研究论文

DOI: 10.12211/2096-8280.2026-100

PINN驱动的连续PURE系统数字孪生模型

杨晨[1]，刘金林[1]，方洲[2]，马红武[1]
（[1] 中国科学院天津工业生物技术研究所，低碳合成工程生物学全国重点实验室，生物设计中心，天津300308；[2] 中国科学院数学与系统科学研究院，数学科学全国重点实验室，北京 100190）

摘要：无细胞蛋白合成系统（Protein Synthesis Using Recombinant Elements，PURE）数字孪生模型，是解析体外自复制系统通量调控和物质循环，并理性设计自复制合成生命的重要工具。然而，连续式PURE体系涉及能量再生、副产物积累及大量短寿命中间态，传统显式动力学模型在长时程模拟中面临数值刚性的难题，纯数据驱动的机器学习方法又因缺乏充分的物理约束导致预测偏差和生物学解释性差。针对上述问题，本研究基于物理约束学习方法提出并构建了具备基元反应精细度的物理信息神经网络驱动的连续PURE系统数字孪生模型。以质量作用动力学、化学计量守恒和开放边界交换的降阶教师模型作为机制化数据源，利用单/多蛋白表达体系的3151/9198个可比代谢物浓度轨迹和扰动分析提取的能量循环、核糖体装载等关键代谢物节点，统一构建浓度监督轨迹和物理约束。在耦合微观中间态浓度和隐藏反应通量的推断结果后，模型实现了基元反应精细度下连续PURE体系代谢物浓度动态轨迹的精准重构。模拟结果显示，验证集中代谢物浓度预测的平均相对误差（Mean Relative Error，MRE）分别为23.57%和0.92%，决定系数（Coefficient of Determination，R^2）分别高达0.9771和0.9863，并合理推断了IF1/IF3介导的翻译起始装配模块等部分隐藏状态。本文的模型为构建稳定运行的自复制体外合成生命系统提供了定量分析的计算平台。

关键词：连续式PURE系统；数字孪生；物理信息神经网络；基元反应；物理约束学习

中图分类号：Q811.4　　**文献标志码：**A

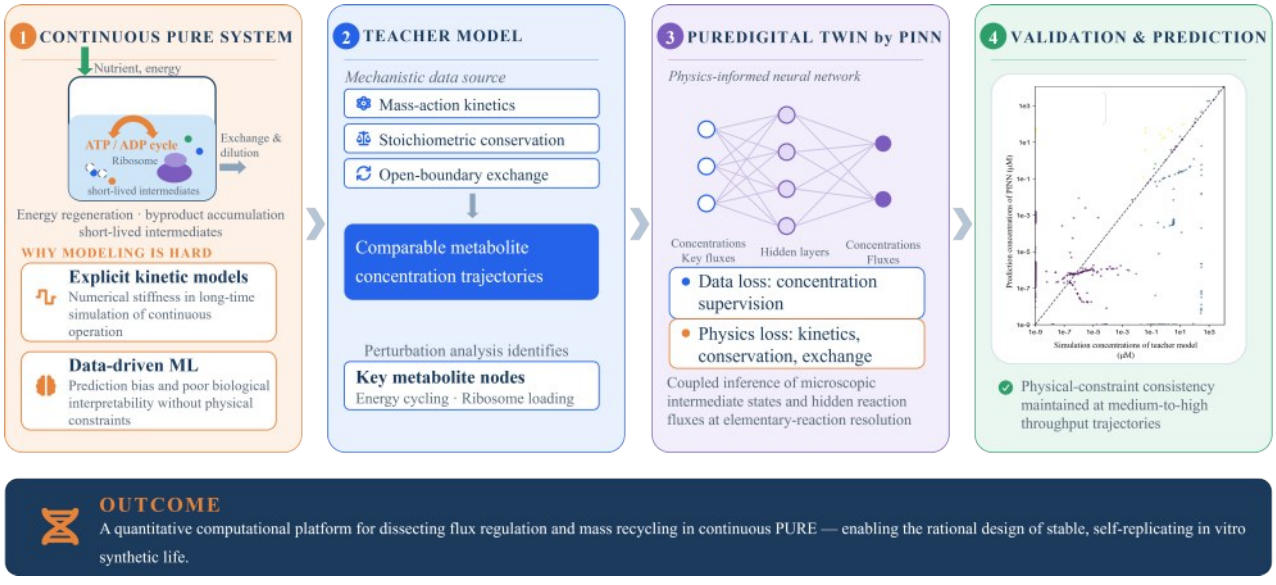
A PINN-Driven Digital Twin Model of a Continuous PURE System

YANG Chen[1]，LIU Jinlin[1]，FANG Zhou[2]，MA Hongwu[1]
（[1]*Biodesign Center，State Key Laboratory of Engineering Biology for Low-Carbon Manufacturing，Tianjin Institute of Industrial Biotechnology，Chinese Academy of Sciences，Tianjin 300308，China*；[2]*State Key Laboratory of Mathematical Sciences，Academy of Mathematics and Systems Science，Chinese Academy of Sciences，Beijing 100190，China*）

Abstract: Digital twin models of Protein Synthesis Using Recombinant Elements (PURE) systems are important tools for quantitatively analyzing flux regulation and material cycling in in vitro self-replicating systems, and for

收稿日期：2026-09-08 修回日期：2026-11-24
基金项目：国家重点研发计划（2025YFA0922800）；中国科学院战略性先导科技专项（XDB0480000）；中国自然科学基金（12501699）；数学科学全国重点实验室开放课题合作团队项目（SLMS-2025-KFKT-TD-01）
引用本文：杨晨，刘金林，方洲，马红武. PINN驱动的连续PURE系统数字孪生模型[J]. 合成生物学，2026，7. DOI：10.12211/2096-8280.2026-100
Citation：YANG Chen，LIU Jinlin，FANG Zhou，MA Hongwu. A PINN-Driven Digital Twin Model of a Continuous PURE System[J]. Synthetic Biology Journal，2026，7. DOI：10.12211/2096-8280.2026-100

providing a computational basis for the rational design of self-replicating synthetic life. However, continuous PURE systems involve energy regeneration, byproduct accumulation, and numerous short-lived intermediates. These features present traditional explicit kinetic models with the challenge of numerical stiffness when long-term simulations are performed. Purely data-driven machine learning methods, meanwhile, may produce biased predictions and exhibit poor biological interpretability because they lack sufficient physical constraints. To address these challenges, this study proposes and develops a Physics-Informed Neural Network (PINN)-driven digital twin model of a continuous PURE system with elementary-reaction resolution based on physics-constrained learning. The model was applied to a deGFP single-protein expression system and a six-protein co-expression system. Using reduced teacher models based on mass-action kinetics, stoichiometric conservation, and open-boundary exchange as mechanistic data sources, this study constructed concentration-supervised trajectories and physical constraints from 3,151 and 9,198 comparable metabolite concentration trajectories in the two expression systems, respectively, together with key metabolite nodes associated with energy cycling, ribosome loading, and other processes identified by perturbation analysis. After coupling the inferred concentrations of microscopic intermediates with the inferred fluxes of hidden reactions, the model achieved accurate reconstruction of the dynamic metabolite concentration trajectories of the continuous PURE system at elementary-reaction resolution. The simulation results showed that, in the validation set, the Mean Relative Errors (MREs) of metabolite concentration prediction was 23.57% and 0.92%, and the Coefficient of Determination (R^2) reached 0.9771 and 0.9863, respectively, for the deGFP and six-protein expression systems. The model also reasonably inferred several hidden states, including those associated with the IF1/IF3-mediated translation initiation assembly module. This model provides a quantitative computational platform for constructing stably operating and self-replicating in vitro synthetic life systems.



Keywords: Continuous PURE system; Digital twin; Physics-Informed Neural Network; Elementary reaction; Physics-constrained learning

1 研究背景

理解生命运作的机理，探索生命本质是“自下而上”合成细胞的终极追求。然而，活体细胞内部错综复杂的代谢旁路与背景调控，极大阻碍了对于生命演化机理的定量认知[1, 2]。为了跨越这

一认知鸿沟，无细胞蛋白合成系统（Protein Synthesis Using Recombinant Elements， PURE）数字孪生已成为合成细胞研究的核心物理原型[3，4]。PURE系统去除了天然细胞内复杂的基因调控网络，仅由高度明确的转录、翻译、氨酰化与能量再生等核心模块组成[5，6]。不仅能够突破天然细胞内分子状态与通量原位监测的屏障，更能通过计算探明系统中的关键控制因素，为自下而上理性设计人工细胞打下坚实的理论基座。

近年来，PURE系统研究沿着“关键组分合成与再生—蛋白质组系统重构—遗传物质与细胞边界协同重构”的模块级重构路线持续推进，为构建具有自维持和自复制能力的合成细胞奠定了重要基础。首先，T7 RNA聚合酶[7]、多顺反子[8]、氨酰-tRNA合成酶（aminoacyl-tRNA Synthetase，aaRS）[9]等关键组分的合成与再生显著增强了中心法则蛋白组分的自我维持能力。在关键组分研究基础上，2026年，Mottaghi等人进一步将PURE体系推向了蛋白质组系统重构。通过同步表达36种非核糖体蛋白，重构出具有功能活性的二代PURE系统[10]。同时，PURE研究还拓展至遗传物质与细胞边界的协同重构。2026年，Restrepo等人在人工脂质体微反应器中，深度集成了φ29噬菌体DNA复制模块与大肠杆菌甘油磷脂合成途径，实现了基因表达驱动的DNA自复制与膜脂合成的协同运行[11]。Yamagishi等人通过表达26种大肠杆菌染色体复制蛋白，重构出环状染色体复制系统，实现累计约28次的倍增，突破大片段DNA体外复制的规模限制[12]。

