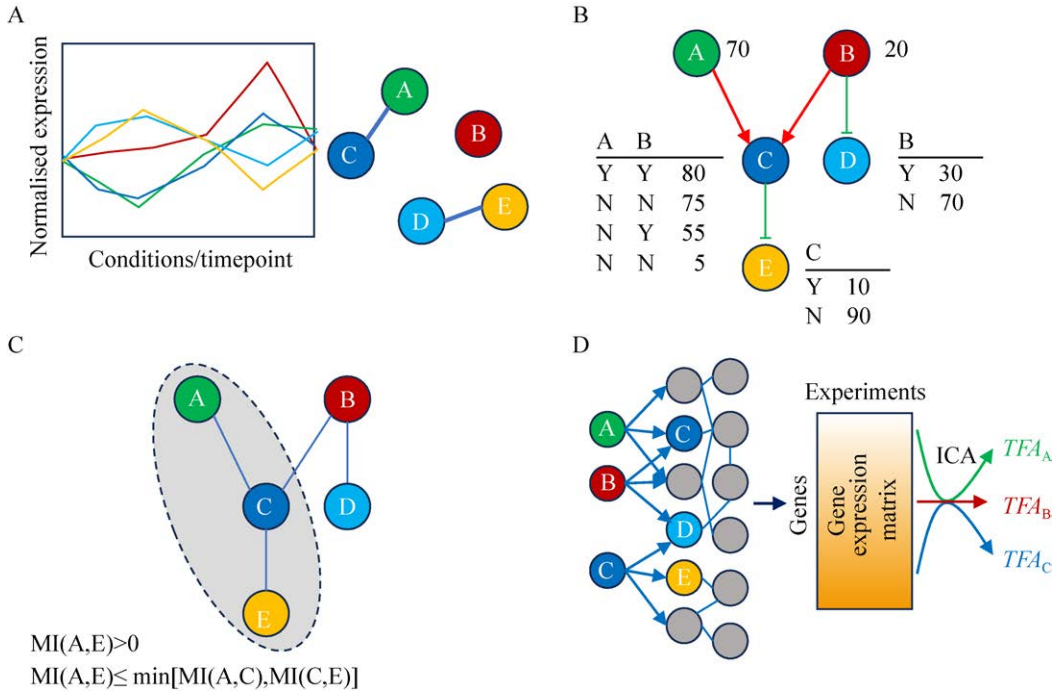


图 3　转录调控网络模型　A：根据基因表达谱相关度将基因聚类，其中基因 A 与 C、D 与 E 聚在一起。B：根据基因表达概率构建贝叶斯调控网络，其中 C、D、E 的条件概率描述了对其调控基因的依赖关系。C：根据基因表达谱计算互信息熵并构建调控网络，其中基因 A 对 E 的互信息小于或等于 A 直接影响 C 和 C 直接影响 E 的最小信息，表示基因 A 可能通过 C 来间接影响 E，因此可在调控网络中去除。D：基于 ICA 推断的转录因子-靶基因网络，其中 A、B 为转录调控因子，D、E 为靶基因，C 既是转录调控因子也是靶基因。

Figure 3　Transcriptional regulatory network model. A: Gene clustering based on correlations between gene expression profiles, grouping genes A and C, as well as D and E. B: Bayesian network representation of the regulation network, with conditional probabilities for genes C, D and E reflecting dependencies on A and B. C: Mutual information network where data processing inequality is applied to eliminate indirect interaction between A and E. D: Transcriptional factor (TF)-target gene (TG) regulatory network inferred using intendent component analysis (ICA), identifying A and B as TFs, D and E as TGs, and C as a dual-function gene.

基于信息方法构建 GRN，核心在于计算基因表达之间的相关性或依赖性。常用的标准包括 Pearson 相关、Spearman 相关、Kendall's Tau 相关、互信息(mutual information, MI)、偏相关和距离相关等(图 3A–3C)。使用相关系数构建基因共表达网络，例如加权共表达方法(weighted gene co-expression network analysis, WGCNA)。ANOVerence 是通过方差分析(ANOVA)计算 η^2 量化基因调控关系；η^2 表示一个基因对另一个基因表达方差的解释比例，通过组间平方和与总平方和的比值得出；基因调控关系的强度通过 η^2 的大小和显著性检验来判断，η^2 越大，调控作用越强[51]。

互信息度量 2 个基因表达之间的关联，$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$，其中 X 和 Y 表示 2 基因的表达水平，$p(x,y)$为联合分布。可处理非线性相关，但无法排除间接相关

的干扰，从而造成高假阳性率。通过不等关系(algorithm for the reconstruction of accurate cellular networks, ARACNe)[52]、生物背景信息(context likelihood of relatedness, CLR)[53]、最小化冗余信息(minimum redundancy networks, mrNET)[54]、偏相关、弹性网络等方法来鉴别和去除间接作用。最大信息系数(maximum information coefficient, MIC)方法通过衡量 X 和 Y 这 2 个基因表达水平之间的所有可能关系，找到能够最大化互信息的网格划分，并以此度量之间的依赖性[55] $\mathrm{MIC}(X,Y)=\max\left(\frac{I(X;Y)}{\log\left(\min\left(|X|,|Y|\right)\right)}\right)$，其中，$|X|$和$|Y|$分别是 X 和 Y 的网格数；MIC 越大，2 个基因间依赖关系越强，越可能存在调控关系。条件互信息(conditional mutual information, CMI)用于衡量 2 个基因表达在第 3 个基因表达下的相互依赖性 $I(X;Y|Z)=\sum_{z\in Z}p(x)\sum_{x\in X}\sum_{y\in Y}p(x,y|z)\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)$；可以区别直接和间接互作，但假阴性较高[56]。部分互信息(part mutual information, PMI)是为了解决 CMI 低估问题而提出的新概念[57]。对于基因 X、Y 直接依赖且强关联第三方 Z，CMI 趋于 0，PMI 则能正确量化 $\mathrm{PMI}(X;Y|Z)=\sum_{x,y,z}p(x,y,z)\log\left(\frac{p(x,y|z)}{p^*(x|z)p^*(y|z)}\right)$。其中 $p(x,y,z)$是联合概率分布，$p^*(x|z)$和 $p^*(y|z)$是考虑了 X 和 Y 之间依赖性的修正条件分布。转移熵(transfer entropy)是 CMI 的特例，旨在量化从一个时间序列到另一个时间序列的信息流动。对于 2 个时间序列 X_t 和 Y_t，定义为 $T_{X\to Y}=H\left(Y_{t+1}|Y_t\right)-H\left(Y_{t+1}|Y_t,X_t\right)$。MICRAT (maximum information coefficient regression and tree-based analysis)是一种结合了 MIC 和决策树的混合方法，通过计算基因对之间的 MIC 值、回归分析以及决策树建模来构建基因调控网络。其数学原理涉及利用 MIC 捕捉基因之间的非线性关系，并通过回归和决策树分析进一步确定调控方向和强度，从而构建一个精确且可解释的 GRN。CMI2NI 通过计算互信息和条件互信息，识别基因对之间的直接调控关系，并过滤掉间接依赖关系，最终构建出 GRN。其数学原理基于互信息和条件互信息的比较，能够有效识别复杂网络中的真实调控关系。但总体上，互信息模型可以推断基因表达间的非线性依赖关系，但不能明确基因之间的调控方向，不能直接判断出 2 个基因中的调控者与被调控者，需要结合其他方法确定调控关系。

机器学习整体上包含有监督和无监督这 2 类算法。随机森林是一种有监督方法，已开发的 GENIE3 、 RF-IM 、 SIRENE 、 iRafNet 、 NetBenchmark、GRADIS 和 GRNBoost 等算法都用于 GRN 推断。这类方法利用特征重要性评估，识别基因之间的潜在调控关系，并逐步优化网络结构，适用于复杂和高维基因表达数据的分析[58]。回归也常用于 GRN 推断，常用方法包括 LASSO 回归、岭回归、Elastic Net、PLS 回归、多元线性回归和非线性回归等。这些方法通过量化基因之间的调控关系，帮助研究人员从高维基因表达数据中推断出 GRN 结构。成分分析如独立成分分析(independent component analysis, ICA) (图 3D)和网络成分分析(network component analysis, NCA)，是用于数据降维和特征提取的无监督学习方法。其主要目标是从一组观察到的数据中分离出彼此独立的潜在信号或成分，ICA 属于特征提取和信号分离的技术，而 NCA 则结合了成分分析与网络推断的特性。将转录组矩阵 X 分解为转录因子活性 A 和转录调控强度 C 的乘积，即 X=A·C。CNN 通过

卷积核提取 DNA 结合基序等局部序列特征，结合池化层降维，常用于识别转录因子结合位点。RNN 方法通过处理基因表达时间序列数据，捕捉动态调控模式，并用于构建随时间变化的 GRN。GNN 方法利用图结构数据，能够整合转录组学、空间基因组学等多组学数据，构建高精度 GRN，尤其适用于处理单细胞数据和三维基因组数据。而生成模型和强化学习方法则提供了创新的网络结构探索和优化手段。这些方法使得在处理基因表达数据和推测 GRN 时，具有更高的准确性和灵活性。

