arXiv:2305.03257v1 [q-bio.QM] 5 May 2023

中国仓鼠卵巢细胞生物反应器的数据驱动与物理信息建模

预印本

Tianqi Cui[1,4], Tom Bertalan[1,4],Nelson Ndahiro[1],Pratik Khare[1],
Michael Betenbaugh[1],Costas Maranas[2],和 Ioannis G. Kevrekidis[1,3]

[1]约翰斯·霍普金斯大学化学与生物分子工程系，美国马里兰州巴尔的摩 21218[2]宾夕法尼亚州立大学化学工程系，美国宾夕法尼亚州大学公园 16802[3]约翰斯·霍普金斯大学应用数学与统计学系，美国马里兰州巴尔的摩 21218[4]这些作者贡献相同。

摘要

补料分批培养是利用哺乳动物细胞培养物生产生物制品的一种成熟操作模式。定量建模结合了某些关键反应步骤的动力学与基于通量平衡分析的优化驱动代谢通量分配，但已知这种方法会导致某些数学不一致性。本文提出一种物理信息数据驱动混合模型（“灰箱”），利用过程数据学习中国仓鼠卵巢（CHO）细胞生物反应器的动态演化模型。该方法既纳入物理定律（例如质量衡算），也纳入代谢通量的动力学表达式。随后利用机器学习（ML）来：(a) 直接学习演化方程（黑箱建模）；(b) 恢复未知物理参数（“白箱”参数拟合）；或——重要的是——(c) 学习部分未知的动力学表达式（灰箱建模）。我们将超定代谢生物物理系统的凸优化步骤编码为架构中的一个可微前馈层，从而将部分物理知识与数据驱动的机器学习连接起来。

1 引言

中国仓鼠卵巢（CHO）细胞广泛应用于生物医学研究，是生产治疗性蛋白质最常用的哺乳动物细胞系 [1]。使用 CHO 细胞的优势在于，它能够为蛋白质疗法（例如治疗性抗体）实现正确的（即哺乳动物特异性）糖基化模式。与传统的分批培养相比，补料分批发酵更常用于这类细胞系，因为它更便于控制某些营养物质的浓度；这些营养物质通过确保前体氨基酸的供应，能够影响目标蛋白质治疗分子的产率或生产率 [2]。然而，缺乏完整、清晰且定量的代谢模型，成为实现准确、精确的系统仿真与控制的障碍。

过去几十年来，融入物理知识的数学模型已广泛应用于细胞代谢分析 [3, 4]。利用稳定碳（即 13C）标记底物技术的代谢通量分析（MFA）是唯一能够提供内部通量信息的技术 [5, 6, 7]。另一方面，通量平衡分析（FBA）能够对整个代谢过程中的碳和能源资源进行全局盘点。通过应用优化原理，可以推导出生物量形成或其他产物（如代谢物或蛋白质）的最大理论产率 [8, 9]。有时，对于某些代谢步骤，现有详细的动力学表达式，可根据代谢物浓度和酶水平准确估算代谢反应的通量 [10]。这需要确定一系列酶学参数的数值。然而，这类表达式通常仅适用于部分反应，因此需要采用混合建模方法：对于其余缺少动力学表达式的反应，则通过优化来确定代谢通量。这会形成一个由反应的化学计量关系决定的常微分方程（ODE）系统。此外，鉴于代谢反应的时间常数通常相对较短（例如，处于

毫秒至秒的范围），与细胞生长和死亡等其他细胞过程相比，伪稳态假设（PSSA）认为，任何细胞内代谢物的积累速率都可以近似为零。对于某些反应，可以基于热力学考量推断其（不可）逆性；对于其他反应，则可以根据化学动力学考量估算反应速率 [11]。

如今，数据驱动方法日益广泛地用于识别复杂的系统动力学，包括传统回归方法、神经网络及其变体 [12, 13, 14, 15, 16]。早在 20 世纪 90 年代初，人们便已知道，将神经网络嵌入数值积分器中，可以有效地近似微分方程，甚至学习对近似物理模型的修正，从而对其进行补充和增强 [13, 12, 17,18, 14]。当底层物理机制不明确时，也可以利用神经网络直接推断系统变量的演化过程 [19, 20, 21, 22, 23]。物理信息神经网络 (PINNs) [24], 、系统生物学信息神经网络 (SBINNs) [25, 26], 及类似架构 [27], , 能够并且已经被用于解决监督学习任务，同时遵循已知的物理定律、系统生物学规律等 [28]。然而，正如下文将要讨论的，代谢模型结构上的模糊性，给利用补料分批培养实验代谢数据中部分已知的物理信息带来了不小的技术困难；而且，这种模糊性会显著影响灰箱神经网络的训练过程，尤其是当其试图根据实验推断此类模型时。本文旨在阐明这些建模歧义的性质，说明它们为何要求对用于识别任务的传统神经网络架构及其训练方式进行调整；并实现能够有效识别代谢动力学和参数的网络，在神经网络训练中发挥物理建模与科学计算之间的协同作用。

2 方法

2.1 生物物理模型的结构

简而言之，下文将使用的混合型中国仓鼠卵巢（CHO）生物反应模型（经过某些修改（见附录D 和附录E，）后，成为文献 [10], 中提出的模型（这是我们的起点））描述了一个连续时间动力系统（相关术语定义见表1：

$$\frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} = \boldsymbol{f}_{\mathbf{eqn:ode}}(\boldsymbol{C}; \boldsymbol{v}(\boldsymbol{C}; \boldsymbol{\alpha})); \tag{1}$$

乍看之下，这些演化方程是简单的常微分方程（方程式（B 的表达式见附录1））；然而，由于计算右端项需要——正如我们将看到的——求解一个优化问题，因此我们需要为这些方程的性质暂时另取一个名称。结合现有文献 [29, 30, 9]，本文将其称为**动**态**通**量**平**衡**分**析（DFBA）方程。