然而，真实的生命过程并非模块级合成与再生，而是由微观分子碰撞、结合、解离驱动高维质量作用动力学网络流组成的长时程运行的完整中心法则系统[13，14]。目前相关系统的研究仍面临核糖体合成与自组装效率低[15]、能量耗散[14]、动态反馈与资源协调机制缺乏[16-18]等瓶颈，极易引发不完整肽链堆积及多级反应失衡。因此，亟须基于微观物理化学细节建立高精度数字孪生计算模型，统一解析化学反应网络（Chemical Reaction Network， CRN）模块间的通量调控、资源分配和物质循环，从而为PURE系统的长时程稳定运行提供设计依据。对此，研究者在构建时间显式积分的动力学模型上展开了深度探索。

早期研究将CRN网络的转录翻译主干简化为少数集总步骤，通过极简动态方程实现了快速数值求解，但粗粒化处理剥离了微观物理化学本质，无法刻画因子循环等细节[19]。之后，CRN网络朝局域显式建模的方向展开，如Nieß等人的混合动力学模型[20]、Matsuura等人的局域翻译网络[21]和Horvath等人的中心碳代谢与外源蛋白质偶联网络[22]。但这些研究仅实现了特定基因序列下局部路径的细致刻画，未能实现“转录-翻译”的全链条贯通。同时为了计算收敛，网络的质量作用动力学模拟需严格依赖刚性ODE求解器，不仅计算开销巨大、而且难以处理高维、复杂网络[22]。为了实现PURE系统的全链条高精度模拟，2026年，Jurado等人利用PURE系统组分明确的优势，将Matsuura等人手工编写的翻译网络，改造成能根据任意DNA序列自动生成，并与逐核苷酸转录途径共享资源的CRN网络，实现了转录-翻译全链条的精细质量作用动力学建模[18]。然而，完全的质量作用动力学模拟使得Jurado等人的网络受限于封闭耗散过程。由于真实生命与外界持续的物质交换使得生命网络呈现短寿命中间态形成/消耗、能量分子快速周转和大分子慢速合成等过程的叠加状态[23，24]，巨大的数值刚性使得Jurado等人的方法难以扩展为连续PURE体系的数字孪生，难以准确追踪跨尺度、深耦合的复杂生命网络的宏观表型、微观通量和稳态边界。

为了突破传统动力学模型在连续PURE体系长时程数字孪生中的数值刚性瓶颈，人工智能辅助求解的智能计算模型愈发兴起。早期研究依赖纯数据驱动，缺乏反应机理及物理定律的显式约束，往往产生违反元素守恒的非物理解[25，26]。针对此局限，物理信息神经网络（Physics-Informed Neural Network， PINN）应运而生。通过将基础物理定律和先验机制嵌入神经网络的损失函数中，赋予模型更强的物理可解释性[27-29]。2022年，De Florio等人通过浓度初值和动力学方程残差约束，提升了复杂刚性反应系统的模拟精度，在罗伯逊刚性化学动力学问题上，PINN的浓度预测绝对误差降至10^{-15}–10^{-5}量级[30]。2026年，Darabi等人针对基因回路资源共享引起的上下文效应，提出嵌

入式PINN框架，通过翻译推进等动力学方程和代谢物监测浓度约束，在包含未知动力学的部分基因回路中，重构了部分隐藏中间体并提升了刚性CRN网络模拟的可解释性[28]。然而，由于连续PURE体系的精细CRN网络存在大量难以观测的隐藏中间体和隐藏反应通量，目前的PINN模型难以将隐藏状态合理纳入物理约束，可能造成预测偏差[31，32]。

针对上述传统动力学求解在连续PURE体系长时程模拟中的数值刚性计算瓶颈，以及纯数据驱动和PINN方法在中间态机理约束方面的不足，本研究基于物理约束学习方法构建了基元反应精细度的PINN驱动的连续PURE系统数字孪生模型。通过建立与分析相对精细、经机制验证的降阶教师模型，从连续PURE体系中系统提取长时程浓度轨迹、关键微观中间态等多层级关键特征、评价指标与约束条件。并显式嵌入质量作用动力学、化学计量守恒和开放边界交换条件，合理推断并耦合了传统实验难以直接观测的微观中间态浓度与反应通量，精准重构了基元反应精细度下连续PURE体系代谢物浓度的动态轨迹。为自下而上探究和理性设计“合成生命体”提供了可计算的数字孪生平台。本研究的代码链接为：https://github.com/Chen-0043/PINN-CRN-PURE。

2 教师模型浓度监督轨迹的模拟

2.1 间歇式教师模型的模拟与准确性验证

针对传统动力学方法在连续PURE体系模拟中的数值刚性瓶颈，以及纯数据驱动和PINN方法因缺乏充分的物理约束导致预测偏差和生物学可解释性差的问题，以deGFP合成体系[33]和Pgi、PfkA、TpiA、GAPDH、Pgk、Eno 6蛋白共表达体系[34]为例，本研究构建了逐核苷酸转录和逐氨基酸翻译的分层降阶教师模型，从而为高维PINN模型提供可靠的物理先验知识与监督来源，如图1（a）所示。

教师模型的构建过程中，通过中心法则实现了已知蛋白质序列向转录-翻译PURE代谢网络的映射。对于特定蛋白质片段，以文献实验结果和数据库为DNA编码序列的主要来源；对于缺少DNA序列的蛋白质片段，采用*E. coli*偏好密码子映射进行反向翻译。在DNA编码序列推断结果基

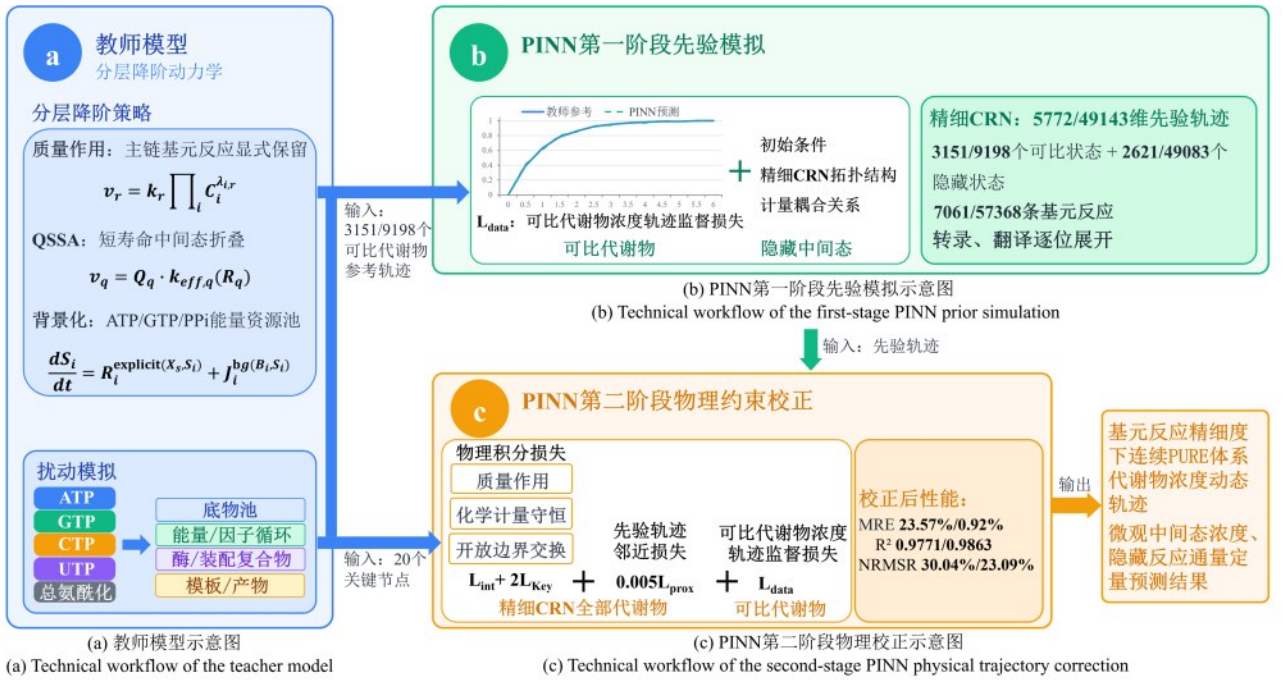


图1 PINN驱动的连续PURE系统数字孪生模型技术路线图

Figure 1 Technical workflow of the PINN-driven digital twin model of a continuous PURE system

础上，按照DNA－mRNA－蛋白质的信息传递关系建立代谢网络拓扑结构。从而为deGFP合成和6蛋白共表达体系分别构建了包含3151和9198个代谢物的教师模型。

构建的教师模型反应模块划分、初始投料浓度、连续补料参数见附表1。具体而言，deGFP单蛋白体系的教师模型主要保留转录、氨酰化和翻译因子循环的基元质量作用反应；将翻译推进步的aa-tRNA递送、肽键形成和核糖体移位等步骤降阶为相邻肽链状态间的一步QSSA反应；并对能量再生、小分子缓冲、NTP池转移及连续补料/流出过程进行背景化处理。由于多蛋白表达通路对能量等组分的共利用，为了降低数值刚性，本研究进一步将Pgi-Eno的6蛋白共表达教师模型的转录过程中逐核苷酸位点的NTP结合、核苷酸加成和PPi释放降阶为一步QSSA反应，其余过程与deGFP基本一致。通过保留逐核苷酸和氨基酸的推进过程，本研究可以直接追踪转录翻译推进状态与资源消耗，并建立教师模型与后续精细CRN之间的序列位置对应。