基因调控过程的动力学特征分析涉及描述基因表达随时间变化的过程，并理解 GRN 如何在不同条件下影响这些变化。这种分析通常依赖于数学模型来捕捉基因调控的动态行为。布尔网络通过使用二值状态和布尔函数，描述基因调控过程中的动力学特征。关键的数学原理包括状态空间的吸引子分析、网络拓扑对动力学的影响、系统的鲁棒性与相变等。通过这些分析，深入理解基因调控网络在不同条件下的全局行为和稳定性，能够帮助揭示生物系统的复杂动力学特性。使用微分方程来描述基因表达随时间的变化。通过稳态分析 $\frac{dx_i}{dt}=0$，帮助理解基因调控网络是否会趋于某种平衡状态，或者在某些情况下可能出现多重稳态或振荡行为。考虑稳态解附近的小扰动 δx_i，系统的线性化方程可以表示为 $\frac{d\delta x_i}{dt}=J_{ij}\cdot\delta x_j$。其中 J 是系统雅可比矩阵(Jacobian matrix)，通过求解其特征值，可以判断扰动是否衰减，从而判定稳态的稳定性。在非线性基因调控网络中，系统可能会展示出复杂的动力学行为，如多稳态、振荡甚至混沌。SDEs 模型中引入 Wiener 过程描述基因表达的随机波动，如噪声诱导的稳态转变或随机振荡。

2.3 信号转导

细胞感知环境变化并通过信号转导调控细胞行为。胞内的双组分系统、G 蛋白偶联受体系统、蛋白激酶系统、磷脂酰肌醇信号通路、环核苷酸系统共同构成了微生物群体感应、趋化运动、细胞分裂、生物膜(biofilm)形成、感受态形成、产孢过程、胁迫响应、代谢调节、底物利用等核心功能网络。信号转导过程复杂、动态多变、专一性高，深入理解信号转导网络对于工程细胞定制设计具有重要意义。

逻辑模型使用布尔逻辑或其他离散逻辑描述信号通路中分子的开关状态，不涉及精准的分子浓度，而是关注信号是否在路径上通过或阻断。例如，在布尔网络中，一个节点的状态可以由其输入节点的状态决定，$x_i(t+1)=f_i(x_1(t),x_2(t),\cdots,x_n(t))$。双组分系统响应环境变化的数学模型主要基于生物化学反应的动力学方程，通常使用常微分方程来描述传感器激酶的磷酸化、磷酸基团的转移、响应调节蛋白的激活及其与 DNA 的结合过程。

$$\frac{d[S^*]}{dx}=k_{ph}[S]-k_{tr}[S^*][R]$$

$$\frac{d[R^*]}{dx}=k_{tr}[S^*][R]-k_{deph}[R^*]$$

其中：[*S*]为 sensor kinase 的浓度，response regulator 为[*R*]，Sensor kinase 自磷酸化的速率为 k_{ph}，磷酸转移速率为 k_{tr}，去磷酸化速率为 k_{deph}。Sensor kinase 的激活函数可以采用 Michaelis-Menten 方程来模拟饱和动力学。

基于 cAMP 的分解代谢物阻遏(carbon catabolite repression, CCR)通过调节细胞内 cAMP 水平来控制基因表达，从而影响代谢物的利用。当存在优先碳源如葡萄糖时，cAMP 水平下降，导致 cAMP 受体蛋白(cAMP-receptor protein, CRP)的活性降低，抑制次要碳源相关基因的表达。在大肠杆菌(*Escherichia coli*)中，葡

葡萄糖通过磷酸转移酶系统(phosphotransferase)抑制腺苷酸环化酶(adenylyl cyclase)活性，降低cAMP的合成，减少CRP-cAMP复合物的形成，进而抑制次要碳源基因的表达。而在枯草芽孢杆菌中，代谢物阻遏由CcpA介导，CcpA与磷酸化的HPr结合后抑制次要碳源基因的表达，葡萄糖的存在促进了这一抑制作用。尽管两者机制类似，但调控蛋白和信号传导路径有所不同。因此模型中包含了葡萄糖摄取与磷酸转移酶系统、cAMP的合成与降解、CRP-cAMP复合物的形成与解离、基因表达的调控等几个部分。约束模型及资源分配(resource allocation)原则可以模拟一定底物组合(如葡萄糖和乙酸，葡萄糖和乳酸)的顺序利用和二次生长(diauxic growth)现象。基于基因调控机制的微分代谢方程(differential algebraic models)则可以用来模拟更广泛的代谢物阻遏抑制以及诱导物排出效应(inducer exclusion)[59]。

在不利环境条件下，枯草芽孢杆菌在芽孢形成和自然感受态之间决策，这一复杂过程涉及双组分系统、转录因子、磷酸化激酶等多种调控蛋白。采用质量作用定律(mass action)和Hill方程形式构建微分动力学方程组，模拟感受态(由群体感应控制的随机开关)和孢子形成(由压力感应调节的定时器)模块以及这2个模块间通过Rap评估系统和AbrB-Rok决策模块的交互，分析细胞响应内外部信号做出形成芽孢还是得到获取外源DNA的决策机制；此外，模型还揭示了通过特定抑制机制产生的Spo0A*和AbrB浓度波动，这可能是细菌变异性和环境适应性的关键来源；这一研究框架为未来探索细菌中复杂信号网络及其在生态和进化中的作用提供了基础[60]。

2.4 群体行为

群感效应(quorum sensing)参与生物膜形成、细胞分化、胞外多糖合成、细胞趋化、运动、抗生素形成等多种生物行为。采用微分方程描述酰基-高丝氨酸内酯(acyl-L-homoserine lactone, AHL)的产生和降解如何影响细菌行为。反应-扩散方程经常用来描述细胞群体的动态变化。

$$\frac{\partial A}{\partial t}=D_A\nabla^2 A+k_N N-k_A A$$

$$\frac{\partial N}{\partial t}=D_N\nabla^2 N+r_N N\left(1-\frac{N}{N_{\max}}\right)f(A)$$

其中：k_N和r_N分别是信号分子合成速率常数以及细胞生长速率，k_A是信号分子降解速率常数；$N_{\max}$是最大生长量；D_A和D_N是信号分子和细胞扩散系数；f(A)表示细胞生长速率r_N对信号分子的响应函数。2001年，Nilsson等[61]建立了一个数学模型，模拟了细胞内和细胞膜上信号分子AHL的浓度随时间的变化；他们发现，AHL浓度变化与群体生长速率、AHL扩散速率和自诱导速率相关；在细菌生长的初始阶段，AHL扩散到细胞外速率低，细胞生长慢，自诱导快，AHL浓度迅速增加，随后达到第1个稳定期；接着细胞生长速率逐渐提高，AHL扩散到细胞外速率快，细胞生长快，AHL浓度逐渐降低，最后进入第2个稳定期。另外也可以使用机器学习方法鉴定共同信号分子介导的细菌间相互作用，从而建立人类肠道微生物菌群的群感效应通讯网络[62]。

描述趋化性的数学方程主要基于偏微分方程，如Keller-Segel模型，具有与群感响应类似的结构。

$$\frac{\partial C}{\partial t}=D_A\nabla^2 C+k_N N-k_A C$$

$$\frac{\partial N}{\partial t}=D_N\nabla^2 N-\nabla\cdot\left(N\chi(C)\nabla C\right)$$

其中：C为化学物质浓度，$\chi(C)$是趋化系数，表示细胞对化学物质梯度的敏感性。扩散项

$D_N\nabla^2N$ 表示细胞随机运动导致的扩散，细胞倾向于从高密度区域向低密度区域扩散，趋化项 $-\nabla\cdot(N\chi(C)\nabla C)$表示细胞在化学物质梯度下的定向移动。化学趋化性使得细胞在化学物质浓度梯度的作用下向(或远离)高浓度区域移动。趋化系数 $\chi(C)$决定了响应的强度和方向。行波解常用来构造反应-扩散方程的解，可将偏微分方程转化为常微分方程组，然后利用相平面技术和打靶法来证明行波解的存在。Agent-based model (ABM)在模拟群感和趋化效应时，通过定义个体(细菌或细胞)的状态和行为规则，结合个体间的局部交互和环境作用，来研究系统整体行为的涌现。这种方法能捕捉复杂的非线性动态，是理解和预测群感效应在复杂生物系统中表现的重要工具。结合 ABM 和 DL 预测了 *E. coli* 和荧光假单胞菌(*Pseudomonas fluorescence*)这 2 种细菌在不同培养基上的相互作用以及随空间环境变化的分布[63]。有限元方法建立了细菌包膜附着模型，分析表明，与附着在厚复合材料相比，绿脓杆菌(*Pseudomonas aeruginosa*)细胞附着在薄复合材料上时，细胞包膜的机械应变和应力增大，从而产生更多的环二鸟苷酸(cyclic diguanylic acid, c-di-GMP)，影响运动和生物膜形成[64]。