这里，$\boldsymbol{C} \in \mathbb{R}^K$（$K = 14$）是实验追踪的变量（为简单起见，我们将其统称为“浓度”，尽管其中可能包括代谢物浓度、细胞密度或其他变量，见表 3）；$\boldsymbol{v} \in \mathbb{R}^N$（$N = E + I = 35$）是所有通量（反应速率，所有反应表达式见附录 C），包括 I 细胞内通量（可根据 $\boldsymbol{C}$ 预先计算）$\boldsymbol{v}_I \in \mathbb{R}^I$（$I = 14$）和 E 细胞外通量 $\boldsymbol{v}_E \in \mathbb{R}^E$ $E = 21$ $\boldsymbol{v}_{E\,ir} \in \mathbb{R}^{E_{ir}}$ $E_{ir} = 14$ $\boldsymbol{v}_{E\,ir} \geq 0$ （）。部分细胞外通量被假定为不可逆（. （）.），而其他通量则被假定为可逆（$\boldsymbol{v}_{E,\ r} \in \mathbb{R}^{E_r}$（$E_r = 7$））；$v$ 是 $\boldsymbol{C}$ 和 $\boldsymbol{\alpha}$的函数，其中 $\boldsymbol{\alpha} \in \mathbb{R}^P$（$P = 45$）是动力学参数。

符号表示	变量	维度
$\boldsymbol{C}$	实验追踪的变量	$K = 14$
$\boldsymbol{\alpha}$	动力学参数	$P = 45$
$\boldsymbol{v}_I$	细胞内通量	$I = 14$
$\boldsymbol{v}_{E,r}$	可逆（细胞外）通量	$E_r = 7$
$\boldsymbol{v}_{E,ir}$	不可逆（细胞外）通量	$E_{ir} = 14$
$\boldsymbol{v}_E$	细胞外通量	$E = E_{ir} + E_r = 21$
$\boldsymbol{v}$	所有通量	$N = E + I = 35$
$\mathbf{S}$	化学计量矩阵	$M \times N = 24 \times 35$

表 1：所有变量的符号表示和维度（此处 $K = 14$，$I = 14$ 和 $E_{ir} = 14$ 恰好如此纯属巧合）。

给定 $\boldsymbol{C}$ 和 $\boldsymbol{\alpha}$，通常有两种截然不同的方法来评估 $\boldsymbol{f}_{\mathbf{eqn:ode}}$ 式（1）。两种方法都包含以下步骤，但在目标函数与约束的具体组合，以及用于强制执行这些条件的优化方法上有所不同。

给定 $\boldsymbol{\alpha}$ 以及一组初始 值 $\boldsymbol{C}_0$ 作为浓度，浓度的时间导数（例如右端项 o f 方程式（1））可以 通过以下步骤进行计算。

1. 根据浓度 $\boldsymbol{C}$ 和给定的动力学方程公式，计算胞内通量速率 $\hat{\boldsymbol{v}}_I \in \mathbb{R}^I\,(I=14)$ 的初步更新值

$$\hat{\boldsymbol{v}}_I = \boldsymbol{f}_{\mathbf{kin}}(\boldsymbol{C};\boldsymbol{\alpha}), \tag{2}$$

其中 $\boldsymbol{f}_{\mathbf{kin}}$： $\mathbb{R}^{K\times P}\mapsto\mathbb{R}^I$（所有动力学表达式的公式见附录D，根据该模型所做的动力学表达式修改见附录E，该模型见 [10]）。

2. 通量必须满足某些约束：
   - 已知的动力学表达式，即方程式（2）。
   - 准稳态假设要求

$$\mathbf{S}\cdot\boldsymbol{v}=0, \tag{3}$$

涉及化学计量矩阵 $\mathbf{S}\in\mathbb{R}^{M\times N}\,(M=24$ 是稳态下代谢物的数量，见附录 F 适用于 $\mathbf{S}$的所有元素）。如果按照 $\mathbf{S}$ 的列分别对应 I 和 E 组分拆分（即 $\mathbf{S}_I=\mathbf{S}\cdot\mathbf{B}_I$，$\mathbf{S}_E=\mathbf{S}\cdot\mathbf{B}_E$ 其中 $\mathbf{B}_I\in\mathbb{R}^{N\times I}$ 和 $\mathbf{B}_E\in\mathbb{R}^{N\times E}$ 是两个指示矩阵，分别表示所有反应的细胞内和细胞外索引），则公式（3）可等价地表示为：

$$\mathbf{S}_I\cdot\boldsymbol{v}_I+\mathbf{S}_E\cdot\boldsymbol{v}_E=0$$

   - 在全部 21 个细胞外通量 $\boldsymbol{v}_E$ 中，已知其中 14 个是不可逆的，这要求 s

$$\boldsymbol{v}_{E,ir}=\mathbf{B}_{ir}\cdot\boldsymbol{v}_E\geq 0, \tag{4}$$

其中，$\mathbf{B}_{ir}\in\mathbb{R}^{E_{ir}\times E}$ 是一个指示矩阵，包含所有细胞外通量中不可逆通量的索引。

请注意，方程（2）和方程（3）合起来包含 38 个独立线性方程，而未知变量 $\boldsymbol{v}$ 的维度只有 35，因此构成了一个超定方程组。为解决这一问题，可以选择使部分方程严格成立，而使其他方程近似成立（例如采用最小二乘意义下的近似）。由此产生了两种截然不同的方法，即“基于动力学的方法”和“基于化学计量的方法”，用于计算胞内通量速率 $\boldsymbol{v}_I$ 和细胞外通量 $\boldsymbol{v}_E$。需要指出的是，对于同一代谢动力学方案的相同初始条件，这两种方法通常会导致截然不同的动态演化。