在教师模型构建结果基础上，本研究通过模型动力学参数的训练优化，验证了模型的模拟精度，模拟结果如图2（a）和（b）所示。

由图2（a）和（b）发现，与文献值相比，模型对deGFP单蛋白与6蛋白体系的平均相对预测误差（Mean Relative Error， MRE）分别为10.85%和18.88%，且模型能够同时再现蛋白质累积平台和由快速上升向平台过渡的总体动态。这说明教师模型在间歇式体系中已经具备了较高的数值模拟精度，能够在不同表达复杂度下提供具有定量依据的浓度参考，进而为连续体系和扰动模拟的扩展提供较为可靠的转录、翻译、能量再生和蛋白质形成等动力学参数。

2.2 连续体系的拟稳态模拟与PINN模型代谢物浓度参考轨迹提取

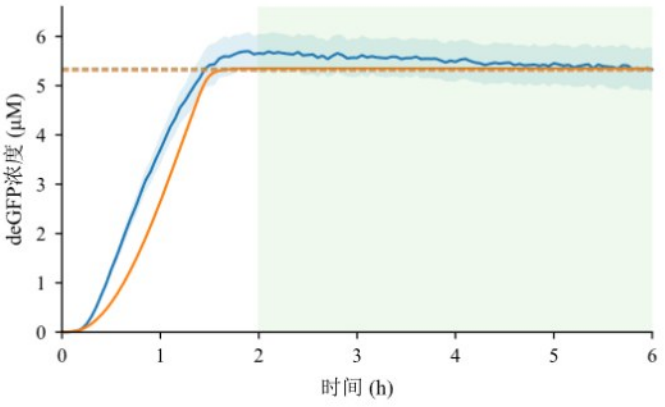
由于自复制生命的维持依赖于开放体系中的非平衡动态稳态，因此，在保持与间歇式验证的动力学参数和反应网络一致的基础上，本研究进一步将教师模型扩展到连续体系。通过增加补料、流出和稀释项，来进行给定开放边界条件的连续式机制外推。计算了交换速率γ为0.50 h^{-1}情况下的长时程运行特征，如图2（c）和（d）所示。

由图2（c）和（d）发现，降阶模型能够稳定重现连续PURE体系长时程运行的稳态特征，48～72 h区间内deGFP和6蛋白浓度的最大漂移均不超过1.57×10^{-3}μM。这说明物质输入、反应消耗和体系洗出能够在教师模型中达到数值稳定的动态平衡。也为deGFP合成体系和6蛋白共表达体系的PINN模型可比代谢物浓度轨迹监督损失L_{data}（衡量PINN相对于教师模型的可比代谢物浓度轨迹总体偏离程度的指标，见附件1，第2节）和整体物理积分损失L_{int}（衡量PINN模拟的可比代谢物浓度轨迹在质量作用动力学、化学计量守恒和开放边界交换约束下的总体物理损失，见附件1，第4节）训练分别提供了3151条和9198条较为可靠的参考轨迹输入与坚实的动力学基础。值得注意的是，为保证教师模型合理性，本研究额外利用Pgi-Eno的6个单蛋白表达体系进行序列特异性参数校准，并将已校准的表达通路整合为包含9198个代谢物的六蛋白共表达教师模型用于后续的PINN模拟分析。而对于Pgi-Eno的6个单蛋白体系，本研究暂未进行PINN模拟分析。

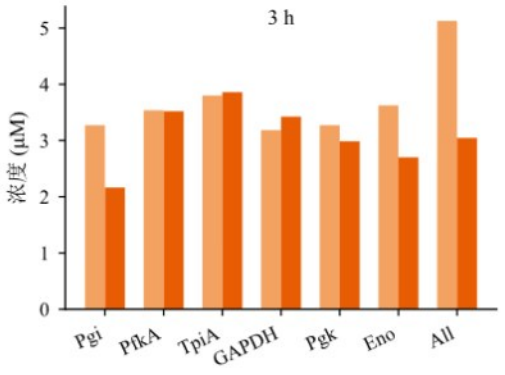
3 教师模型扰动模拟与PINN提取的关键机制约束信息

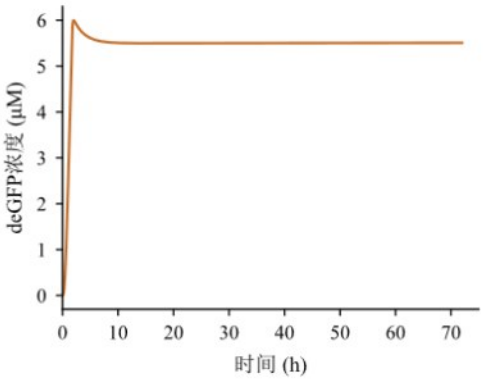
3.1 教师模型扰动模拟

为识别自复制表达子系统中影响资源分配、组分再生和长期稳定性的关键因素，并为PINN初步筛选关键特征与关键约束，本研究在deGFP教师模型连续模拟的基础上进一步进行了扰动分析，如图1（a）所示。鉴于多蛋白表达对能量分配、核糖体装载和翻译因子循环等干扰，为保证模型合理性，本研究未对6蛋白体系进行扰动模拟，而是将deGFP的扰动分析的关键特征与关键约束应用于6蛋白体系。根据连接的反应数、同一反应中的邻域代谢物数、覆盖参与的机制模块数和连接的QSSA与背景化的反应数4个指标进行评分，并根据评分选取可进行投料和连续补料的代谢物/模块进行扰动模拟。评分结果见附表2，最终选择了

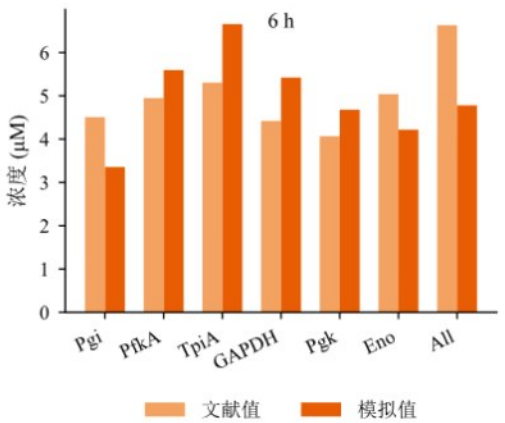


(a) deGFP单蛋白间歇体系的教师模型模拟值与文献值
(a) Simulated and literature values for the deGFP protein batch synthesis system

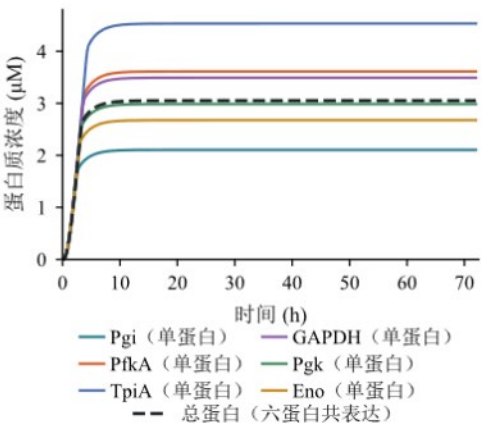




(c) deGFP单蛋白连续体系的教师模型模拟结果
(c) Simulation results for the deGFP protein continuous synthesis system



(b) 3 h与6 h时间点的6蛋白间歇体系的教师模型模拟值与文献值（注：Pgi-Eno表示6个单表达体系，All表示共表达体系）
(b) Simulated and literature values for the six-protein batch system synthesis at 3 h and 6 h (Note: Pgi–Eno denote the six single-protein expression systems and All denotes the co-expression system)



(d) 6蛋白连续体系的教师模型模拟结果（注：实线表示单表达体系，虚线表示共表达体系）
(d) Simulation results for the six-protein continuous synthesis system (Note: solid lines represent the single-protein expression systems and the dashed line represents the co-expression system).

图 2 教师模型的模拟结果

Figure 2 Simulation results of the teacher model

兼具转录底物与翻译供能支撑作用的ATP、GTP，作为转录原料的CTP、UTP，以及由总aaRS与氨基酸供给构成的总氨酰化模块进行扰动模拟。这三类扰动分别对应模板形成与能量分配、转录负荷和aa-tRNA供给，覆盖了蛋白质合成由核酸合成到肽链延伸的主要资源传递环节。

以Jurado等人的deGFP生产体系的初始投料浓度做为反应腔初始浓度和外部交换容器的浓度基线，对参与扰动的底物/模块分别进行0、0.01、0.03、0.06、0.1、0.22、0.25、0.5、1、2、3、4倍的浓度扰动，根据已经达到拟稳态的6h deGFP浓度（见附表2）确定最优浓度倍数，发现在当前模型参数和边界条件下，ATP、GTP、CTP、UTP和总氨酰化模块的最优浓度倍数分别为0.5、3、0.01、0.01和0.22。然后分别计算0倍浓度和相邻过最优浓度下，相比于最优浓度点的教师模型全代谢网络尺度的相对代谢物浓度和相对通量，关键结果如图3所示。

由图3发现，所有扰动均呈现出由供给不足到动态适配再到过量错配的共同变化规律。图3（a）模拟结果显示，低于适配区间时，各扰动均表现为蛋白质合成关键链路未能贯通的断链型错配，但具体断点有所不同：ATP或GTP不足同时限制能量供给及mRNA形成过程，CTP或UTP不足阻碍完整mRNA的形成，总氨酰化模块不足则限制aa-tRNA供给和肽链持续延伸。随着供给增加，mRNA、核糖体-mRNA复合物、翻译位置链推进、翻译因子循环和终止释放逐步衔接，系统进入资源投入能够较有效转化为deGFP生成通量的适配状态。