2.5 生化代谢

工程细胞的生化代谢网络是生物制造的核心。精准的代谢模型是设计构建高效细胞代谢网络的基础。当前常用的代谢模型包括约束模型和动力学模型。1993 年，Varma 等[65]首创了基于约束的化学计量模型(constraint-based stoichiometric model, CBM)方法(图 4)。2000 年，Schilling 与 Palsson 构建了嗜血流感菌(*Haemophilus influenzae*)模型，这是首个基因组尺度代谢模型(genome-scale metabolic model, GSMM or GEM)[66]。至今，文献中已报道了超过 6 000 个 GSMM，横跨真核、古菌、细菌三域[67]；BiGG、Metabolic Atlas、BioModels 等模型数据库均有收录。另外微生物组研究领域也开发了一系列基因组尺度代谢模型库，包括 AGORA、APOLLO 等；其中，APOLLO 包含了 247 092 个微生物 GSMM[68]。一些模式生物如大肠杆菌、枯草芽孢杆菌、酿酒酵母以及人等已有多个模型。最大的 GSMM 为人类细胞模型 Recon3D，包含 13 543 个代谢反应、4 140 个代谢物、3 288 个开放读码框(open reading frame, ORF)[69]。在此基础上，研究人员还开发了特定人类细胞系和组织的模型[70]。

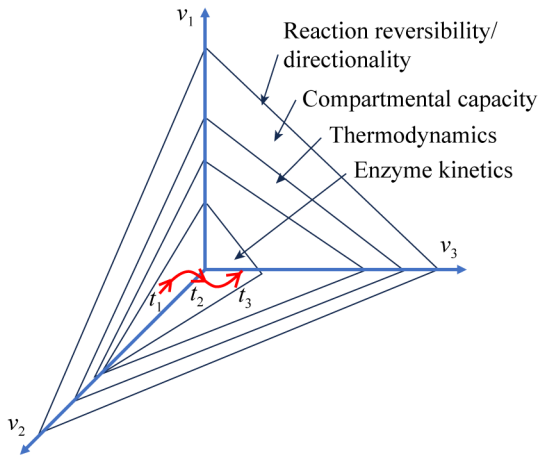


图 4 基于多层约束的代谢模型解空间 红色曲线表示基于常微分方程组的代谢模型的解。

Figure 4 Biochemical model solution space with multi-layer constrains. Red line represents the solution of ODEs.

GSMM 集合了细胞内所有的生化代谢反应，使用以线性规划为数学基础的通量平衡分析(flux balance analysis, FBA)，在物质守恒($\mathbf{S}\cdot\mathbf{v}=0$)、反应可逆性、给定底物利用速率($v_j<b_j$)等约束条件下，计算使目标函数取得最值，如生长速率最大(max $v_{biomass}$)的决策变量代谢流分布。计算相对简单，所需资源少。构建一个高质量 GSMM 一般从基因组注释信息出发，参照 KEGG、BioCyc、Brenda 等生化代谢数据库，

建立基因-酶-反应的对应关系、代谢物和反应列表，最终写成基于 SBML 标准的模型文件。GSMM 构建的标准化流程包括多达 96 个步骤[71]。可以使用 COBRA Toolbox、cobrapy、sybil、GECKO 等工具读取模型开展计算分析。同时，结合如 Biolog 实验等生长表型和突变体文库信息不断修正模型，最终达到对生长代谢表型的精准预测。目前也发展了多种 GSMM 自动构建工具，包括 carveME、KBase、AuReMe、gapseq 等。模型质控方面也发展了 Memote 等一系列标准和工具[72]。

最大化生长作为目标函数预测代谢流，是基于生长快的细菌易取得进化优势的强假设，这可能会与真实情况有所偏差，例如细胞生长较慢，这一假设可能并不适用，因此经常采用能量产生最大化作为目标函数[73]。使用通量可变分析(flux variability analysis, FVA)方法可以计算代谢流可变范围，使用随机方法在解空间随机抽样，可以无偏刻画解析代谢特征，这样的方法有马尔可夫链蒙特卡洛(Markov chain Monte Carlo, MCMC)采样、artificial centering hit-and-run (AHCR)等。简约通量平衡分析(parsimony FBA)等方法通过优化目标函数并最小化总代谢通量或蛋白质总量，消除冗余途径，从而得到全局代谢网络的最优通量分布。可利用表型相平面(phenotype phase plane)分析从完全好氧、微好氧和厌氧情况下细胞的代谢表型。可以通过模拟单基因敲除(single gene deletion)分析基因必需性。可利用最小代谢调整(minimisation of metabolic adjustment, MOMA)、最小调控开关(regulatory on/off minimization, ROOM)等方法计算基因扰动下代谢流的重分布。可以通过动态通量平衡分析(dynamic flux balance analysis, dFBA)方法计算生长动态过程。

一般情况下，FBA 的求解空间较大，引入其他约束可以缩小解空间，包括：(1) 热力学约束。例如采用吉布斯自由能变(Δ_rG)限制反应方向的 TMFA (thermodynamics-based flux analysis)、能量产生和耗散平衡的 EBA (energy balance analysis)、能量耗散优化的 TOS (thermodynamic optimum searching)算法等[74-76]。(2) 酶总量约束。例如酶总量不超过细胞干重的 50%[77]，细胞器(如线粒体)酶的总体积不超过该细胞器总容积，细胞膜上膜蛋白所占总面积不超过膜的总面积(从发酵和呼吸过程对细胞膜资源的竞争进行约束建模，尝试开发基于膜资源经济型原则的代谢模型)解读溢流代谢，细胞内所有蛋白的总量不超过细胞的总容积(FBAwMC 算法)[78]。(3) 酶动力学约束。单位酶的催化速率不超过其 k_{cat}，且受到酶饱和程度的约束，其中，MOMENT (metabolic modeling with enzyme kinetics)、ECMpy、GECKO 等方法通过引入代谢酶动力学约束($v \le v_{max}=k_{cat}E$)以提高代谢模型预测准确度[79]。(4) 酶量约束。定量蛋白质组学数据以及蛋白复合体数据作为约束酶量的上限，如 GECKO。(5) 转录调控约束。如整合了基于布尔代数的调控信息的 rFBA (regulatory FBA)、整合了转录调控及信号转导的 iFBA (integrated flux balance analysis)[80]、整合了调控概率的 PROM (probabilistic regulation of metabolism)。(6) 多组学约束。如使用转录组学数据作为约束的 MADE、INIT 使高表达量的酶催化反应的代谢通量限定必须大于某个阈值、iMAT 将转录组学和蛋白质组学数据与基因组规模代谢网络模型集成在一起，以预测酶的代谢通量，和 Repitide，使用代谢组学数据作为约束的 uFBA，使用脂质组学作为约束，同时使用转录组学和代谢组学作为约束 GIMME 等。(7) 整合动力学模块的 dFBA 可以计算细胞代谢的动态变化。

GSMM 在工程细胞设计构建中展现出卓越

的预测指导能力。基于优化算法的设计方法已形成体系化发展，主要包括采用双层规划的OptKnock、OptReg、OptORF 等[81-83]；基于混合整数线性规划的 OptDesign、OptRAM、OptStrain、NIHBA 等[84-87]；对比野生型与工程菌种代谢差异从而发现改造靶点的 OptForce、RoboKoD 等[88-89]；以及应用启发式优化算法与系统搜索策略的 OptGene、FastKnock 等[90-91]。使用模型指导工程细胞设计构建，显著提高了包括有机酸、氨基酸、脂肪酸、蛋白等一系列产物的产量。例如，基于大肠杆菌 iJO1366 模型的酪氨酸氨基转移酶(*tyrB*)和天冬氨酸转氨酶(*aspC*)基因双敲除策略，使 D-phenyllactate 产量从 0.55 g/L 提升至 1.62 g/L[92]。利用酿酒酵母 Yeast8 模型筛选出 84 个血红素合成相关靶点，经实验验证与组合优化后，通过 40 个遗传改造使产量提升 40 倍[93]。应用酵母 pcSecYeast 模型预测 α-淀粉酶高产靶点时，116 个候选靶点中有 18 个经实验验证，其中 14 个显著提升酶产量，阳性率达 77.8%[94]。这些成果凸显了 GSMM 指导代谢工程的多层次应用价值：从单基因敲除到多靶点协同调控，从初级代谢物到复杂重组蛋白，为智能工程细胞设计提供了理论框架和技术范式。

GSMM 在基础生物学和合成生物学中得到了广泛应用。然而，其基础是代谢稳态假设，仅能模拟细胞代谢的一个断面。代谢动力学模型(metabolic kinetic model)可以计算代谢物浓度以及代谢酶活性随时间的连续变化。一般动力学模型由若干常微分方程和代数方程构成，包含初始条件、状态变量和速率方程等几个关键组成要素，采用数值积分方法求解。模型多用于计算中心代谢的动态调控、底物利用过程。可以精确模拟代谢物浓度随时间的变化曲线。动力学模型也可用于代谢控制分析(metabolic control analysis, MCA)，计算弹性系数、控制系数等关键变量[95]。之前的动力学模型多对中心碳代谢进行模拟[96-97]。目前最大的代谢动力学方程是大肠杆菌的 k-ecoli457 模型，包含 457 个代谢方程、337 个代谢物以及 295 个底物水平调控关系[98]。代谢动力学方程的关键是参数估计，可以使用遗传算法(genetic algorithm)等优化算法，利用大量定量代谢组学、通量组学数据训练得到。代谢动力学方程也可以和 FBA 方法整合，以动力学模块驱动，通过类数值积分算法模拟细胞全局代谢的动力学过程，例如前文提到的 rFBA、iFBA 等方法。模型可用来解析 CCR 效应的机制[59]。基于动力学模型的代谢工程优化取得新突破。最新开发的 NOMAD 框架通过整合代谢组学、热力学及发酵动力学数据，构建了包含 80 万个候选模型的代谢网络集合，经多阶段筛选获得 10 个符合生理稳定性、动态响应特性和表型鲁棒性的高质量模型；该框架创新性地采用网络响应分析结合混合整数线性规划，在约束代谢物浓度扰动和酶活性调控幅度的前提下，系统枚举多靶点协同改造方案[99]。模型预测芳香族前体合成限速酶 DDPA 上调耦合谷氨酸代谢节点 GLUDy 下调的三靶点组合，在维持 90%基准生长率的同时使邻氨基苯甲酸产率提升 93%；模拟验证显示，NOMAD 设计方案在生物反应器中的终产物浓度(0.78 g/L)显著优于传统单靶点改造菌株(0.44 g/L)，且包含 8 个经文献验证的有效靶点(如 SHKK、PPS 等)，证实了该框架对复杂代谢网络调控效应的精准预测能力[99]。