(a) 基于动力学的方法：我们使动力学方程（2）严格成立，并使化学计量方程（3）近似成立，从而得到

$$\begin{cases}\boldsymbol{v}_I=\hat{\boldsymbol{v}}_I,\\ \boldsymbol{v}_E=\operatorname{argmin}_{\boldsymbol{v}_E}||\mathbf{S}_I\cdot\hat{\boldsymbol{v}}_I+\mathbf{S}_E\cdot\boldsymbol{v}_E||_2^2,\ \text{s.t. }\mathbf{B}_{ir}\cdot\boldsymbol{v}_E\geq 0.\end{cases} \tag{5}$$

这里可以看出，该优化问题是一个带约束的线性最小二乘问题（这意味着它实际上是一个凸优化问题）。此外，如果忽略这些约束，就可以通过计算$\mathbf{S}_E$的伪逆，得到 $\boldsymbol{v}_E$ 的解析解：

$$\boldsymbol{v}_E=-(\mathbf{S}_E)^+\cdot\mathbf{S}_I\cdot\hat{\boldsymbol{v}}_I.$$

（b）化学 计量法 [10]：我们精确满足化学计量方程（3），然后 d 再近似 满足动力学方程（2），从而得到

$$(\boldsymbol{v}_I,\boldsymbol{v}_E)=\operatorname{argmin}_{\boldsymbol{v}_I,\boldsymbol{v}_E}||\boldsymbol{v}_I-\hat{\boldsymbol{v}}_I||_2^2,\ \text{s.t. }\mathbf{S}_I\cdot\boldsymbol{v}_I+\mathbf{S}_E\cdot\boldsymbol{v}_E=0,\mathbf{B}_{ir}\cdot\boldsymbol{v}_E\geq 0. \tag{6}$$

这是一个带线性约束的最小二乘优化问题。如果忽略不等式约束，我们可以通过拉格朗日乘子法得到（$\boldsymbol{v}_I$，$\boldsymbol{v}_E$）的解析解，详情请参见附录G。

为改善这两个嵌套优化问题的数值计算，我们实际上会对给定的$\hat{\boldsymbol{v}}_I$数值进行缩放：将其除以 1000，以调整其数值范围，使其在进入任一优化问题时达到 $\sim 1-10$，并在退出前将所得通量 $\boldsymbol{v}_I$ 和 $\boldsymbol{v}_E$ 乘以相同的因子。这不会改变解，但能改善数值条件性。

优化步骤之后，我们得到

$$\boldsymbol{v}=\mathbf{B}_I\cdot\boldsymbol{v}_I+\mathbf{B}_E\cdot\boldsymbol{v}_E=\boldsymbol{f}^{(\ell)}_{\text{optim}}(\hat{\boldsymbol{v}}_I), \tag{7}$$

其中 $\ell\in\{k,s\}$ 表示我们采用的是动力学方法还是化学计量学方法来求解通量。

3. 组合方程（1）、方程（7）和方程（2）以创建

$$\frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} = \boldsymbol{f}_{\mathbf{eqn:ode}}(\boldsymbol{C}; \boldsymbol{f}^{(\ell)}_{\mathrm{optim}}(\boldsymbol{f}_{\mathbf{kin}}(\boldsymbol{C}; \boldsymbol{\alpha}))). \tag{8}$$

由于通量 $\boldsymbol{v}$ 是每个 $E+I$ 反应的反应速率，因此这可直接从这些反应的化学计量关系得出。

无论采用第一种还是第二种方法，随后都可以使用误差控制积分器对所得方程组进行积分，以获得所有浓度的完整时间序列，例如，

$$\boldsymbol{C}(t = t_{n+1}) = \boldsymbol{f}_{\mathbf{int}}(\boldsymbol{C}(t = t_n); \boldsymbol{f}_{\mathbf{eqn:ode}}) = \boldsymbol{C}(t = t_n) + \int_{t_n}^{t_{n+1}} \boldsymbol{f}_{\mathbf{eqn:ode}}(\boldsymbol{C}; \boldsymbol{v}(\boldsymbol{C}; \boldsymbol{\alpha}))\, \mathrm{d}t, \tag{9}$$

其中 $\{t_i : \ i = 0, \ 1, \ 2, \ \cdots\}$ 是等间隔时间戳的集合。不过，值得注意的是，当不同约束变为活跃约束时，可能会（而且确实会）出现速率不连续的情况（见图 3）。还要注意，生物反应器的典型操作规程通常要求在特定时间点添加物质（例如营养物质），从而导致系统状态随时间出现不连续。我们将在下文第3.1节说明这两种不连续情况。

本文所依据的若干重要成果已在先前的研究中确立，最初应用于大肠杆菌 [9]，随后扩展至较新的哺乳动物生物制造目标 [10]。除了我们上文展示的内层（优化）问题约束外，[9] 还对其外层（时间积分）问题施加了其他约束，例如代谢物非负以及通量变化率限制。在其动态（或静态）优化方法（DOA，或 SOA）中，他们确定了整个轨迹（或单个轨迹片段，其中通量保持恒定）上的通量。我们的模拟可以视为一种 SOA，其中片段就是单个积分步（ [10]中也是如此）。

在开始之前，先说明一下模型梯度的计算：许多高精度积分器需要系统雅可比矩阵以及关于参数的灵敏度。这对于微分代数方程组（带有等式约束的微分方程）的积分也很重要。此外，这些梯度（关于状态变量和/或参数）对于辨识任务至关重要：训练神经网络，使其近似系统方程和/或根据数据估计系统参数。如上文所述，我们的演化方程并非简单的显式常微分方程，而是其右端项由求解一个取决于当前状态的优化问题得出。因此，与显式右端项的情况相比，对这些常微分方程进行精确求值（以及计算其灵敏度和变分）并不那么直接。