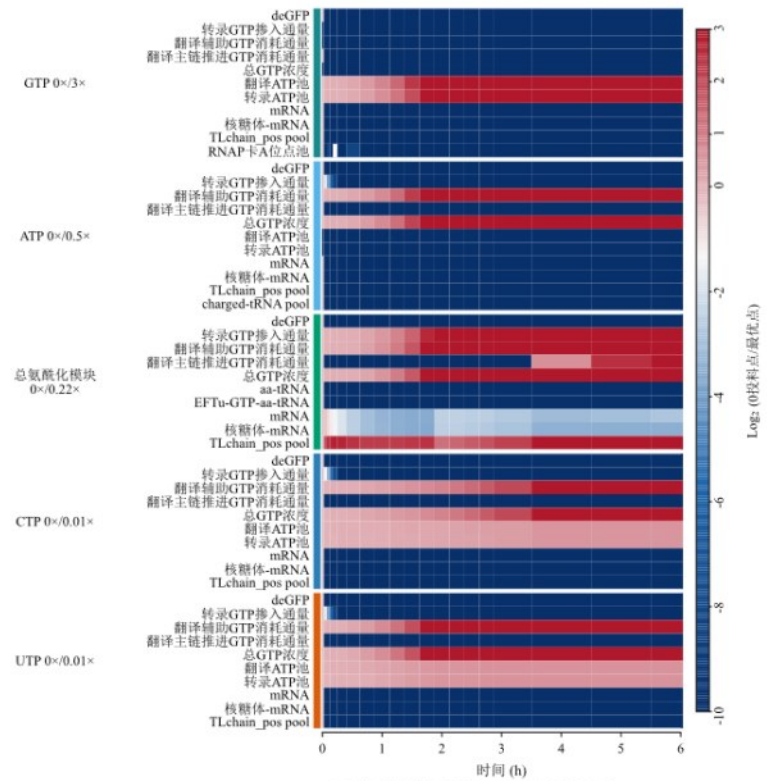
图3（b）模拟结果显示，超过适配区间后，扰动结果进一步表现出两类错配机制。GTP过量时，转录延伸GTP掺入及翻译过程的GTP相关因子循环被同时增强，继而透支ATP，使含ATP位点的转录推进受阻；导致未成熟的转录中间态积累；另一方面，ATP相关的氨酰化供给、能量支撑和翻译链承接能力同步减弱，使翻译延伸与终止释放过程放缓。除GTP外，ATP、CTP/UTP和总氨酰化模块在超过各自适配点后则呈现相似的错配机制：大量供给首先推动mRNA模板形成或aa-tRNA等供给，使翻译主链推进短暂增强；增强的主链随后占用大量GTP，导致转录过程和翻译辅助过程因GTP可用资源下降而受阻。进而使mRNA或核糖体-mRNA复合物逐渐成为承接瓶颈，翻译主链推进回落。随着主链GTP消耗减弱，转录和翻译辅助过程的GTP可用资源回升，转录和翻译辅助过程也逐步回升，系统最终进入新平衡。这说明，过量供给造成的并非资源总量不足，而是资源在转录、翻译辅助和翻译主链之间的时序分配失衡。

综合上述结果，连续PURE体系的蛋白质生成由转录底物供给、ATP/GTP能量状态、氨酰化供给、核糖体装载和翻译因子循环之间的动态适配共同决定。对于需要持续再生功能蛋白的PURE系统，控制目标应当是避免任何资源轴过弱造成断链，或过强造成自身过推和承接不足。

3.2 PINN模型提取关键机制约束信息

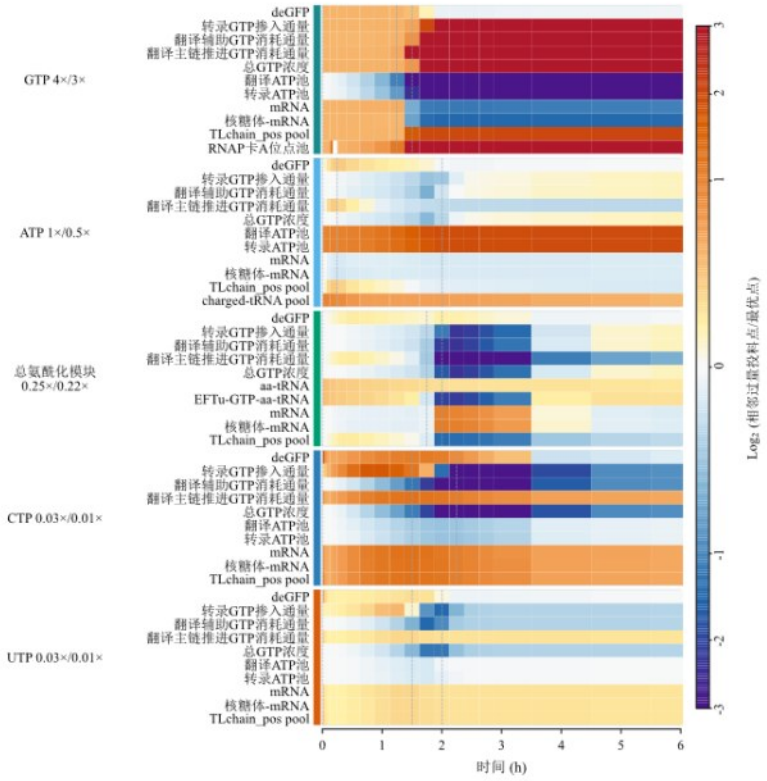
基于deGFP教师模型扰动模拟结果，本研究将识别出的关键机制信息分别转化为关键代谢物约束。以GTP、ATP、CTP、UTP和总aaRS供给能力五类入口扰动为基础，追踪其在教师模型中牵动的下游瓶颈过程。结果显示，这些扰动主要影响转录/翻译GTP参与、转录ATP参与/能量循环、副产物累积、转录产物/翻译模板形成、核糖体装载、翻译起始与氨酰化、翻译延伸与因子循环以及蛋白质产物形成过程。接下来选取每类过程中的底物、关键循环因子、关键酶复合物、装载复合物和产物，确定了GTP、ATP、PPi、核糖体–mRNA复合物、延伸因子复合物、aaRS和deGFP等20个关键代谢物节点用于PINN关键代谢物增强型物理积分损失L_{key}的输入（衡量PINN模拟的关键代谢物浓度轨迹在物理约束下的总体损失的指标，见附件1，第4节）的训练，关键代谢物节点见附表2。

六蛋白共表达体系则以deGFP体系确定的20个关键代谢物为模板，按照其在能量供给、翻译因子循环、核糖体装载和蛋白产物形成等过程中的作用，映射至六蛋白模型的对应状态，最终确定31个关键代谢物状态用于构建L_{key}，具体关键代谢物与对应关系见附表2。



(a) GTP、ATP、CTP、UTP和总氨酰化模块零投料对PURE体系的影响

(a) Effects of zero GTP, ATP, CTP, UTP, and total aminoacylation module feeding on the PURE system



(b) GTP、ATP、CTP、UTP和总氨酰化模块过量投料对PURE体系的影响

(b) Effects of excessive GTP, ATP, CTP, UTP, and total aminoacylation module feeding on the PURE system.

图 3 教师模型扰动模拟结果

Figure 3 Perturbation simulation results of the teacher model

4 PINN先验轨迹生成与物理约束校正

4.1 精细CRN网络构建与代谢物先验浓度轨迹的生成

为使数字孪生模型的预测同时反映资源分配与通量调控关系，本研究根据教师模型的网络拓扑结构构建了精细化CRN网络。以QSSA和背景化反应的反应物、产物及净化学计量关系为边界，补回被折叠的底物结合、中间复合物形成、链推进和产物释放等步骤，并将能量再生、核苷酸转换及副产物处理过程展开为相应反应，从而为deGFP合成与6蛋白共表达体系分别构建了包含7061条基元反应/5772个代谢物和57368条基元反应/49143个代谢物的精细化CRN网络。

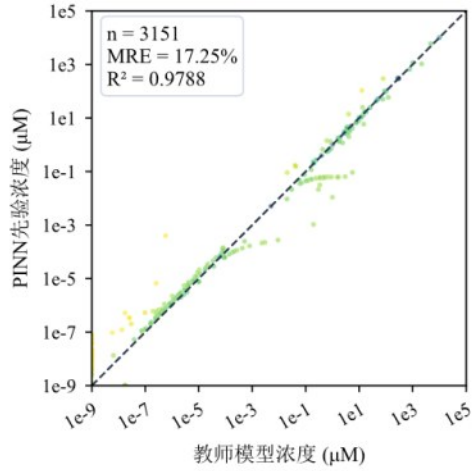
在精细化CRN网络构建基础上，利用教师模型提供的可比代谢物浓度轨迹，根据L_{data}对PINN模型进行第一阶段训练和评价，生成精细CRN网络的先验代谢物浓度轨迹。如图1（b）所示，通过离散数据点进行训练（注意：本研究的deGFP合成体系和6蛋白共表达体系的训练数据集分别覆盖0-5.00 h和0-5.50 h）；6 h浓度作为验证集。结合构建的精细CRN网络（见附表3）生成并输出deGFP单蛋白/6蛋白共表达体系的5772/49143维代谢物先验浓度轨迹。其中，3151/9198个可比代谢物具有教师模型浓度参考并通过L_{data}监督，其余代谢物作为精细CRN网络中的隐藏代谢物，由初始条件、网络拓扑及代谢物之间的计量耦合关系共同推断。先验模拟结果如图4所示。