以上介绍的是各个细胞过程的数学模拟方法。细胞是一个复杂系统，蛋白-核酸、蛋白-蛋白、蛋白-代谢物间存在十分复杂的相互作用，这种相互作用存在于不同的细胞过程之间，如代谢-翻译、代谢-转录、转录-代谢等。细胞

采用这种模块间调控来协同各个细胞过程。其次，模型主要包含代谢酶和转运蛋白的编码基因，在大肠杆菌等原核生物基因组中仅占 1/3 左右[100]，在酵母中约占 1/5[101]，基因组覆盖度较低，模型不能计算除代谢外的其他细胞过程。因此，不仅需要在单一过程水平上进行数学模拟，还需要整合多个过程，模拟细胞整体的变化。

3 多细胞过程的数学建模

3.1 约束模型

依照约束化学计量模型框架，尝试整合包括生化代谢在内的多细胞过程，实现更大范围的模拟计算。其中一个典型代表是 GECKO 酶约束模型[102]；这个模型引入了转录、翻译、蛋白折叠与分泌等多个细胞过程，扩大了原有生化代谢模型的模拟范围；并引入酶动力学和蛋白质组学约束，酶总量以及酶平均饱和度 σ，提高了模型的预测准确度。这一方法被成功应用于酿酒酵母、枯草芽孢杆菌、大肠杆菌、乳酸乳球菌(*Lactococcus lactis*)的研究中[103]。然而，GECKO 模型虽然拓展了模拟范围，但不同细胞过程之间仅有前体供需联系，例如代谢网络合成 GTP 和氨基酸作为蛋白翻译过程的底物，同时回收翻译产生的 GDP，并没有考虑翻译速率和代谢速率之间的关系。研究者们根据生物学机制引入额外约束实现过程耦合。例如，ETFL (expression and thermodynamics-enabled flux model)方法在 GSMM 基础上添加大分子合成、降解和生长稀释间的平衡约束、酶量约束、大分子合成速率约束、核糖体及 RNA 聚合酶等大分子机器总量约束以及热力学约束，计算基因表达和生长代谢速率，同时预测大分子胞内浓度变化[104]。这种方法目前用于 *E. coli* 模型中。RBA (resource balance analysis)方法引入分子合成、降解、大分子合成需求及生长稀释间的平衡约束，酶动力学约束、大分子合成速率约束、分子折叠约束以及细胞区室约束，可以准确预测生化代谢和基因表达速率。目前，模型已被应用于枯草芽孢杆菌、大肠杆菌[105]、需钠弧菌 (*Vibrio natriegens*)[106]、钩虫贪铜菌(*Cupriavidus necator*)等细菌中[107]，并且研发了相关工具 RBApy。另一类 RBA 方法则引入了模块之间的约束[108]。全基因组代谢与基因表达模型 (genome-scale model of metabolism and macromolecular expression, ME)的基础是 RNA/蛋白比例随生长速率的经验线性关系[109]，即 $\frac{R}{P}=k\mu+\lambda$。同时引入模块之间的约束关系，例如，合成核糖体的速率要足够快以满足蛋白质翻译需求，合成 RNA 聚合酶的速率要足够快以满足 RNA 转录需求，RNA 合成速率要快于其降解的速率，mRNA 合成速率要满足转录速率要求，tRNA 合成速率要满足 tRNA charging 需求，大分子复合体形成速率要满足其使用需求。代谢酶合成速率要满足其催化生化代谢反应的要求。细胞体积和表面积的增长速率是比生长速率的函数，脂质合成受到细胞膜表面积约束。由于约束方程中出现了比生长速率μ与反应速率 v 的乘积，使问题变成了非线性规划，大幅增加了求解难度。目前，这一模型架构在大肠杆菌、永达尔梭菌(*Clostridium ljungdahlii*)、乳酸乳球菌等细菌中得到了应用。

3.2 微分动力学模型

以上这些约束模型虽然扩展了模拟的范围，但依然存在稳态、强假设、生物过程简单化的缺点。全细胞混合动力学模型整合了微分动力学方程、概率方程、代谢方程、线性方程等多种数学方法刻画细胞多个生命过程，通过类数值积分算法实现了细胞组分连续动态变化的数值模拟。2012 年 Covert 团队从 900 篇文献

出发，针对有 500 多个基因的生殖道支原体(*Mycoplasma genitalium*)建立了第一个全细胞模型，包含 DNA 复制、转录、翻译、代谢等 28 个细胞过程模块，以及 DNA、RNA、蛋白、代谢物、细胞形态等 16 组状态变量，通过类数值积分算法实现了一个完整细胞周期的动态模拟，然而由于当时计算资源的匮乏，计算一个完整细胞周期需要约 1 d 的时间[15]。随后，Maritan 等[110]结合 AlphaFold 2 等工具，构建了对应的细胞 3D 模型。构建的全细胞模型可以用来和实验结果比较从而发现不吻合的地方，例如预测的与实验获得的单基因敲除菌株的比生长速率，从而发现注释错误的基因[111]。由于生殖道支原体的动力学参数相对较少，相关定量研究不多，模型大量采用了大肠杆菌等其他模式生物的动力学参数，因此模型质量和预测准确率上存在一定问题[112]。随后，Khodayari 等[98]使用 *E. coli* 专有数据，构建了基因组规模的 *E. coli* 的动力学模型。2015 年，Palsson 团队[113]采用 MASS 模型框架构建了人红细胞的全细胞动力学模型，准确模拟了生化代谢等细胞动态过程。2020 年，Liu 团队[114]针对酿酒酵母建立了一个全细胞模型 WM_S288C，包含了 975 个代谢物、6 156 个反应以及 6 447 个基因；该模型参照了 Covert 的模型结构，在模型中包含了 15 个细胞状态和 26 个细胞过程；该模型必需基因预测准确度达到 70%，对比 FBA 仅有 30%；利用该模型阐明了酿酒酵母中基因型和表型间的关系，预测了细胞周期内的资源分配和细胞行为，以及识别了细胞内核苷酸的调节机制。2021 年，Pelletier 等[115]针对全人工合成的具有 493 个基因的最小细胞 JCVIsyn3A 构建了全细胞动力学模型，与 Covert 的模型相比，其考虑了核糖体、DNA 等大分子在胞内的空间分布，而且采用了随机概率模型与决定论动力学模型相结合的方法以提高对低拷贝分子动态变化模拟的可靠性。2020 年，Covert 团队参考 1 200 篇文献，成功构建了 *E. coli* 全细胞混合动力学模型，包含 1 214 个基因、超过 10 000 个数学方程以及超过 19 000 个参数，其中 100%的参数由实验获得，与之前的生殖道支原体模型相比，新模型改进了细胞分裂模块，使得模型可以准确模拟细胞连续生长过程，突破了之前模型仅能模拟一个周期的限制，这之后，Covert 团队开展了 *E. coli* 全细胞模型计划[16-17]；并逐步添加模块，改进模型。随后，相关研究增加了 tRNA 氨酰化、肽链延伸和 N 端甲硫氨酸消除[116]、生长控制机制包括 ppGpp、氨基酸合成和翻译动力学[117]，扩大模型模拟的规模并提高模型预测的准确度。Skalnik 等[118]使用全细胞动力学模型研究了 *E. coli* 对抗生素作用的响应异质性机制。同时 Agmon 等[119]开发了 Vivarium、WholeCellKB 等工具开展细胞动力学模拟。KETCHUP 工具用来推断动力学模型参数新的基因功能(与模型预测不符的)[120]。

4 面临问题和解决思路

细胞是一个典型的复杂系统，构建数学模型刻画细胞生命过程存在着若干挑战。首先，测量的生物学数据不足。细胞由核酸、蛋白质、脂质等众多组分构成。一些大分子存在多种形式：例如 DNA 存在着甲基化修饰，RNA 存在着甲基化、假尿苷、双胞嘧啶修饰，蛋白质存在磷酸化、乙酰化、糖基化、甲基化、泛素化等不同修饰，代谢物可形成不同电离形式。受限于技术，大多工程细胞的细胞组分并未完全检测到，或仅能定性，无法定量。例如，基于液质联用(liquid chromatography-mass spectrometry, LC-MS)的代谢组学可以鉴定工程细胞内超