2.2 黑箱模型

我们的黑箱模型是嵌入数值积分方案（例如前向欧拉方案或龙格–库塔模板）中的多层感知机（MLP），其中 MLP（$\mathrm{NN}_\mathrm{b}(\cdot;\boldsymbol{\theta})$）用于学习常微分方程（ODE）的右端项（RHS）：

$$\tilde{\boldsymbol{C}}(t = t_{n+1}) = \boldsymbol{f}_{\mathbf{int}}(\boldsymbol{C}(t = t_n); \mathrm{NN}_\mathrm{b}) = \boldsymbol{C}(t = t_n) + \int_{t_n}^{t_{n+1}} \mathrm{NN}_\mathrm{b}(\boldsymbol{C}; \boldsymbol{\theta})\, \mathrm{d}t. \tag{10}$$

生成数据集的详细信息见第3.1节。请注意，此处右端项仅取决于系统状态；也可以让神经网络公式（10）依赖于物理输入参数（例如喂养条件或基础基因表达速率），方法是将这些参数作为额外输入提供给神经网络函数。如果之后希望针对某项额外的全局目标（例如最大化生物量生成）优化操作条件，这一点将十分重要。这种可能性已在早期研究中得到展示[17]；此处不再赘述。

2.3 白箱与灰箱模型

2.3.1 模型结构

与纯粹数据驱动的黑箱模型相比，白箱模型和灰箱模型能够利用已有的物理知识，因此只有模型中的未知部分需要训练。本文中的这两种模型，其结构与我们认为的真实生物物理模型相似（见图1），差异仅限于初步细胞内通量的计算 $\hat{}\boldsymbol{v}_I$。白箱模型假设部分动力学参数 $\boldsymbol{\alpha}$ 未知或需要校准，而灰箱模型则认为部分动力学表达式 $\boldsymbol{f}_{\mathbf{kin}}$

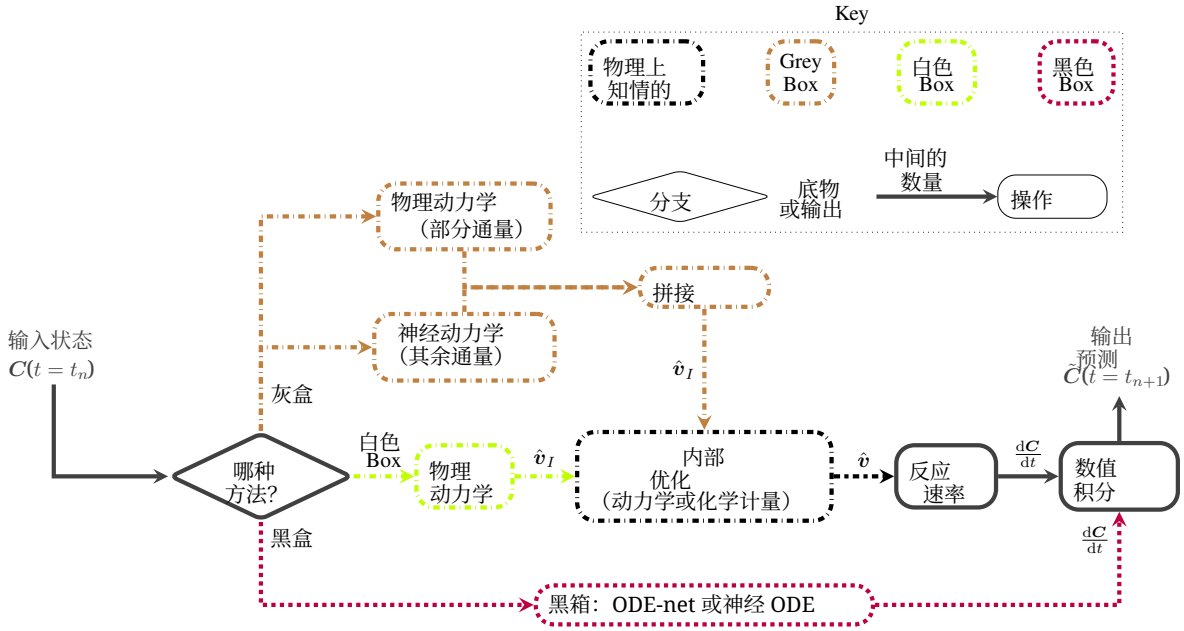


图1：**白箱、灰箱和黑箱内部架构**。 运算以方框表示，数据或预测不加方框表示，重要的命名中间体标注在边上。颜色和图案用于区分灰箱、白箱和黑箱方法各自不同的模型路径，以及三者共有的路径。

没有已知的函数形式，因此用神经网络近似来替代这些表达式。自然也可以构建一种混合模型的混合版本，同时包含未知的（“白箱”）动力学参数和未知的（“黑箱”）动力学表达式，从而形成整体上的灰箱模型。

尽管白箱模型表面上与典型的参数拟合问题相似，但由于其评估过程中包含内部优化步骤，传统拟合方法（如一般线性最小二乘法）无法很好地适用于我们的框架。因此，我们将采用基于梯度的拟合方法，并使用第2.3.2节所述的可微凸优化层。这里提醒读者，原始生物物理模型包含两种不同的通量计算方法（见式（7））， 因此，我们也构建了两种版本的白箱或灰箱框架：即使用动力学版本模型学习由动力学方法生成的数据集，并使用化学计量版本模型学习由化学计量方法生成的数据集。

2.3.2 凸优化中的梯度计算

特别是在本节中，我们需要定义一些术语。

模型指用于进行预测的可微程序：包括右端项评估（白箱、灰箱或黑箱），可能需要嵌入式凸优化程序方程（5）或方程（6）（***ECOP***）；以及在数值积分步骤中使用这些右端项。

此 ECOP 的底物包括两者

$$\begin{matrix}\text{constants} & \mathbf{S}_I, \mathbf{S}_E, \text{ and } \mathbf{B}_{ir}; & \text{and} \\ \text{outputs from upstream modules} & \hat{\boldsymbol{v}}_I & \text{(function evaluations).}\end{matrix} \tag{11}$$