图4（a）和（c）模拟结果显示，先验轨迹能够较好地逼近教师模型的浓度分布。在验证集中，deGFP和6蛋白共表达体系的可比代谢物浓度预测的MRE分别为17.25%和0.02%，决定系数（Coefficient of determination，R^2）分别为0.9788和0.9995。说明浓度点监督能够形成较为合理的代谢物状态模拟先验。然而，图4（b）和（d）模拟结果显示，可比代谢物的轨迹尚未充分符合精细CRN网络物理约束的动力学方向，物理积分变化均方根s_i^{rhs}>0 μM（衡量代谢物浓度轨迹的变化幅度的指标，见附件1，第4节）的代谢物浓度轨迹归一化均方根残差（Normalized Root Mean Square Residual of trajectory simulation，NRMSR，衡量任意参与评价的代谢物浓度轨迹在物理约束下的总体损失的指标，见附件1，第5节）分别为186.21%和101.71%。为避免预测积分变化接近零时相对残差被分母过度放大，进一步选取s_i^{rhs}>1 μM的轨迹进行评价，NRMSR分别为105.90%和102.77%。这说明浓度点仍不足以保证预测轨迹沿质量作用动力学、化学计量守恒和开放边界交换共同规定的物理轨迹演化。

4.2 物理轨迹校正

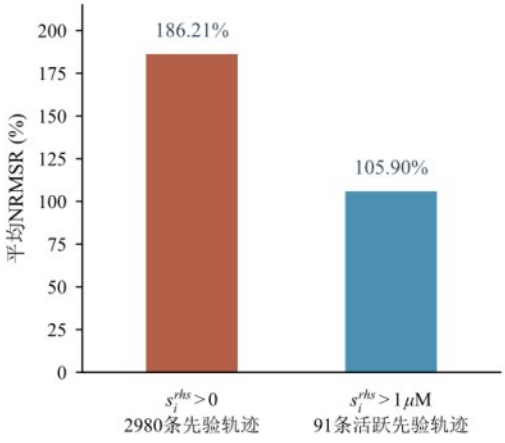
为了使PINN预测结果在保持浓度拟合精度的同时，尽可能靠近精细CRN网络的质量作用动力学、化学计量守恒和开放边界交换等物理约束的动力学方向。在获得全代谢物先验浓度轨迹后，本研究进行了第二阶段物理轨迹校正。将神经网络第一阶段输出的全状态先验轨迹转换为离散的时间—浓度数据，并直接校正各时间点的浓度数值，此过程不再将校正结果反馈至神经网络。如图1（c）所示，第二阶段将deGFP单蛋白和6蛋白共表达体系各时间节点的5772和49143维预测浓度代入包含7061和57368条基元反应的精细CRN网络，计算反应通量、化学计量净变化和开放边界交换贡献，并在相邻时间窗口内计算各代谢物的物理约束偏离残差，用来反映PINN预测的单一代谢物浓度轨迹与精细CRN网络物理约束计算的浓度轨迹之间的偏离程度（见附件1，第4节）。

在此基础上，以deGFP单蛋白和6蛋白共表达体系的3151和9198个可比代谢物的0-6 h $e_{m,\,i}$构建整体物理积分损失L_{int}，并对扰动分析确定的关键代谢物额外构建关键代谢物增强型物理积分损失L_{key}。L_{int}和L_{key}在反馈可比代谢物状态的同时通过化学计量耦合关系和基元反应通量将当前状态逐层传导至隐藏代谢物，从而校正精细CRN的全部代谢物浓度（见附件1，第1、4节）。同时，设置轨迹邻近损失L_{prox}来衡量校正后的0-6 h全代谢物浓度远离上一段优化状态浓度的指标（见附件1，第4节）。在此基础上，通过0-5.50 h覆盖时间点的L_{data}进一步对可比代谢物施加教师浓度轨迹的直接监督。第二阶段轨迹校正中，L_{data}、L_{int}、L_{key}和L_{prox}



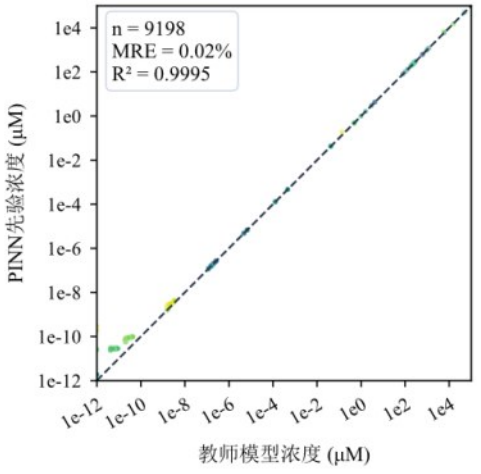
(a) deGFP合成体系验证集可比代谢物浓度PINN先验预测值与教师模型浓度

(a) PINN-predicted prior concentrations and teacher-model concentrations of comparable metabolites in the validation set for the deGFP protein synthesis system



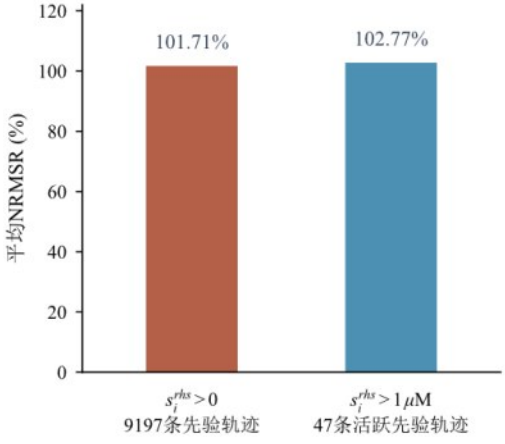
(b) deGFP合成体系先验轨迹NRMSR

(b) NRMSR of the prior trajectories for the deGFP protein synthesis system



(c) 6蛋白共表达体系验证集可比代谢物浓度PINN先验预测值与教师模型浓度

(c) PINN-predicted prior concentrations and teacher-model concentrations of comparable metabolites in the validation set for the six-protein co-expression system



(d) 6蛋白共表达体系先验轨迹NRMSR

(d) NRMSR of the prior trajectories for the six-protein co-expression system.

图4 PINN先验模拟结果

Figure 4 Prior simulation results of the PINN

的权重分别设为1、1、2和0.005。该设置使关键代谢物在整体物理约束基础上获得二倍的附加校正。L_{prox}采用较小权重，用于限制第二阶段校正发生无必要的大幅偏移，从而在改善物理一致性的同时保持轨迹校正的连续性和数值稳定性。

轨迹校正结束时，各损失项分别为：

deGFP单蛋白体系：

$$L_{data} = 3.86 \times 10^{-2}\ L_{int} = 18.91$$
$$L_{key} = 0.30\ L_{prox} = 2.90 \times 10^{-4}$$

按照上述权重计算得到的总损失为：

$$L_{total} = L_{data} + L_{int} + 2L_{key} + 0.005L_{prox} = 19.56$$

6蛋白共表达体系：

$$L_{data} = 1.14 \times 10^{-2}\ L_{int} = 5.47 \times 10^{-2}$$
$$L_{key} = 0.22\ L_{prox} = 2.00 \times 10^{-6}$$

按照上述权重计算得到的总损失为：

$$L_{total} = L_{data} + L_{int} + 2L_{key} + 0.005L_{prox} \approx 0.50$$

从四项损失的构成来看，两个体系在轨迹校正结束时的剩余偏差来源有所不同。deGFP单蛋白

体系中，L_{int}高达18.91，约占总损失的96.70%，说明该体系第二阶段校正的主要任务是提高全体可比代谢物浓度轨迹与质量作用动力学、化学计量守恒和开放边界交换共同规定的物理积分轨迹之间的一致性；L_{key}较低，说明在整体物理约束之外，关键代谢物增强校正能够将扰动分析识别出的瓶颈节点维持在较低的物理约束偏离水平。由于这些节点集中反映能量循环、模板形成、核糖体装载、翻译因子循环和氨酰化等持续蛋白质合成过程中的关键控制环节，其物理一致性的提高有助于保证模型对连续PURE体系核心运行状态的可解释预测。

6蛋白共表达体系中，L_{int}降至5.47×10^{-2}，在四项基础损失中的占比仅约10.86%，说明全体可比代谢物轨迹的整体物理积分偏差已得到较大程度压低。相比之下，L_{key}占比大幅增大，$2L_{key}$达到了总损失的86.87%，说明剩余的物理约束偏差主要集中于关键代谢物节点。这可能由多条蛋白表达通路对能量、核苷酸、核糖体及翻译因子等组分的共同利用形成的强耦合特征导致。

两个体系的L_{data}均保持在较低水平，表明物理轨迹校正没有使可比状态大幅偏离教师模型的浓度参考；极低的L_{prox}说明校正状态的参数处于变化极小的区间。因此，第二阶段校正在合理保留先验信息和教师模型监督信息的同时，提高了输出结果与物理约束之间的一致性。

4.3 耦合微观中间态浓度—隐藏反应通量的物理约束轨迹一致性评价

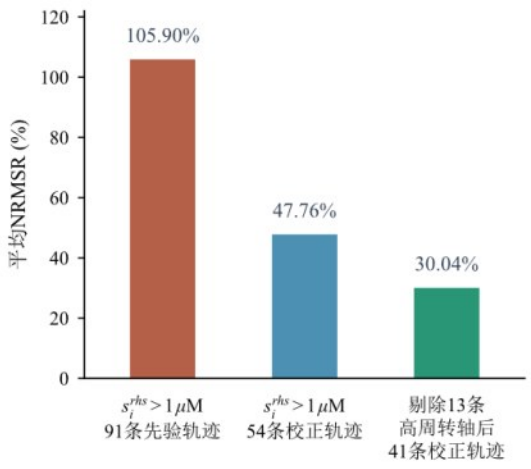
经第二阶段轨迹物理约束校正，PINN输出结果如图5所示。图5（a）和（c）模拟结果显示，deGFP体系和6蛋白共表达体系中s_i^{rhs}>1 μM的轨迹数目分别从91、47条降至54、34条，轨迹NRMSR分别降为47.76%和23.09%，表明物理轨迹校正显著降低了PINN模拟与精细CRN网络物理约束动力学方向之间的偏离，但对于deGFP体系，全部代谢物轨迹的动力学重现程度仍偏低。