过 5 000 个特征峰，然而，搜索数据库仅能初步鉴定约 1 000 个代谢物[121]。脂质组学研究受限于标准品，能够准确定量的分子较少，远远低于细胞膜中脂质的丰富组成。构建模型时，变量数远远少于细胞实际的分子种类。另外，蛋白与 DNA、蛋白与 RNA、蛋白与蛋白、蛋白与小分子间存在着复杂的相互作用。当前组学技术多侧重于细胞组分定性定量分析，而对互作研究不够，模型中分子互作的类型和数量也很有限。因此，需要开发高通量高精度的组学分析方法，提高细胞组分以及组分互作定量分析的覆盖度和精确度，才能进一步地构建对应的数学模型。其次，机制不清，缺少准确的数学模型。模型是在深刻认识机理前提下对细胞生命过程的数学抽象化描述。例如，基因组尺度代谢模型是基于代谢反应速度快，代谢网络可以形成平衡的认知，利用线性规划方法求解生化代谢微分方程组稳态解 $\frac{dc}{dt}=Sv=0$；Michaelis-Menten 方程是对酶-底物复合体形成慢平衡、同时快速催化生成产物的数学概括。细胞是一个复杂系统，存在复杂性、涌现性。生物学复杂的本质导致清晰模拟细胞过程可能非常困难。目前，对大多数细胞过程的机理认知仅停留在定性描述，还没有总结概括出数理机制。另一方面，工程细胞需要兼顾产率、产量、得率、能量利用速率/效率、碳素效率、生产强度、原子经济性等生产属性，以及辅因子平衡、还原力平衡、能量供需平衡、遗传稳定性、耐受性、适应性、稳定性、协同性、适配性、正交性等工业属性。这些属性分子基础也不清楚，制约了对其数学刻画和优化。因此，需要开展细胞生命过程的定量分析，深入挖掘其分子机制。第三，精确机制模型复杂度过高，需要大量的参数估计。目前，广泛使用的基因组尺度代谢模型虽然构建简单，但仅能计算某个稳态下的得率，而非发酵产物浓度的动态变化；微分动力学方程可以模拟细胞组分的动态变化，但所需参数众多，构建难度大。因此，需要增加海量生物数据，尤其是关联多组学数据用于参数估计；同时使用人工智能方法学习公共数据，预测参数，或辅助参数估计；另外，可以根据实际需要适当简化模型。第四，细胞多层级过程的动态耦合与建模异质性。细胞内同时进行着 DNA 复制、转录、翻译、大分子修饰与降解、生化代谢、膜形成等过程，它们之间动态相互关联，但时间跨越秒级(代谢)、分级(转录、翻译、复制)、小时级(生长，蛋白稳定性)等多个尺度，空间上也有胞质(生化代谢)、细胞器(转录、翻译、复制)、细胞膜(膜形成)等不同分布。细胞组分物化性质相差较大，大分子(如多聚酶复合体)的扩散限制与空间可及性导致其反应动力学难以用经典质量作用定律描述，迫使模型对转录翻译等过程采用离散事件驱动的非连续数学框架。同时，多组学观测数据在时间颗粒度与空间覆盖度上的割裂，使得全细胞建模需要开发时空对齐算法与跨模态降维技术，从而实现原子级分子机制与宏观表型参数(发酵产率、生物量增长)的因果串联。这种数据-模型的双向尺度鸿沟，本质上是生命系统“局部精确性”与“全局一致性”不可兼得的计算困局，可通过开发混合建模范式，如机理方程约束下的机器学习校正来突破。最后，模型的实验验证不足，显著影响了可靠性和适用性。许多现有的模型在构建过程中，由于资源和时间有限，实验验证往往被简化处理，甚至在某些情况下完全跳过。实验验证的缺乏不仅削弱了模型对现实系统的模拟能力，还增加了模型结果不确定性的风险。此外，现阶段模型的构建和优化过程中，依赖实验数据进行的反馈和

迭代升级相对不足，未能充分利用实验数据对模型进行精准修正和调试。这种局限性使得模型的预测能力难以在动态和复杂的生物系统中得到广泛应用，亟待进一步解决。

2010 年以来，人工智能(artificial intelligence, AI)在生物制造与合成生物学研究中发挥着越来越重要的作用。运用机器学习能够从海量生物数据中提取关键信息预测细胞表型。在多肽设计、蛋白设计和基因编辑等特定领域，AI 已显著提升了研究的效率与精度。这一趋势不仅展现了 AI 在生物科学中的巨大潜力，也预示着生物学与计算科学的深度融合，将为未来生物技术的发展开辟新的路径。细胞是一个复杂的动力学系统，AI 也是一种复杂系统，规模法则和涌现性是跨越不同复杂系统的共性规律。AI 模型可以视作细胞复杂系统的通用压缩器，将细胞生成的生物大数据压缩进 AI 模型中，并通过归因分析和可解释 AI 技术来揭示细胞的运作规律。当前 AI 算法集中在蛋白元件设计以及生成式基因组设计方面，在细胞设计方面还没有显著进展。研究人员最近使用化学储备池计算(chemical reservoir computation)准确模拟了 *E. coli* 中心碳代谢的动态变化。聚糖反应是一种自组织化学反应网络，其优势是不需耗费大量资源计算反应机理，而只需对输出变量的线性加权做出预测。细胞中反应复杂，目前大多数的反应机制未知。因此，可以尝试使用化学储备池计算替代未知机理的反应，整体上实现细胞生命过程的准确预测[122]。在构建机理模型时结合 AI 技术，不仅可以快速推断未知规律并充分利用生物大数据，还能通过机理模型补充 AI 的机制，提高其可解释性。两者相辅相成，优势互补。

构建 AI 模型需要高质量的大数据。基于深度学习的 AlphaFold 模型依赖于高质量的序列-结构关联数据，其中序列数据主要来自 UniProt 等数据库，结构数据则来自 PDB 等数据库，这些数据源自数十年的实验研究成果的积累。同理，精准的工程细胞 AI 大模型的构建也需要高质量、标准化且具备因果关联的实验数据。在生物过程中，工程细胞的基因组背景是内因，包括基因表达的强化或弱化、基因敲除与插入，以及基因组的位点突变、大片段删除与敲入等遗传扰动。发酵过程的环境因素构成外因，包括孔板、摇瓶、发酵罐等不同发酵体系、培养基组成、取样时间、温度、溶氧、pH、渗透压以及胞外代谢物组等。细胞的基因组、转录组、蛋白组、代谢组、通量组、蛋白-DNA、蛋白-蛋白、蛋白-代谢物等相互作用网络的动态变化，是内外因共同作用下在分子微观水平上的体现。在宏观层面，细胞生长、胁迫耐受、发酵产量、转化率和生产强度等表型特征则是这种微观变化的最终表现。构建高效的细胞 AI 大模型的关键之一，在于获得全面且高质量的因果关联、确保宏观表型与微观分子层面紧密联系。即便是一个简单的原核细胞，也包含成千上万种不同的组分，并构成错综复杂的相互作用网络，其动态变化远比单个蛋白的折叠问题更为复杂。现有的数据库和文献中数据来源不同，数据质量参差不齐，实验条件千差万别，因此无法满足数据需求；同时，手工生成数据也难以满足规模化与标准化要求，亟须颠覆传统数据获取模式。因此，最优的解决方案是聚焦于特定的底盘细胞，依托先进大设施，规模化生产标准化、高质量的生物大数据，从而推动 AI 模型的精准构建与应用。大设施整合自动化高通量生物反应器、集成式多组学分析平台、智能化数据采集与管理系统以及大数据存储与处理平台等系统，能够在严格可控的条件下，高效、系统化、批量化地对工程细胞开展实验。

通过高通量实验设计与自动化执行，确保实验条件的一致性与可控性，减少人为误差，从而提高数据的可重复性和可靠性。总之，利用大设施进行高效的数据获取，不仅能够大幅提高实验效率和数据质量，还能形成标准化的内外因果关联，宏观微观对应的标准化数据资源库，为工程细胞 AI 大模型的开发和优化提供坚实的支撑。

5 结论与展望

精准全面的数字细胞模型是深入理解细胞生命活动、高效设计构建工程细胞的关键。本文总结了使用网络模型、概率模型、动力学模型等数学架构刻画生长分裂、形态发生、DNA 复制、转录调控、信号转导、群体效应、生化代谢等细胞过程的方式，并探讨了如何整合这些模块以构建全细胞模型的研究进展。此外，文章还讨论了在数学模型中描述细胞生命过程所面临的关键技术难题，探讨了未来机理与 AI 双驱动模型的可能性，为未来工程生物的设计与构建提供了的理论基础。

作者贡献声明

朱岩：构思总体框架、文献调研、综述撰写；孙际宾：内容补充修订、终稿审核。

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] CARRUTHERS DN, LEE TS. Translating advances in microbial bioproduction to sustainable biotechnology[J]. Frontiers in Bioengineering and Biotechnology, 2022, 10: 968437.

[2] 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 industry is the key development direction of bioeconomy[J]. China Biotechnology, 2022, 42(5): 4-5 (in Chinese).

[3] CARBONELL P, JERVIS AJ, ROBINSON CJ, YAN CY, DUNSTAN M, SWAINSTON N, VINAIXA M, HOLLYWOOD KA, CURRIN A, RATTRAY NJW, TAYLOR S, SPIESS R, SUNG R, WILLIAMS AR, FELLOWS D, STANFORD NJ, MULHERIN P, Le FEUVRE R, BARRAN P, GOODACRE R, et al. An automated Design-Build- Test-Learn pipeline for enhanced microbial production of fine chemicals[J]. Communications Biology, 2018, 1: 66.