对于每次调用右端项时求解 ECOP，所有这些都将视为常量。

参数在此指可由我们的外层训练循环修改的量，包括动力学方程 $\boldsymbol{\alpha}$（即方程（2）的动力学参数（当我们进行白盒参数估计或灰盒参数估计时）；以及神经网络参数 $\boldsymbol{\theta}$，即可训练权重和偏置（当我们训练灰盒网络以

恢复未知的函数依赖关系时）：

$$\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{\theta}. \tag{12}$$

根据具体实验，我们将这些参数分为可训练参数和不可训练（固定）参数。

该输出 的 ECOP 是报告的收敛数值，即该问题求解的变量 的数值，包括两者 h

$$\text{the fluxes } \boldsymbol{v}_E \text{ and } \boldsymbol{v}_I; \quad \text{and, possibly auxiliary variables (described below)} \quad \boldsymbol{r}. \tag{13}$$

在训练过程中，我们希望损失函数（参见后续章节中针对特定实验所述的式（16））关于所有可训练参数（式（12））均可微。这要求 模型预测 可微，因此模型中的每个步骤（包括 ECOP）都必须关于这些参数可微。

为了在 PyTorch 等基于梯度的计算框架中实现这一点，我们采用了 cvxpylayers 软件包，该软件包专为此类问题开发 [31]。该软件包本身使用 cvxpy（一种嵌入 Python 的凸优化问题建模语言） [32, 33, 34] 和 diffcp（一种用于计算锥规划导数的 Python 软件包；锥规划是凸规划的一种特殊情形） [35, 36, 37]。

为使这些软件包能够正确计算 ECOP 的输出方程（13）相对于其所有输入方程（11）的梯度，问题必须具备某些结构特征。具体而言，问题需要改写为符合规范凸规划（DCP）和规范参数化规划（DPP）的规则。DCP 是一种构造凸规划的体系，它将常见的凸函数（例如 x^2、$|x|$）与复合规则和组合规则（例如 $f \circ g$ 在 f 和 g 均为凸函数时也是凸函数；凸函数的非负线性组合仍为凸函数）结合起来。遵循这些规则后，库便可自动判断整个问题是否确实为凸问题。

DPP 是 DCP 的子集，它进一步要求 ECOP 的所有表达式相对于 ECOP 输入方程（11）均为仿射表达式。已有研究证明 [34]，符合 DPP 要求的凸规划可以通过可逆变换转化为锥规划（并可从`diffcp`获取其导数信息）。因此，DPP 主要用于依赖底物的凸规划，使整个程序无需实际展开优化循环并通过反向传播，即可实现可微分。

由于 DCP 要求 ECOP 中的表达式相对于问题底物是仿射的（式 (11)），因此两个底物的乘积不是可接受的表达式。这意味着，例如，式 (5) 目标函数中的 $(\mathrm{S}_I \cdot \hat{}\boldsymbol{v}_I + \mathrm{S}_E \cdot \boldsymbol{v}_E)^T \cdot (\mathrm{S}_I \cdot \hat{}\boldsymbol{v}_I + \mathrm{S}_E \cdot \boldsymbol{v}_E)$ 必须重新表述。解决方法是在式 (14) 和式 (15) 中增加另一个变量 $\boldsymbol{r}$，并对这个附加变量施加等式约束，从而使新定义的问题与原问题等价。此外，在动力学情形下，为避免直接出现底物乘积 $\mathrm{S}_I \cdot \hat{}\boldsymbol{v}_I$，我们需要将 $\boldsymbol{v}_I$ 纳入优化变量，但随后必须将式 (5) 中原本预先优化的表达式 $\boldsymbol{v}_I = \hat{}\boldsymbol{v}_I$ 改为式 (14) 中实际的等式约束。

总之，对于动力学方法，我们将式 (5) 改写为

$$\begin{aligned} \min_{\boldsymbol{r}, \boldsymbol{v}_I, \boldsymbol{v}_E} \quad & ||\boldsymbol{r}||_2^2 \\ \text{s.t.} \quad & \mathbf{B}_{ir} \cdot \boldsymbol{v}_E \geq 0 \\ & \boldsymbol{r} = \mathbf{S}_I \cdot \hat{\boldsymbol{v}}_I + \mathbf{S}_E \cdot \boldsymbol{v}_E \\ & \boldsymbol{v}_I - \hat{\boldsymbol{v}}_I = 0 \end{aligned} \tag{14}$$

对于化学计量方法，我们将方程（6）改写为

$$\begin{aligned} \min_{\boldsymbol{r}, \boldsymbol{v}_I, \boldsymbol{v}_E} \quad & ||\boldsymbol{r}||_2^2 \\ \text{s.t.} \quad & \mathbf{B}_{ir} \cdot \boldsymbol{v}_E \geq 0 \\ & \boldsymbol{r} = \boldsymbol{v}_I - \hat{\boldsymbol{v}}_I \\ & \mathbf{S}_I \cdot \hat{\boldsymbol{v}}_I + \mathbf{S}_E \cdot \boldsymbol{v}_E = 0 \end{aligned} \tag{15}$$

以方程（11）为底物。

经过这些修改，这两个问题均符合 DPP；因此，我们能够计算问题输出（式（13））相对于问题底物（式（$\boldsymbol{v}_I$ 和 $\boldsymbol{v}_E$）的导数（尤其是极小化点），也能根据需要在更大的 PyTorch 反向传播中计算向量－雅可比积，最终得到相对于参数（式（12））的损失梯度。