进一步对deGFP体系各轨迹的误差分布及其反应组成进行检查发现，较大的误差主要集中于GTP核苷酸循环轨迹、mRNA中间态轨迹、能量再生轨迹以及部分aaRS的氨基酸活化中间态轨迹等13条代谢物轨迹。这些代谢物直接参与核苷酸转换、能量再生或酶促中间态转换，其浓度变化对局部反应通量偏差较为敏感。因此，本研究进一步进行了分层统计：除去这13条轨迹后，其余41条轨迹的NRMSR为30.04%。结果表明，deGFP体系的物理偏差主要集中在快速生成、快速消耗和连续交换同时存在的高周转资源轴，而不是均匀分布于整个CRN网络。

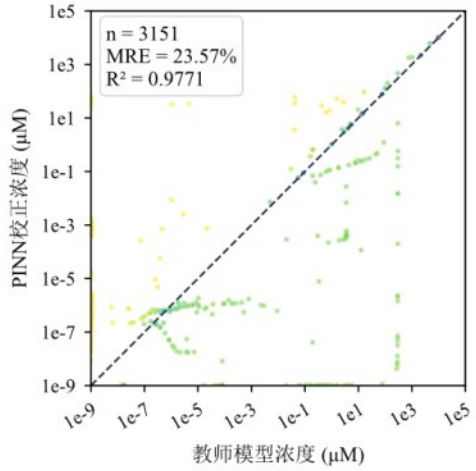
图5（b）和（d）模拟结果显示，验证集中，deGFP体系和6蛋白共表达体系可比代谢物浓度预测的MRE分别为23.57%和0.92%，R^2分别为0.9771和0.9863。与物理校正前相比，模型允许浓度误差适度增加，以换取预测轨迹与精细CRN网络物理约束动力学方向的一致性，这体现了浓度重构精度与物理轨迹约束之间的平衡。

上述轨迹模拟误差与浓度预测误差并非仅根据可比代谢物自身的浓度变化计算。在先验轨迹模拟结束后，模型首先将精细CRN网络中包括微观中间态在内的全代谢物预测浓度代入质量作用动力学算子，计算相应的隐藏基元反应通量。再通过化学计量守恒和开放边界交换汇总得到可比代谢物在相邻时间窗口内的理论积分变化$R_i(t)$，并据此构建$e_{m,i}$，对全代谢物轨迹进行第二阶段校正。此过程同时反映了微观中间态浓度、隐藏反应通量与可比代谢物浓度轨迹之间的联合一致性，使原本不可由教师模型直接评估的通路获得了可量化的指标。

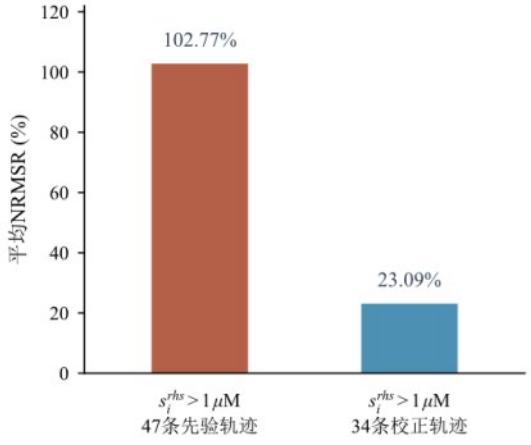
通过分析微观中间态轨迹，发现在deGFP体系中，PINN对于IF1/IF3介导的翻译起始装配模块（教师模型的QSSA子模块）和总能量再生模块（教师模型的背景化子模块）的微观中间态轨迹取得了物理约束动力学较为一致的结果：NRMSR分别为28.11%和54.66%。在6蛋白共表达体系中，CTP和UTP转录底物掺入子模块的NRMSR分别为32.96%和35.19%。结合可比代谢物浓度预测的精度，这些结果说明，模型不仅通过微观中间态浓度和隐藏反应通量的推断在整体上较为合理地支撑可比代谢物的动态演化，还能够在不依赖短寿命中间态和快速循环小分子的刚性ODE长时间积分求解的前提下，合理恢复教师模型的刚性反应通路中由基元质量作用动力学解释的隐藏中间态



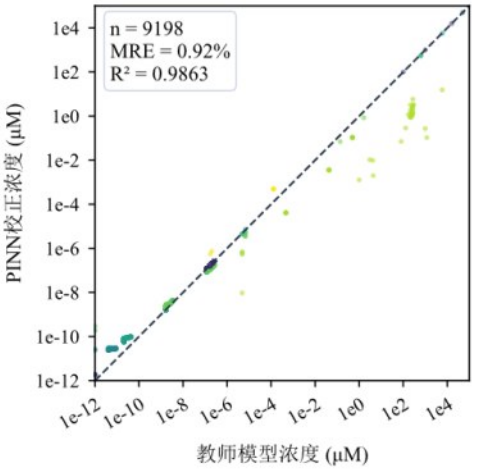
(a) deGFP合成体系s_i^{rhs}>1 μM 的校正轨迹NRMSR
(a) NRMSR of trajectories with $s_i^{rhs} > 1$ μM for the deGFP protein synthesis system after physical trajectory correction



(b) 校正后deGFP合成体系验证集可比代谢物浓度PINN预测值与教师模型浓度
(b) PINN-predicted concentrations and teacher-model concentrations of comparable metabolites in the validation set for the deGFP protein synthesis system after physical trajectory correction



(c) 6蛋白共表达体系s_i^{rhs}>1 μM 的校正轨迹NRMSR
(c) NRMSR of trajectories with $s_i^{rhs} > 1$ μM for the six-protein co-expression system after physical trajectory correction



(d) 校正后6蛋白共表达体系验证集可比代谢物浓度PINN预测值与教师模型浓度
(d) PINN-predicted concentrations and teacher-model concentrations of comparable metabolites in the validation set for the six-protein co-expression system after physical trajectory correction.

图5 物理轨迹校正后的PINN模拟结果

Figure 5 PINN simulation results after physical trajectory correction

轨迹，为基元反应精细度下内部状态与反应通量的机制化解析提供了依据。

本研究通过神经网络先验轨迹生成与物理约束校正，显著缓解了传统刚性ODE直接求解完整精细CRN时对高维稀疏线性系统进行长时程逐步积分的计算瓶颈（见附表4），并一定程度解决了纯数据驱动人工智能模型对中间态机理约束方面不足的问题。实现了基元反应精细度下连续PURE体系代谢物浓度的动态轨迹的精准重构。

5 总 结

建立高精度PURE系统数字孪生模型，是深入解析合成细胞物质循环与通量调控，理性设计自复制合成生命的重要路径。针对传统动力学模型

在连续体系长时程模拟中的数值刚性计算瓶颈，以及纯数据驱动和PINN模型在中间态机理约束方面的不足，本研究基于物理约束学习方法，构建了PINN驱动的连续PURE系统数字孪生模型，实现了基元反应精细度下连续体系代谢物浓度动态轨迹的精准重构。模型通过分层降阶教师模型提供浓度监督轨迹和关键节点约束信息，通过精细CRN网络的基元反应提供质量作用动力学、化学计量守恒和开放边界交换约束，形成了“整体浓度轨迹—关键节点—物理约束关系”的一体化建模框架。结果表明，教师模型能够为连续PURE体系提供可靠的长时程参考轨迹，扰动分析进一步识别了影响蛋白质合成的能量循环、核糖体装载、翻译因子循环等关键代谢物节点。在此基础上，PINN模型通过“神经网络状态先验—物理轨迹校正”的两阶段流程，保障了可比代谢物浓度预测合理性和模拟轨迹与物理约束动力学方向之间的一致性。验证集中代谢物浓度预测的MRE低至0.92%，R^2高达0.9863，可比代谢物中高通量轨迹的NRMSR低至23.09%。更重要的是，模型将微观中间态浓度、隐藏基元反应通量和可比代谢物浓度变化联合起来，使教师模型中因QSSA和背景化处理而不可直接评估的隐藏状态获得了可量化的物理轨迹指标。QSSA子模块中IF1/IF3介导的翻译起始装配模块、CTP、UTP转录底物掺入子模块和背景化子模块中总能量再生模块的微观中间态轨迹取得了物理约束动力学较为一致的结果：NRMSR分别为28.11%、32.96%、35.19%和54.66%。本模型赋予了PURE系统在基元反应精细度下高分辨率预测隐蔽中间态的能力，为自下而上理性设计全链条可控、稳定自维持的合成细胞提供了可计算的物理底座与分析工具。

致谢：本论文得到国家重点研发计划（2025YFA0922800）、中国科学院战略性先导科技专项（XDB0480000）、中国自然科学基金（12501699）和数学科学全国重点实验室开放课题合作团队项目（SLMS-2025-KFKT-TD-01）的资助。

参 考 文 献

[1] SNOECK S, GUIDI C, DE MEY M. "Metabolic burden" explained: stress symptoms and its related responses induced by (over)expression of (heterologous) proteins in Escherichia coli [J]. Microbial Cell Factories, 2024, 23(1): 96.

[2] GUIDRY J, BOWMAN G R. Programmable cell differentiation in budding yeast uncouples reproductive and metabolic tasks [J]. bioRxiv preprint, 2025, doi: https://doi.org/10.64898/2025.12.04.692473.

[3] ROTHSCHILD L J, AVERESCH N J H, STRYCHALSKI E A, et al. Building Synthetic Cells — From the Technology Infrastructure to Cellular Entities [J]. ACS Synthetic Biology, 2024, 13(4): 974-97.

[4] HOFMANN M, ABDO F, BORKOWSKI O, et al. Reconstituting alternative life using the test-bed of cell-free systems [J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2025, 380(1936): 20240293.