[4] SHENDURE J, BALASUBRAMANIAN S, CHURCH GM, GILBERT W, ROGERS J, SCHLOSS JA, WATERSTON RH. DNA sequencing at 40: past, present and future[J]. Nature, 2017, 550(7676): 345-353.

[5] TAMARA S, den BOER MA, HECK AJR. High-resolution native mass spectrometry[J]. Chemical Reviews, 2022, 122(8): 7269-7326.

[6] WANG DQ, HE PS, WANG ZJ, LI GY, MAJED N, GU AZ. Advances in single cell Raman spectroscopy technologies for biological and environmental applications[J]. Current Opinion in Biotechnology, 2020, 64: 218-229.

[7] O'CONNOR E, MICKLEFIELD J, CAI YZ. Searching for the optimal microbial factory: high-throughput biosensors and analytical techniques for screening small molecules[J]. Current Opinion in Biotechnology, 2024, 87: 103125.

[8] CARBONELL P, RADIVOJEVIC T, GARCÍA MARTÍN H. Opportunities at the intersection of synthetic biology, machine learning, and automation[J]. ACS Synthetic Biology, 2019, 8(7): 1474-1477.

[9] LEE SY, KIM HU. Systems strategies for developing industrial microbial strains[J]. Nature Biotechnology, 2015, 33(10): 1061-1072.

[10] HOOK PW, TIMP W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology[J]. Nature Reviews Genetics, 2023, 24(9): 627-641.

[11] BAYSOY A, BAI ZL, SATIJA R, FAN R. The technological landscape and applications of single-cell multi-omics[J]. Nature Reviews Molecular Cell Biology, 2023, 24(10): 695-713.

[12] BRESSAN D, BATTISTONI G, HANNON GJ. The dawn of spatial omics[J]. Science, 2023, 381(6657): eabq4964.

[13] SCHUBERT OT, RÖST HL, COLLINS BC, ROSENBERGER G, AEBERSOLD R. Quantitative proteomics: challenges and opportunities in basic and applied research[J]. Nature Protocols, 2017, 12(7): 1289-1294.

[14] PEREZ de SOUZA L, ALSEEKH S, SCOSSA F, FERNIE AR. Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research[J]. Nature Methods, 2021, 18(7): 733-746.

[15] 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.
[16] SUN G, AHN-HORST TA, COVERT MW. The *E. coli* whole-cell modeling project[J]. EcoSal Plus, 2021, 9(2): eESP00012020.
[17] 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.
[18] CHEN Y, BANERJEE D, MUKHOPADHYAY A, PETZOLD CJ. Systems and synthetic biology tools for advanced bioproduction hosts[J]. Current Opinion in Biotechnology, 2020, 64: 101-109.
[19] MACHADO D, COSTA RS, ROCHA M, FERREIRA EC, TIDOR B, ROCHA I. Modeling formalisms in systems biology[J]. AMB Express, 2011, 1: 45.
[20] KOUTROULI M, KARATZAS E, PAEZ-ESPINO D, PAVLOPOULOS GA. A guide to conquer the biological network era using graph theory[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 34.
[21] PENG C, CHEN Q, TAN SJ, SHEN XT, JIANG C. Generalized reporter score-based enrichment analysis for omics data[J]. Briefings in Bioinformatics, 2024, 25(3): bbae116.
[22] LIU F, HEINER M, GILBERT D. Fuzzy Petri nets for modelling of uncertain biological systems[J]. Briefings in Bioinformatics, 2020, 21(1): 198-210.
[23] KADELKA C, BUTRIE TM, HILTON E, KINSETH J, SCHMIDT A, SERDAREVIC H. A meta-analysis of Boolean network models reveals design principles of gene regulatory networks[J]. Science Advances, 2024, 10(2): eadj0822.
[24] ZHAO MY, HE WY, TANG JJ, ZOU Q, GUO F. A comprehensive overview and critical evaluation of gene regulatory network inference technologies[J]. Briefings in Bioinformatics, 2021, 22(5): bbab009.
[25] BORDBAR A, MONK JM, KING ZA, PALSSON BO. Constraint-based models predict metabolic and associated cellular functions[J]. Nature Reviews Genetics, 2014, 15(2): 107-120.
[26] DU DZ, PARDALOS PM, WU W. Mathematical theory of optimization[M]. Nonconvex Optimization and Its Applications: Springer Science & Business Media, 2013.
[27] WILKINSON DJ. Stochastic Modelling for Systems Biology[M]. Third Edition. Milton: Chapman and Hall/CRC, 2018.
[28] DAIGLE BJ Jr, SRINIVASAN BS, FLANNICK JA, NOVAK AF, BATZOGLOU S. Current progress in static and dynamic modeling of biological networks[M]//Systems Biology for Signaling Networks. New York, NY: Springer New York, 2010: 13-73.
[29] LU J, EMRAH Ş, SILVER A, YOU LC. Advances and challenges in programming pattern formation using living cells[J]. Current Opinion in Chemical Biology, 2022, 68: 102147.
[30] CHAUDHURI O, COOPER-WHITE J, JANMEY PA, MOONEY DJ, SHENOY VB. Effects of extracellular matrix viscoelasticity on cellular behaviour[J]. Nature, 2020, 584(7822): 535-546.
[31] ANTONAKOUDIS A, BARBOSA R, KOTIDIS P, KONTORAVDI C. The era of big data: genome-scale modelling meets machine learning[J]. Computational and Structural Biotechnology Journal, 2020, 18: 3287-3300.
[32] ASHYRALIYEV M, FOMEKONG-NANFACK Y, KAANDORP JA, BLOM JG. Systems biology: parameter estimation for biochemical models[J]. The FEBS Journal, 2009, 276(4): 886-902.
[33] TYSON JJ, NOVAK B. A dynamical paradigm for molecular cell biology[J]. Trends in Cell Biology, 2020, 30(7): 504-515.
[34] BERGMANN FT, HOOPS S, KLAHN B, KUMMER U, MENDES P, PAHLE J, SAHLE S. COPASI and its applications in biotechnology[J]. Journal of Biotechnology, 2017, 261: 215-220.
[35] HEIRENDT L, ARRECKX S, PFAU T, MENDOZA SN, RICHELLE A, HEINKEN A, HARALDSDÓTTIR HS, WACHOWIAK J, KEATING SM, VLASOV V, MAGNUSDÓTTIR S, NG CY, PRECIAT G, ŽAGARE A, CHAN SHJ, AURICH MK, CLANCY CM, MODAMIO J, SAULS JT, NORONHA A, 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.
[36] 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.
[37] SHIN J, PORUBSKY V, CAROTHERS J, SAURO HM. Standards, dissemination, and best practices in systems biology[J]. Current Opinion in Biotechnology, 2023, 81: 102922.
[38] GHENU AH, MARREC L, BANK C. Challenges and pitfalls of inferring microbial growth rates from lab cultures[J]. Frontiers in Ecology and Evolution, 2024, 11: 1313500.
[39] ADHUKHAN S, VILLA R, SARKAR U. Microbial production of succinic acid using crude and purified glycerol from a Crotalaria juncea based biorefinery[J]. Biotechnology Reports (Amst), 2016, 10: 84-93.
[40] KENKRE V and KUPERMAN M. Applicability of the Fisher equation to bacterial population dynamics[J].