2.4 自回归损失

对于时间序列数据所蕴含动力学的监督学习，一种方法是将前一时间步的真实预测值／输出作为当前时间步的底物，这便是教师强制方法（在 [38] 中也称为教授强制）。另一种方法是将模型在前一时间步的预测作为底物，这称为“自回归训练”。事实上，如果将连续的数据轨迹划分为每段 M 个时间步的片段，并将 M 降至 2，就会发现前者实际上是后者的一种特例。因此，一般而言，我们采用自回归模型结构（但另见式（10）），这意味着模型的前向传播可写为

$$\tilde{\boldsymbol{C}}(t=t_{i+1}) = \mathrm{Model}(\tilde{\boldsymbol{C}}(t=t_i)), \tilde{\boldsymbol{C}}(t=t_0) = \boldsymbol{C}(t=t_0), \tag{16}$$

其中，Model 可以表示黑箱、白箱或灰箱右端项的积分。因此，两条时间序列之间的均方损失（MSE）可计算为

$$\mathrm{MSE}(\{\tilde{\boldsymbol{C}}(t=t_i)\}, \{\boldsymbol{C}(t=t_i)\}) = \frac{1}{KL}\sum_{j=1}^{L} ||\tilde{\boldsymbol{C}}(t=t_j) - \boldsymbol{C}(t=t_j)||_2^2, \tag{17}$$

其中，$L+1$ 是数据集 $\{\boldsymbol{C}(t=t_i)$ 的长度：$i=0,\ 1,\ 2,\ \cdots,\ L\}$ 且 $K=14$ 是状态向量的维度，如表1所示。

3 结果

本节将介绍本文中进行的表 2 所列各项计算实验。首先，我们将说明后续各项参数辨识和神经网络训练实验所采用的数据生成过程。我们还将分析内层优化问题中约束的影响，考察约束切换为激活状态（或从激活状态切换为非激活状态）的情况。接着，我们将通过一个黑箱示例开展实际训练实验。所有训练实验均包含动力学和化学计量变体。随后，我们将进行白箱辨识，分别采用两个自由参数和五个自由参数的变体。最后，我们将通过灰箱建模结合这两项任务：首先使用神经网络替换 $\boldsymbol{f}_{\mathbf{kin}}$中的一个动力学表达式；然后重复这一过程，同时允许将一个物理参数 $\boldsymbol{\alpha}$ 设为可训练参数。

实验	章节
数据生成	§3.1
黑箱	§3.2
白箱（2 个参数未知）	§3.3.1
白箱（5 个参数未知）	§3.3.2
灰箱（1 个表达式未知）	§3.4.1
灰箱（1 个表达式未知 + 1 参数未知）	§3.4.2

表2：本文计算实验的汇总。

3.1 数据生成

我们首先针对各种初始条件模拟短期轨迹，并将这些流收集为一个数据集 用于一组参数值。随后，我们严格按照第 2.1 节所述，在 PyTorch 中实现神经网络模型，并对其进行训练，使其匹配这些流。

该数据集由完整模型方程（ N_{run} ）的瞬态过程组成，或等价地，由方程（1）组成；其初始条件（IC）是在均值附近施加高斯扰动得到的，而均值本身则沿一条中心名义轨迹（NT）在时间上均匀随机采样。每个变量的标准差与该变量在 NT 中的变化范围成正比。也就是说，IC 样本表示为

$$\left\{ \boldsymbol{C}^{(i)}(t=0) = \begin{bmatrix} C_1^{(i)}(t=0) & \sim & \mathcal{N}(\bar{C}_1(t=t_i), \sigma_1) \\ & \vdots & \\ C_K^{(i)}(t=0) & \sim & \mathcal{N}(\bar{C}_K(t=t_i), \sigma_K) \end{bmatrix} \middle| i = 1, 2, \cdots, N_{\mathrm{run}} \right\}, \tag{18}$$

其中，名义轨迹 $\bar{\boldsymbol{C}}(t) = (\bar{C}_1(t)\ \bar{C}_2(t)\ \cdots,\ \bar{C}_K(t))^T$始于一组特定的初始条件，这些条件是在实验室实验中测得的。进料事件通过状态不连续来实现。此