[5] LIAO C, LI H, ZHENG H, et al. Optimization and Application of the Purified Cell-Free System [J]. ChemBioChem, 2025, 26 (11): e202500176.

[6] HUNT A C, RASOR B J, SEKI K, et al. Cell-Free Gene Expression: Methods and Applications [J]. Chemical Reviews, 2024, 125(1): 91-149.

[7] LAVICKOVA B, LAOHAKUNAKORN N, MAERKL S J. A partially self-regenerating synthetic cell [J]. Nature Communications, 2020, 11(1): 6340.

[8] LIBICHER K, HORNBERGER R, HEYMANN M, et al. In vitro self-replication and multicistronic expression of large synthetic genomes [J]. Nature Communications, 2020, 11 (1): 904.

[9] HAGINO K, ICHIHASHI N. In Vitro Transcription/Translation-Coupled DNA Replication through Partial Regeneration of 20 Aminoacyl-tRNA Synthetases [J]. ACS Synthetic Biology, 2023, 12(4): 1252-63.

[10] MOTTAGHI S S, MAERKL S J. PURE makes PURE: reconstitution of the PURE cell-free system from self-synthesized proteins [J]. Nature Communications, 2026, 17(1): 6756.

[11] RESTREPO SIERRA A M, RAMIREZ GOMEZ F, VAN TONGEREN M, et al. A synthetic cell with integrated DNA self-replication and lipid biosynthesis [J]. Nature Communications, 2026, 17(1): 2727.

[12] YAMAGISHI Y, SONOYAMA Y, KAWAKAMI N, et al. A self-replicating artificial module-genome that generates bacterial chromosome replication system in vitro [J]. Nucleic Acids Research, 2026, 54(13): gkag663

[13] KEARNEY T, FLEGG M B. Particle-based simulation of non-elementary bimolecular kinetics [J]. Mathematical Biosciences, 2026, 391: 109582.

[14] JURADO Z, PANDEY A, MURRAY R M. Nucleotide-Level Chemical Reaction Network Modeling Enables Quantitative Prediction of Reconstituted Cell-Free Expression Systems [J]. ACS Synthetic Biology, 2026, 15(7): 3005-16.

[15] KOSAKA Y, MIYAWAKI Y, MORI M, et al. Autonomous ribosome biogenesis in vitro [J]. Nature Communications, 2025, 16(1): 514.
[16] YADAV S, PERKINS A J P, LIYANAGEDERA S B W, et al. ATP Regeneration from Pyruvate in the PURE System [J]. ACS Synthetic Biology, 2025, 14(1): 247-56.
[17] GIAVERI S, ABIL Z, KOHYAMA S, et al. Building a Synthetic Cell Together [J]. Nature Communications, 2025, 16 (1): 7488.
[18] DE CAPITANI J, MUTSCHLER H. The Long Road to a Synthetic Self-Replicating Central Dogma [J]. Biochemistry, 2023, 62(7): 1221-32.
[19] STöGBAUER T, WINDHAGER L, ZIMMER R, et al. Experiment and mathematical modeling of gene expression dynamics in a cell-free system [J]. Integrative Biology, 2012, 4 (5): 494.
[20] NIEß A, FAILMEZGER J, KUSCHEL M, et al. Experimentally Validated Model Enables Debottlenecking of in Vitro Protein Synthesis and Identifies a Control Shift under in Vivo Conditions [J]. ACS Synthetic Biology, 2017, 6(10): 1913-21.
[21] MATSUURA T, TANIMURA N, HOSODA K, et al. Reaction dynamics analysis of a reconstituted Escherichia coli protein translation system by computational modeling [J]. Proceedings of the National Academy of Sciences, 2017, 114(8): E1336 – E44.
[22] HORVATH N, VILKHOVOY M, WAYMAN J A, et al. Toward a genome scale sequence specific dynamic model of cell-free protein synthesis in Escherichia coli [J]. Metabolic Engineering Communications, 2020, 10: e00113.
[23] JI W, QIU W, SHI Z, et al. Stiff-PINN: Physics-Informed Neural Network for Stiff Chemical Kinetics [J]. The Journal of Physical Chemistry A, 2021, 125(36): 8098-106.
[24] HUANG X, KANDRIS K, KATSOU E. Training stiff neural ordinary differential equations in data-driven wastewater process modelling [J]. Journal of Environmental Management, 2025, 373: 123870.
[25] JI W, DENG S. Autonomous Discovery of Unknown Reaction Pathways from Data by Chemical Reaction Neural Network [J]. The Journal of Physical Chemistry A, 2021, 125(4): 1082-92.
[26] DöPPEL F A, VOTSMEIER M. Robust mechanism discovery with atom conserving chemical reaction neural networks [J]. Proceedings of the Combustion Institute, 2024, 40(1-4): 105507.
[27] KISBENEDEK L, DREXLER D A, KOVáCS L A. Physics-informed neural network for parameter identification of a tumor model [J]. IFAC PapersOnLine, 2025, 59(11): 168-73.
[28] DARABI A, AN Z, AL-RADHAWI M A, et al. Combining model-based and data-driven models: An application to synthetic biology resource competition [J]. Mathematical Biosciences, 2026, 396: 109649.
[29] ZHANG S, ZHANG C, WANG B. CRK-PINN: A physics-informed neural network for solving combustion reaction kinetics ordinary differential equations [J]. Combustion and Flame, 2024, 269: 113647.
[30] DE FLORIO M, SCHIASSI E, FURFARO R. Physics-informed neural networks and functional interpolation for stiff chemical kinetics [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2022, 32(6): 063107.
[31] ALAM M N, ANURAG A, GANGWAR N, et al. Physics - informed neural networks guided modelling and multiobjective optimization of a mAb production process [J]. The Canadian Journal of Chemical Engineering, 2024, 103(3): 1319-34.
[32] BAJAJ C, MCLENNAN L, ANDEEN T, et al. Recipes for when physics fails: recovering robust learning of physics informed neural networks [J]. Machine Learning: Science and Technology, 2023, 4(1): 015013.
[33] JURADO Z, PANDEY A, MURRAY R M. A chemical reaction network model of PURE[J]. bioRxiv: 2023.08.14.553301, 2023.
[34] SATO G, MIYAZAWA S, DOI N, et al. Cell-Free Protein Expression by a Reconstituted Transcription – Translation System Energized by Sugar Catabolism [J]. Molecules, 2024, 29(13): 2956.



通讯作者： 马红武，1970年，男，博士，研究员，博士生导师，研究方向：主要研究方向包括基因组规模代谢网络的构建质控及其结构和功能分析、多约束细胞模型的构建、代谢途径和代谢工程改造策略设计、生物计算设计算法和网站工具开发、基于人工智能的蛋白功能预测和菌种设计等。
E-mail：ma_hw@tib. cas. cn



第一作者： 方洲，1992年，男，博士，现任中国科学院数学与系统科学研究院副研究员，博士生导师。研究方向：研究工作聚焦于生命系统控制与计算，包括生物系统滤波与辨识算法设计、复杂动力学系统分析、基因线路设计等。
E-mail：zhfang@amss. ac. cn



第一作者： 杨晨，1996年，女，博士，博士后，研究方向：人工合成细胞数字孪生系统的智能模拟、设计与构建。
E-mail：yangchen@tib. cas. cn

作者贡献声明：
杨晨：确定研究对象范围、收集数据、采集、清洗与分析数据，设计研究思路、设计研究方案、研究方案可行性调查分析、实施模拟分析研究过程，文献调研与整理，起草论文、修订论文、审核论文。刘金林：参与模拟分析研究过程实施，参与数据收集、采集、清洗与分析。方洲：提出概念、基本框架、提出研究方向、设计论文框架。马红武：提出概念、基本框架、提出研究方向、设计论文框架，参与论文修订、论文最终版本修订。
所有作者均已阅读并同意文章的已发布版本。
机构审查委员会声明：
不适用。
知情同意声明：
不适用。
数据可用性声明：
如需实验数据，请联系本文通讯作者。
利益冲突：
没有利益冲突。

附件1

模型构建与模拟方法

1 教师模型分层降阶混合动力学建模

教师模型采用分层降阶动力学方法。模型以转录、翻译、氨酰化、翻译因子循环、核糖体循环及能量小分子代谢等生化过程为基础，首先保留主要反应的计量关系和资源耦合关系；随后根据反应模块的时间尺度差异和计算刚性，将网络显式划分为质量作用模块、背景化资源模块和准稳态近似（Quasi-Steady-State Approximation，QSSA）有效模块。

其中，质量作用动力学用于描述模型中显式保留的基元反应。假设在充分混合体系中，某一反应发生的瞬时通量与所有反应物的碰撞速率成正比。因此，对于一般反应：

$$\sum_i \lambda_{i,r} X_i \rightarrow \sum_i \lambda'_{i,r} X_i \qquad (1)$$

其质量作用通量可写为：

$$v_r = k_r \prod_i C_i^{\lambda_{i,r}} \qquad (2)$$

其中，v_r 表示第 r 个反应的通量，k_r 表示该反应的速率常数，C_i 表示代谢物 X_i 的当前浓度，$\lambda_{i,r}$ 表示 X_i 在该反应中的反应物计量数。反应通量确定后，各代谢物浓度随时间 t 的变化由所有反应通量按计量系数叠加得到：

$$\frac{dC_i}{dt} = \sum_r (\lambda'_{i,r} - \lambda_{i,r}) v_r \qquad (3)$$

背景化处理用于 ATP、GTP 和 PPi 等被快速消耗和再生的小分子资源池，使它们在不展开快速循环细节的情况下，仍然作为动态资源池影响转录和翻译速率。

在模型中，一个被背景化的小分子资源池可以概括写成：

$$\frac{dS_i}{dt} = R_i^{\mathrm{explicit}(X_s, S_i)} + J_i^{\mathrm{bg}(B_i, S_i)} \qquad (4)$$