Physical Review E, 2003, 67(5): 051921.
[41] HARRIS LK, THERIOT JA. Relative rates of surface and volume synthesis set bacterial cell size[J]. Cell, 2016, 165(6): 1479-1492.
[42] HO PY, AMIR A. Simultaneous regulation of cell size and chromosome replication in bacteria[J]. Frontiers in Microbiology, 2015, 6: 662.
[43] COOPER S, HELMSTETTER CE. Chromosome replication and the division cycle of *Escherichia coli* Br[J]. Journal of Molecular Biology, 1968, 31(3): 519-540.
[44] ZHENG H, BAI Y, JIANG ML, TOKUYASU TA, HUANG XL, ZHONG FJ, WU YQ, FU XF, KLECKNER N, HWA T, LIU CL. General quantitative relations linking cell growth and the cell cycle in *Escherichia coli*[J]. Nature Microbiology, 2020, 5(8): 995-1001.
[45] DiNAPOLI KT, ROBINSON DN, IGLESIAS PA. Tools for computational analysis of moving boundary problems in cellular mechanobiology[J]. Wiley Interdisciplinary Reviews Systems Biology and Medicine, 2020: e1514.
[46] POIRIER CC, NG WP, ROBINSON DN, IGLESIAS PA. Deconvolution of the cellular force-generating subsystems that govern cytokinesis furrow ingression[J]. PLoS Computational Biology, 2012, 8(4): e1002467.
[47] SUROVTSEV IV, MORGAN JJ, LINDAHL PA. Kinetic modeling of the assembly, dynamic steady state, and contraction of the FtsZ ring in prokaryotic cytokinesis[J]. PLoS Computational Biology, 2008, 4(7): e1000102.
[48] LI FT, LONG T, LU Y, OUYANG Q, TANG C. The yeast cell-cycle network is robustly designed[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(14): 4781-4786.
[49] OKAZ E, ARGÜELLO-MIRANDA O, BOGDANOVA A, VINOD PK, LIPP JJ, MARKOVA Z, ZAGORIY I, NOVAK B, ZACHARIAE W. Meiotic prophase requires proteolysis of M phase regulators mediated by the meiosis-specific APC/CAma1[J]. Cell, 2012, 151(3): 603-618.
[50] NARULA J, KUCHINA A, ZHANG F, FUJITA M, SÜEL GM, IGOSHIN OA. Slowdown of growth controls cellular differentiation[J]. Molecular Systems Biology, 2016, 12(5): 871.
[51] KÜFFNER R, PETRI T, TAVAKKOLKHAH P, WINDHAGER L, ZIMMER R. Inferring gene regulatory networks by ANOVA[J]. Bioinformatics, 2012, 28(10): 1376-1382.
[52] MARGOLIN AA, NEMENMAN I, BASSO K, WIGGINS C, STOLOVITZKY G, DALLA FAVERA R, CALIFANO A. ARACNE an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context[J]. BMC Bioinformatics, 2006, 7(Suppl 1): S7.
[53] FAITH JJ, HAYETE B, THADEN JT, MOGNO I, WIERZBOWSKI J, COTTAREL G, KASIF S, COLLINS JJ, GARDNER TS. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles[J]. PLoS Biology, 2007, 5(1): e8.
[54] MEYER PE, KONTOS K, LAFITTE F, BONTEMPI G. Information-theoretic inference of large transcriptional regulatory networks[J]. EURASIP Journal on Bioinformatics & Systems Biology, 2007, 2007(1): 79879.
[55] RESHEF DN, RESHEF YA, FINUCANE HK, GROSSMAN SR, McVEAN G, TURNBAUGH PJ, LANDER ES, MITZENMACHER M, SABETI PC. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062): 1518-1524.
[56] ZHANG XJ, ZHAO XM, HE K, LU L, CAO YW, LIU JD, HAO JK, LIU ZP, CHEN LN. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information[J]. Bioinformatics, 2012, 28(1): 98-104.
[57] ZHAO J, ZHOU YW, ZHANG XJ, CHEN LN. Part mutual information for quantifying direct associations in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2016, 113(18): 5130-5135.
[58] MORDELET F, VERT JP. *SIRENE*: supervised inference of regulatory networks[J]. Bioinformatics, 2008, 24(16): i76-i82.
[59] KREMLING A, GEISELMANN J, ROPERS D, JONG HD. Understanding carbon catabolite repression in *Escherichia coli* using quantitative models[J]. Trends in Microbiology, 2015, 23(2): 99-109.
[60] SCHULTZ D, WOLYNES PG, BEN JACOB E, ONUCHIC JN. Deciding fate in adverse times: sporulation and competence in *Bacillus subtilis*[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(50): 21027-21034.
[61] NILSSON P, OLOFSSON A, FAGERLIND M, FAGERSTROM T, RICE S, KJELLEBERG S, STEINBERG P. Kinetics of the AHL regulatory system in a model biofilm system: how many bacteria constitute a “quorum”?[J]. Journal of Molecular Biology, 2001, 309(3): 631-640.
[62] WU SB, FENG J, LIU CJ, WU H, QIU ZK, GE JJ, SUN SY, HONG X, LI YK, WANG XN, YANG AD, GUO F, QIAO JJ. Machine learning aided construction of the quorum sensing communication network for human gut microbiota[J]. Nature Communications, 2022, 13(1): 3079.
[63] LEE JY, SADLER NC, EGBERT RG, ANDERTON CR, HOFMOCKEL KS, JANSSON JK, SONG HS. Deep learning predicts microbial interactions from self-organized spatiotemporal patterns[J]. Computational and Structural Biotechnology Journal,

2020, 18: 1259-1269.
[64] WANG LY, WONG YC, CORREIRA JM, WANCURA M, GEIGER CJ, WEBSTER SS, TOUHAMI A, BUTLER BJ, O'TOOLE GA, LANGFORD RM, BROWN KA, DORTDIVANLIOGLU B, WEBB L, COSGRIFF-HERNANDEZ E, GORDON VD. The accumulation and growth of *Pseudomonas aeruginosa* on surfaces is modulated by surface mechanics *via* cyclic-di-GMP signaling[J]. NPJ Biofilms and Microbiomes, 2023, 9(1): 78.
[65] VARMA A, BOESCH BW, PALSSON BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates[J]. Applied and Environmental Microbiology, 1993, 59(8): 2465-2473.
[66] SCHILLING CH, PALSSON BO. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis[J]. Journal of Theoretical Biology, 2000, 203(3): 249–283.
[67] GU CD, KIM GB, KIM WJ, KIM HU, LEE SY. Current status and applications of genome-scale metabolic models[J]. Genome Biology, 2019, 20(1): 121.
[68] HEINKEN A, HULSHOF TO, NAP B, MARTINELLI F, BASILE A, O'BROLCHAIN A, O'SULLIVAN NF, GALLAGHER C, MAGEE E, McDONAGH F, LALOR I, BERGIN M, EVANS P, DALY R, FARRELL R, DELANEY RM, HILL S, McAULIFFE SR, KILGANNON T, FLEMING RMT, et al. APOLLO: a genome-scale metabolic reconstruction resource of 247 092 diverse human microbes spanning multiple continents, age groups, and body sites[J]. bioRxiv, 2023: 2023.10.02.560573.
[69] BRUNK E, SAHOO S, ZIELINSKI DC, ALTUNKAYA A, DRÄGER A, MIH N, GATTO F, NILSSON A, PRECIAT GONZALEZ GA, AURICH MK, PRLIĆ A, SASTRY A, DANIELSDOTTIR AD, HEINKEN A, NORONHA A, ROSE PW, BURLEY SK, FLEMING RMT, NIELSEN J, THIELE I, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism[J]. Nature Biotechnology, 2018, 36(3): 272-281.
[70] FOGUET C, XU Y, RITCHIE SC, LAMBERT SA, PERSYN E, NATH AP, DAVENPORT EE, ROBERTS DJ, PAUL DS, Di ANGELANTONIO E, DANESH J, BUTTERWORTH AS, YAU C, INOUYE M. Genetically personalised organ-specific metabolic models in health and disease[J]. Nature Communications, 2022, 13(1): 7356.
[71] THIELE I, PALSSON BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction[J]. Nature Protocols, 2010, 5(1): 93-121.
[72] LIEVEN C, BEBER ME, OLIVIER BG, BERGMANN FT, ATAMAN M, BABAEI P, BARTELL JA, BLANK LM, CHAUHAN S, CORREIA K, DIENER C, DRÄGER A, EBERT BE, EDIRISINGHE JN, FARIA JP, FEIST AM, FENGOS G, FLEMING RMT, GARCÍA-JIMÉNEZ B, HATZIMANIKATIS V, et al. MEMOTE for standardized genome-scale metabolic model testing[J]. Nature Biotechnology, 2020, 38(3): 272-276.
[73] NILSSON A, NIELSEN J. Genome scale metabolic modeling of cancer[J]. Metabolic Engineering, 2017, 43: 103-112.
[74] BEARD DA, LIANG SD, QIAN H. Energy balance for analysis of complex metabolic networks[J]. Biophysical Journal, 2002, 83(1): 79-86.
[75] HENRY CS, BROADBELT LJ, HATZIMANIKATIS V. Thermodynamics-based metabolic flux analysis[J]. Biophysical Journal, 2007, 92(5): 1792-1805.
[76] NIEBEL B, LEUPOLD S, HEINEMANN M. An upper limit on Gibbs energy dissipation governs cellular metabolism[J]. Nature Metabolism, 2019, 1(1): 125-132.
[77] 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.
[78] BEG QK, VAZQUEZ A, ERNST J, de MENEZES MA, BAR-JOSEPH Z, BARABÁSI AL, OLTVAI ZN. 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.
[79] ADADI R, VOLKMER B, MILO R, HEINEMANN M, SHLOMI T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters[J]. PLoS Computational Biology, 2012, 8(7): e1002575.
[80] COVERT MW, XIAO N, CHEN TJ, KARR JR. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*[J]. Bioinformatics, 2008, 24(18): 2044-2050.
[81] BURGARD AP, PHARKYA P, MARANAS CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization[J]. Biotechnology and Bioengineering, 2003, 84(6): 647-657.
[82] PHARKYA P, MARANAS CD. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems[J]. Metabolic Engineering, 2006, 8(1): 1-13.
[83] KIM J, REED JL. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains[J]. BMC Systems Biology, 2010, 4: 53.
[84] JIANG SY, OTERO-MURAS I, BANGA JR, WANG Y, KAISER M, KRASNOGOR N. OptDesign: identifying optimum design strategies in strain engineering for