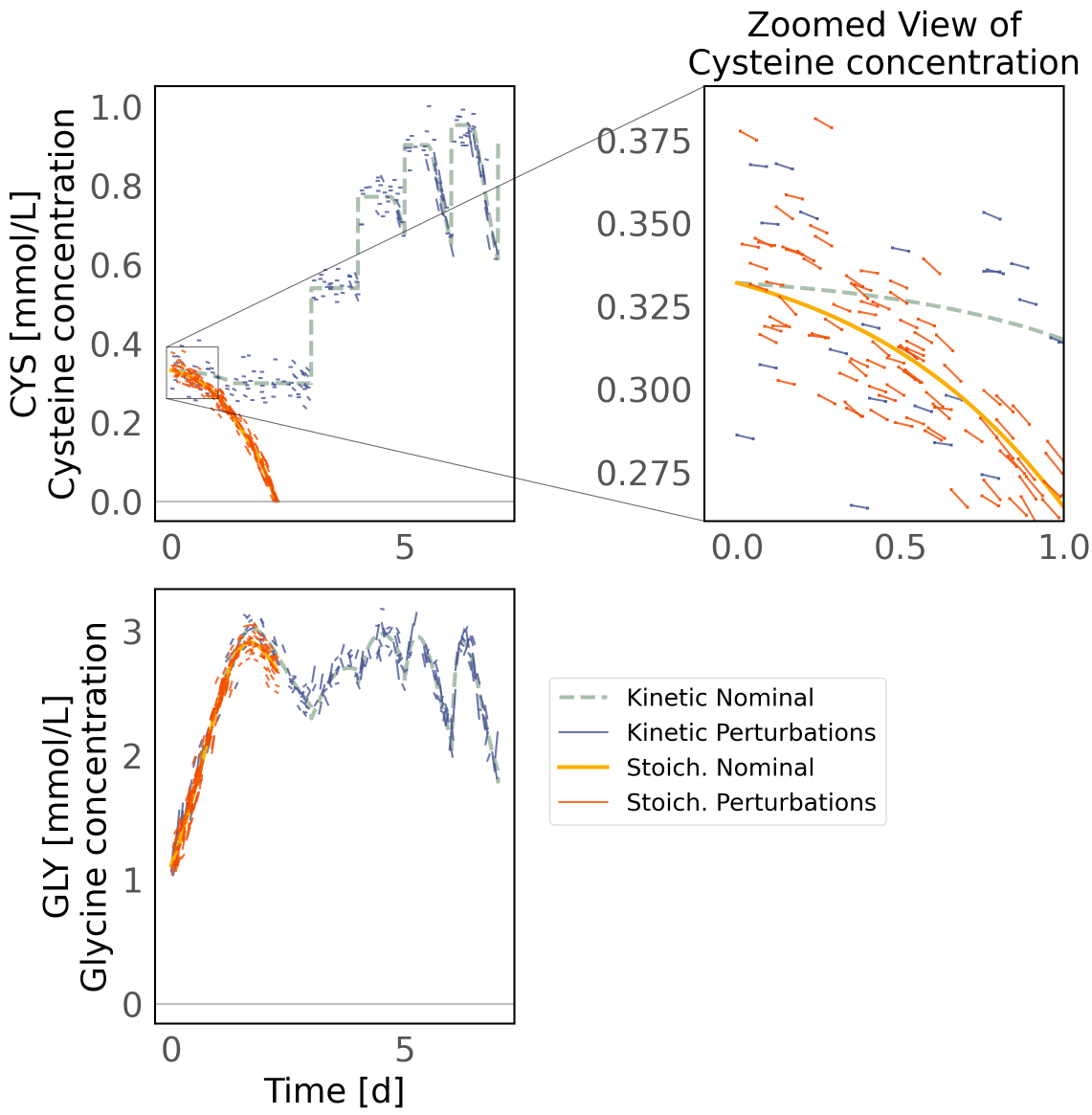


图2：名义轨迹（动力学和化学计量）与从扰动初始条件采样的短时流叠加。仅显示了少数 K 变量的轨迹。所有变量的轨迹见图15。图例中使用颜色区分不同的曲线。

名义轨迹（包括其“动力学”和“化学计量”积分）见图 2；“化学计量”NT 在状态变为负值之前不久停止（因此，在发生任何进料事件之前）。

随后对式（18）中的每个初始条件进行精确模拟（采用阶数为8(5,3）的显式龙格－库塔方法[39]，绝对容差为 10^{-8}，相对容差为 10^{-7}），模拟的时间范围通常显著短于整条名义轨迹（图4 和图5中的圆点），由此得到由若干轨迹“窗口”构成的一个“片段”的数据。

3.1.1 模拟过程中检测约束活动状态的转迁

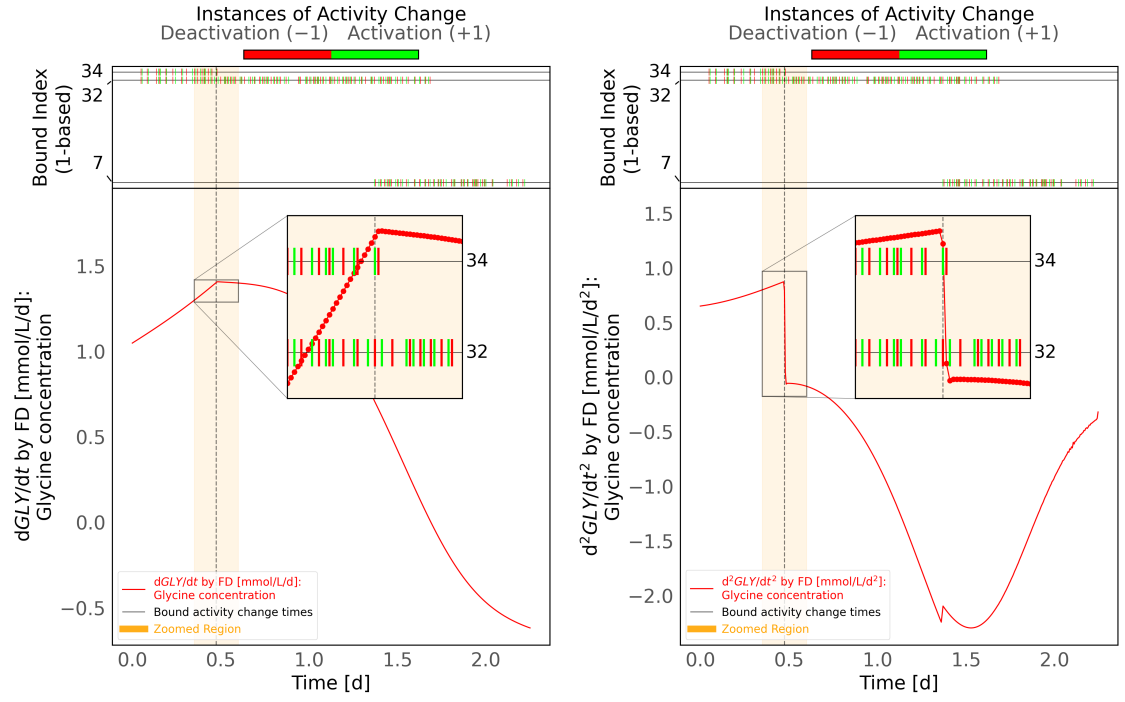
第 2.1 节中描述的内部优化包括对计算得出的一些通量施加边界（下界表示不可逆反应）。根据模拟变量 $\boldsymbol{C}$ 的当前状态，最优无约束