其中，S_i 表示第 i 个被背景化处理的小分子资源池浓度，X_s 表示显示保留的代谢物浓度集合，$R_i^{\mathrm{explicit}(X_s, S_i)}$ 表示显式保留反应对 S_i 浓度变化的合计影响，B_i 表示背景化模块中用于维持第 i 个小分子浓度的等效储备量，$J_i^{\mathrm{bg}(B_i, S_i)}$ 表示背景化模块与显式小分子池之间的等效净交换通量。

QSSA 有效模块用于处理转录-翻译主链的短寿命中间态。对于生成后很快被消耗的局部中间态，不作为独立代谢物保留，而是折叠进每个位置步的有效推进速率系数。QSSA 处理后的局部通量可概括写成：

$$v_q = Q_q \cdot k_{eff,q}(R_q) \qquad (5)$$

其中，v_q 表示从 Q_q 向下一位置状态推进的有效通量，Q_q 表示第 q 个显式位置状态浓度，R_q 表示影响第 q 步推进的局部资源浓度集合，包括该步

所需底物、因子或复合物的浓度，$k_{eff,\ q}(R_q)$表示第q步的QSSA有效推进速率系数。

对于连续式PURE体系，教师模型引入了补料和洗出模块，把反应腔与外部容器之间的持续交换，近似成一个膜分隔的连续进料/出料过程。对可连续补料的小分子代谢物，净交换项表示为：

$$\frac{dE_i}{dt} = F_i(E) + \gamma\cdot(E_{i,feed} - E_i) \quad (6)$$

其中，E_i表示反应体系内第i个可连续补料的小分子代谢物的浓度，γ表示交换速率常数，$F_i(E)$表示内部反应引起的第i个代谢物的净浓度变化速率，E表示反应体系的代谢物浓度向量，$E_{i,\ feed}$表示外部储液池中第i个可连续补料的小分子代谢物的浓度。

2 PINN第一阶段先验模拟代谢物浓度轨迹提取方法

教师模型作为PINN的机制化数据监督来源，在质量作用动力学、化学计量守恒和开放边界交换约束下运行，为PINN二阶段模拟提供可比代谢物浓度轨迹：

$$\delta_i(t) = log\left(C_i^P(t) + \varepsilon_C\right) - log\left(C_i^T(t) + \varepsilon_C\right) \quad (7)$$

$$L_{data} = \frac{1}{N_{data,\ train}}\sum_{X_i \in \Omega_{cmp}}\ \sum_{t \in T_{train}}\ \delta_i(t)^2 \quad (8)$$

其中，$\delta_i(t)$为第i个代谢物在时间t的对数浓度预测误差，$C_i^P(t)$为PINN模拟的t时刻下代谢物X_i的浓度，ε_C为避免低浓度数值奇异的微小常数。$C_i^T(t)$为教师模型模拟的t时刻下代谢物X_i的浓度，L_{data}为可比代谢物浓度轨迹监督损失，$N_{data,\ train}$为训练采样点总数。Ω_{cmp}为教师模型与基元反应精细度的CRN网络（以下简称精细CRN网络）中可一一对应的代谢物集合，T_{train}为训练集时间集合。

在第一阶段先验模拟中，PINN神经网络输出精细CRN的完整状态向量。其中，可比代谢物浓度轨迹通过L_{data}监督。其余隐藏状态则由初始条件、网络拓扑及代谢物之间的计量耦合关系共同推断，形成待校正的全状态先验轨迹。然后，通过第二阶段对全状态轨迹进行物理约束校正。

3 PINN第二阶段物理校正过程的关键代谢物提取方法

本研究通过机制扰动分析，扰动教师模型的不同组分，比较动力学相关指标的变化，确定关键代谢物。随后将关键代谢物用于PINN第二阶段增强型物理积分损失L_{key}的输入。

4 PINN第二阶段物理校正过程的多目标损失计算方法

PINN第二阶段在先验模拟基础上进行轨迹校正。将全代谢物离散时间节点浓度作为待校正变量。每次校正时，模型首先把完整节点状态代入精细CRN网络的质量作用动力学方程，计算第r条反应通量，见（2）式：随后，所有反应通量通过化学计量矩阵转换为各代谢物的净生成或消耗，并叠加连续体系的开放边界交换项，得到物理约束算子给出的动力学变化方向：

$$R\quad(\mathrm{t}) = \sum_r\left(\lambda'_{i,r} - \lambda_{i,r}\right)v_r(t) + q_i^{ex}(t)t \in T_{eval}(11)$$

其中，R（t）表示第i个代谢物在t时刻下物理约束算子算出的浓度变化速率，$\lambda'_{i,\ r}$和$\lambda_{i,\ r}$分别表示第r个反应中第i个代谢物的生成和消耗的化学计量系数，$v_r(t)$表示质量作用动力学计算的t时刻下的第r个基元反应的通量，$q_i^{ex}(t)$表示t时刻下第i个代谢物的开放边界交换通量。

接下来对每个代谢物计算相邻节点之间的物理约束偏离残差，用于衡量预测轨迹是否沿物理约束允许的轨迹演化，并进一步计算出整体物理积分损失与关键代谢物增强型物理积分损失来衡量物理约束轨迹偏离程度。

$$e_{m,i} = C_i(t + \Delta t) - C_i(t) - \Delta t/2\cdot[R_i(t) + R_i(t + \Delta t)] \quad (12)$$

$$L_{int} = \frac{1}{N_{cmp}(N_{eval} - 1)}\sum_{i \in \Omega_{cmp}}\ \sum_{m \in [1,2,\cdots,N_{eval}-1]}\ \left(\frac{e_{m,i}}{s_i^{rhs}}\right)^2 \quad (13)$$

$$L_{key} = \frac{1}{N_{key}(N_{eval} - 1)}\sum_{i \in \Omega_{key}}\ \sum_{m \in [1,2,\cdots,N_{eval}-1]}\ \left(\frac{e_{m,i}}{s_i^{rhs}}\right)^2 \quad (14)$$

其中，$e_{m,\ i}$表示第m个相邻时间窗口内，第i

个代谢物的物理约束偏离残差，$C_i(t)$表示第i个代谢物在时间节点t的PINN预测浓度，$\triangle t$表示相邻时间间隔。L_{int}表示整体物理积分损失，N_{cmp}表示教师模型与精细CRN网络中可一一对应的代谢物数目，$N_{eval}-1$表示纳入通量评价的相邻时间窗口数目，s_i^{rhs}表示物理方程预测的第i个代谢物在各相邻时间窗口内积分变化的均方根，用于将绝对物理约束偏离残差归一化为相对物理约束偏离残差。L_{key}表示关键代谢物增强型物理积分损失，Ω_{key}为关键代谢物集合，N_{key}表示关键代谢物数目。

为避免每一段校正使浓度轨迹发生过大的变化，构建了轨迹邻近损失：

$$L_{prox}=\frac{1}{N_x N_{eval}}\sum_{i\in[1,2,\cdots,N_x]}\sum_{m\in[1,2,\cdots,N_{eval}]}[log_{10}(C_i(t)+\varepsilon_C)-log_{10}(C_i^{ref}(t)+\varepsilon_C)]^2 \quad (15)$$

其中，L_{prox}表示轨迹邻近损失，N_x表示精细CRN网络的总代谢物数目，N_{eval}表示纳入通量评价的时间点数目，$C_i^{ref}(t)$表示本段校正开始前的第i个代谢物在时间节点t的浓度，ε_C为低浓度保护常数。

L_{data}、L_{int}、L_{key}与L_{prox}一起建立了总损失，用来对节点浓度求梯度并迭代优化：

$$L_{total}=\omega_{data}L_{data}+\omega_{int}L_{int}+\omega_{key}L_{key}+\omega_{prox}L_{prox} \quad (16)$$

其中，L_{total}表示总损失，ω_{data}、ω_{int}、ω_{key}、ω_{prox}分别表示教师模型可比代谢物浓度轨迹监督损失、整体物理积分损失、关键代谢物增强型物理积分损失、轨迹邻近损失的权重。最终，梯度计算使反应通量、化学计量关系和边界交换通量随浓度校正而同步趋向更合理的物理状态。

5 PINN模拟准确性的评价方法

本研究采用验证集代谢物浓度预测的平均相对误差（Mean Relative Error， MRE）、决定系数（Coefficient of determination， R^2）和代谢物浓度轨迹模拟归一化均方根残差（Normalized Root Mean Square Residual of trajectory simulation，NRMSR）来评价PINN模拟的准确性：

$$MRE=\frac{1}{N_{data,val}}\sum_{X_i\in\Omega_{cmp}}\sum_{t\in T_{val}}\frac{|C_i^P(t)-C_i^T(t)|}{\max[C_i^T(t),\varepsilon_C]} \quad (17)$$

$$R^2=1-\frac{\sum_{X_i\in\Omega_{cmp}}\sum_{t\in T_{val}}(C_i^T(t)-C_i^P(t))^2}{\sum_{X_i\in\Omega_{cmp}}\sum_{t\in T_{val}}(C_i^T(t)-\overline{C^T})^2} \quad (18)$$

$$NRMSR=\frac{1}{N_{dcmp}}\sum_{i\in\Omega_{dcmp}}\sqrt{\frac{1}{N_{eval}-1}\sum_{m\in[1,2,\cdots,N_{eval}-1]}(\frac{e_{m,i}}{s_i^{rhs}})^2} \quad (19)$$

其中，$N_{data,\ val}$表示验证采样点总数，T_{val}表示验证集时间集合，$\overline{C^T}$为教师模型代谢物浓度均值，N_{dcmp}为参与轨迹评价的代谢物数目，Ω_{dcmp}为参与轨迹评价的代谢物集合。