biochemical production[J]. ACS Synthetic Biology, 2022, 11(4): 1531-1541.
[85] SHEN FZ, SUN RL, YAO J, LI J, LIU Q, PRICE ND, LIU CG, WANG Z. OptRAM: in-silico strain design *via* integrative regulatory-metabolic network modeling[J]. PLoS Computational Biology, 2019, 15(3): e1006835.
[86] PHARKYA P, BURGARD AP, MARANAS CD. OptStrain: a computational framework for redesign of microbial production systems[J]. Genome Research, 2004, 14(11): 2367-2376.
[87] JIANG SY, WANG Y, KAISER M, KRASNOGOR N. NIHBA: a network interdiction approach for metabolic engineering design[J]. Bioinformatics, 2020, 36(11): 3482-3492.
[88] RANGANATHAN S, SUTHERS PF, MARANAS CD. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions[J]. PLoS Computational Biology, 2010, 6(4): e1000744.
[89] STANFORD NJ, MILLARD P, SWAINSTON N. RobOKoD: microbial strain design for (over)production of target compounds[J]. Frontiers in Cell and Developmental Biology, 2015, 3: 17.
[90] ASADOLLAHI MA, MAURY J, PATIL KR, SCHALK M, CLARK A, NIELSEN J. Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering[J]. Metabolic Engineering, 2009, 11(6): 328-334.
[91] HASSANI L, MOOSAVI MR, SETOODEH P, ZARE H. FastKnock: an efficient next-generation approach to identify all knockout strategies for strain optimization[J]. Microbial Cell Factories, 2024, 23(1): 37.
[92] YANG JE, PARK SJ, KIM WJ, KIM HJ, KIM BJ, LEE H, SHIN J, LEE SY. One-step fermentative production of aromatic polyesters from glucose by metabolically engineered *Escherichia coli* strains[J]. Nature Communications, 2018, 9(1): 79.
[93] ISHCHUK OP, DOMENZAIN I, SÁNCHEZ BJ, MUÑIZ-PAREDES F, MARTÍNEZ JL, NIELSEN J, PETRANOVIC D. Genome-scale modeling drives 70-fold improvement of intracellular heme production in *Saccharomyces cerevisiae*[J]. Proceedings of the National Academy of Sciences of the United States of America, 2022, 119(30): e2108245119.
[94] 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.
[95] de LEEUW M, MATOS MRA, NIELSEN LK. Omics data for sampling thermodynamically feasible kinetic models[J]. Metabolic Engineering, 2023, 78: 41-47.
[96] JAHAN N, MAEDA K, MATSUOKA Y, SUGIMOTO Y, KURATA H. Development of an accurate kinetic model for the central carbon metabolism of *Escherichia coli*[J]. Microbial Cell Factories, 2016, 15(1): 112.
[97] KHODAYARI A, ZOMORRODI AR, LIAO JC, MARANAS CD. A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data[J]. Metabolic Engineering, 2014, 25: 50-62.
[98] KHODAYARI A, MARANAS CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains[J]. Nature Communications, 2016, 7: 13806.
[99] NARAYANAN B, WEILANDT D, MASID M, MISKOVIC L, HATZIMANIKATIS V. Rational strain design with minimal phenotype perturbation[J]. Nature Communications, 2024, 15(1): 723.
[100] MONK JM, LLOYD CJ, BRUNK E, MIH N, SASTRY A, KING Z, TAKEUCHI R, NOMURA W, ZHANG Z, MORI H, FEIST AM, PALSSON BO. iML1515, a knowledgebase that computes *Escherichia coli* traits[J]. Nature Biotechnology, 2017, 35(10): 904-908.
[101] ZHANG C, SANCHEZ BJ, LI F, EIDEN CWQ, SCOTT WT, LIEBAL UW, BLANK LM, MENGERS HG, ANTON M, RANGEL AT, MENDOZA SN, ZHANG L, NIELSEN J, LU H, KERKHOVEN EJ. Yeast9: a consensus genome-scale metabolic model for *S. cerevisiae* curated by the community[J]. Molecular Systems Biology, 2024, 20(10): 1134-1150.
[102] CHEN Y, GUSTAFSSON J, TAFUR RANGEL A, ANTON M, DOMENZAIN I, KITTIKUNAPONG C, LI FR, YUAN L, NIELSEN J, KERKHOVEN EJ. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0[J]. Nature Protocols, 2024, 19(3): 629-667.
[103] CHEN Y, van PELT-KLEINJAN E, van OLST B, DOUWENGA S, BOEREN S, BACHMANN H, MOLENAAR D, NIELSEN J, TEUSINK B. Proteome constraints reveal targets for improving microbial fitness in nutrient-rich environments[J]. Molecular Systems Biology, 2021, 17(4): e10093.
[104] 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.
[105] OFTADEH O, HATZIMANIKATIS V. Genome-scale models of metabolism and expression predict the metabolic burden of recombinant protein expression[J]. Metabolic Engineering, 2024, 84: 109-116.
[106] COPPENS L, TSCHIRHART T, LEARY DH, COLSTON SM, COMPTON JR, HERVEY WJ 4th, DANA KL, VORA GJ, BORDEL S, LEDESMA-AMARO R. *Vibrio natriegens* genome-scale modeling reveals insights into halophilic adaptations and resource allocation[J]. Molecular Systems Biology, 2023, 19(4): e10523.
[107] JAHN M, CRANG N, JANASCH M, HOBER A, FORSSTRÖM B, KIMLER K, MATTAUSCH A,

CHEN Q, ASPLUND-SAMUELSSON J, HUDSON EP. Protein allocation and utilization in the versatile chemolithoautotroph *Cupriavidus necator*[J]. eLife, 2021, 10: e69019.

[108] DINH HV, MARANAS CD. Evaluating proteome allocation of *Saccharomyces cerevisiae* phenotypes with resource balance analysis[J]. Metabolic Engineering, 2023, 77: 242-255.

[109] SCOTT M, GUNDERSON CW, MATEESCU EM, ZHANG ZG, HWA T. Interdependence of cell growth and gene expression: origins and consequences[J]. Science, 2010, 330(6007): 1099-1102.

[110] MARITAN M, AUTIN L, KARR J, COVERT MW, OLSON AJ, GOODSELL DS. Building structural models of a whole *Mycoplasma* cell[J]. Journal of Molecular Biology, 2022, 434(2): 167351.

[111] SANGHVI JC, REGOT S, CARRASCO S, KARR JR, GUTSCHOW MV, BOLIVAL B Jr, COVERT MW. Accelerated discovery *via* a whole-cell model[J]. Nature Methods, 2013, 10(12): 1192-1195.

[112] KARR JR, WILLIAMS AH, ZUCKER JD, RAUE A, STEIERT B, TIMMER J, KREUTZ C, WILKINSON S, ALLHOOD BA, BOT BM, HOFF BR, KELLEN MR, COVERT MW, STOLOVITZKY GA, MEYER P. Summary of the DREAM8 parameter estimation challenge: toward parameter identification for whole-cell models. PLoS Computational Biology, 2015, 11(5): e1004096.

[113] BORDBAR A, McCLOSKEY D, ZIELINSKI DC, SONNENSCHEIN N, JAMSHIDI N, PALSSON BO. Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics[J]. Cell Systems, 2015, 1(4): 283-292.

[114] YE C, XU N, GAO C, LIU GQ, XU JZ, ZHANG WG, CHEN XL, NIELSEN J, LIU LM. Comprehensive understanding of *Saccharomyces cerevisiae* phenotypes with whole-cell model WM_S288C[J]. Biotechnology and Bioengineering, 2020, 117(5): 1562-1574.

[115] PELLETIER JF, SUN LJ, WISE KS, ASSAD-GARCIA N, KARAS BJ, DEERINCK TJ, ELLISMAN MH, MERSHIN A, GERSHENFELD N, CHUANG RY, GLASS JI, STRYCHALSKI EA. Genetic requirements for cell division in a genomically minimal cell[J]. Cell, 2021, 184(9): 2430-2440.e16.

[116] CHOI H, COVERT MW. Whole-cell modeling of *E. coli* confirms that *in vitro* tRNA aminoacylation measurements are insufficient to support cell growth and predicts a positive feedback mechanism regulating arginine biosynthesis[J]. Nucleic Acids Research, 2023, 51(12): 5911-5930.

[117] AHN-HORST TA, MILLE LS, SUN G, MORRISON JH, COVERT MW. An expanded whole-cell model of *E. coli* links cellular physiology with mechanisms of growth rate control[J]. NPJ Systems Biology and Applications, 2022, 8(1): 30.

[118] SKALNIK CJ, CHEAH SY, YANG MY, WOLFF MB, SPANGLER RK, TALMAN L, MORRISON JH, PEIRCE SM, AGMON E, COVERT MW. Whole-cell modeling of *E. coli* colonies enables quantification of single-cell heterogeneity in antibiotic responses[J]. PLoS Computational Biology, 2023, 19(6): e1011232.

[119] AGMON E, SPANGLER RK, SKALNIK CJ, POOLE W, PEIRCE SM, MORRISON JH, COVERT MW. Vivarium: an interface and engine for integrative multiscale modeling in computational biology[J]. Bioinformatics, 2022, 38(7): 1972-1979.

[120] HU MQ, SUTHERS PF, MARANAS CD. KETCHUP: Parameterizing of large-scale kinetic models using multiple datasets with different reference states[J]. Metabolic Engineering, 2024, 82: 123-133.

[121] MAYERS JR, VARON J, ZHOU RR, DANIEL-IVAD M, BEAULIEU C, BHOSLE A, GLASSER NR, LICHTENAUER FM, NG J, VERA MP, HUTTENHOWER C, PERRELLA MA, CLISH CB, ZHAO SD, BARON RM, BALSKUS EP. A metabolomics pipeline highlights microbial metabolism in bloodstream infections[J]. Cell, 2024, 187(15): 4095-4112 e21.

[122] BALTUSSEN MG, DE JONG TJ, DUEZ Q, ROBINSON WE, HUCK WTS. Chemical reservoir computation in a self-organizing reaction network[J]. Nature, 2024, 631(8021): 549-555.