通量可能不在有界域内。此时，受约束的最优解将落在约束边界上（甚至可能落在多个约束边界的交点上）。

在此类事件发生之初或结束时，$\boldsymbol{v}$ 的轨迹可能会出现尖角。为了探究这一现象对系统动力学的影响，我们报告每个模拟时间步中哪些通量边界处于活跃状态，哪些处于非活跃状态。关于此类模拟的事件驱动版本，请参阅[30]，中的软件包，其中考虑了若干失效模式（除我们研究中出现的失效模式外），包括不可行的内部优化问题，以及存在多个解的问题（从而导致集值微分方程）。一般而言，事件可以是“时间事件”或“状态事件”；前者只需在特定时间准确停止积分。另一方面，当连续状态满足某些条件时，就会发生状态事件。在某些情况下，可以通过求解某些插值多项式的零点来检测此类事件 [30, 40]；如果事件条件无法表示为连续函数的根，则会带来进一步的困难（例如图3中的情况，其中一个布尔量会在事件发生时改变）。我们计划在今后的工作中探究如何恰当地分析此类事件检测问题（以及如何在事件之间进行积分；在每对事件之间，可能需要修改右端项求值，以确保活跃边界必然得到满足）。我们采用了一种较不复杂的可视化方法，如图3所示：图3展示了约束活跃状态的这些变化，并将其与两个关键模拟变量的一阶（3c、3a）和二阶（3d、3b）时间导数按时间对齐呈现。此处展示的是化学计量模型的一段短时模拟：约束转为活跃时以短暂的绿色刻度标记，转为非活跃时则以短暂的红色刻度标记。请注意二阶导数中的跳变，以及浓度演化的一阶导数中的尖角。

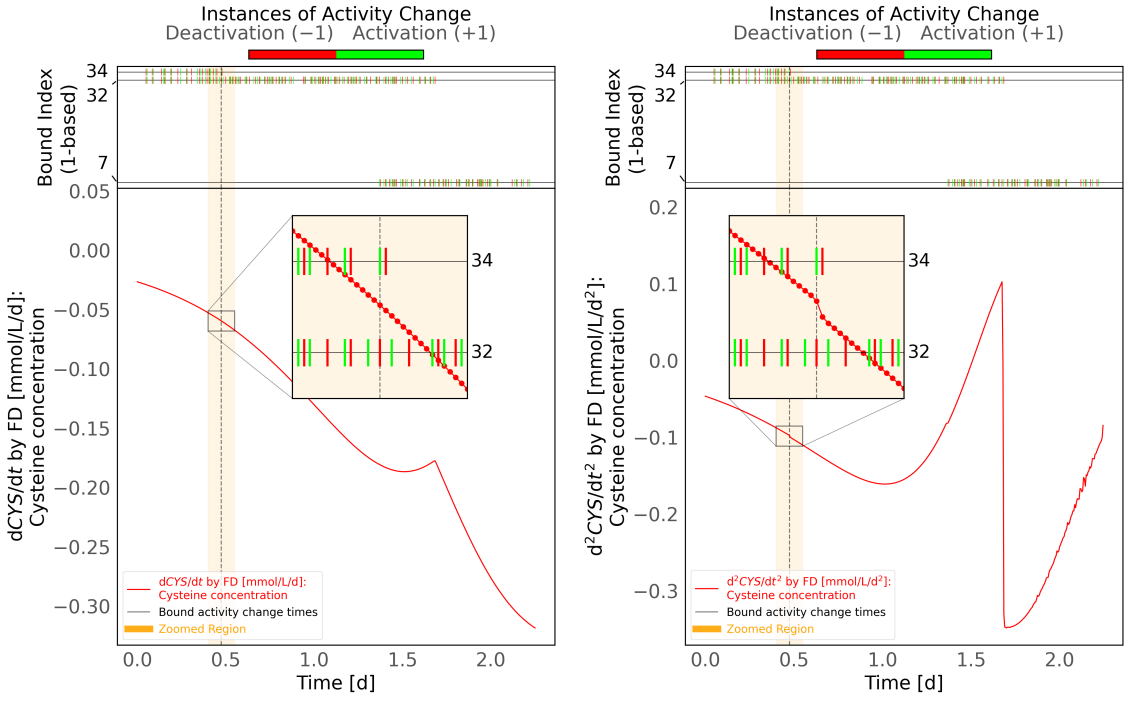
在这次运行中，我们观察到三个通量（7、32 和 34）触及了各自对应的下界。更具体地说，在 $t^* \approx 0.5$ 附近，当反应 34 的边界持续处于非活跃状态时（因此，相应的通量远离零），我们看到半胱氨酸和甘氨酸演化的二阶时间导数出现了明显的向下不连续。请注意，反应 34 是 NADH 的（不可逆）分解反应（参见附录F中的完整化学计量矩阵，或 [10] 中反应网络图的相关部分）；此时，其通量持续从 0 变为正值。因此，我们预计 $\mathrm{d}v_{34}/\mathrm{d}t$ 可能会在 t^* 处发生不连续。此外，生成 GLY 的一个（可逆）反应以 NADH 为底物。如果该反应速率在此时为正，那么它的一种底物会突然变得不那么充足。因此，由于 $\mathrm{dGLY}/\mathrm{d}t \sim -v_{34}$，且 $\mathrm{d}v_{34}/\mathrm{d}t$ 不连续，我们可以预期 $\mathrm{d}(\mathrm{dGLY}/\mathrm{d}t)/\mathrm{d}t$ 也会不连续，这在图 3b 中清晰可见。

对于半胱氨酸，也可以做出类似的解释：它同样依赖 NADH 作为底物，并且其二阶导数也存在不连续点（图 3d）。值得注意的是，半胱氨酸在通量 32 下界激活时出现的不连续幅度大得多。还要注意，甘氨酸后来还出现了第二个不连续点（约在 t 1.4）；这与通量 7 有关，不过涉及的是不同的代谢途径。



(a) 甘氨酸；一阶导数

(b) 甘氨酸；二阶导数



(c) 半胱氨酸；一阶导数

(d) 半胱氨酸；二阶导数

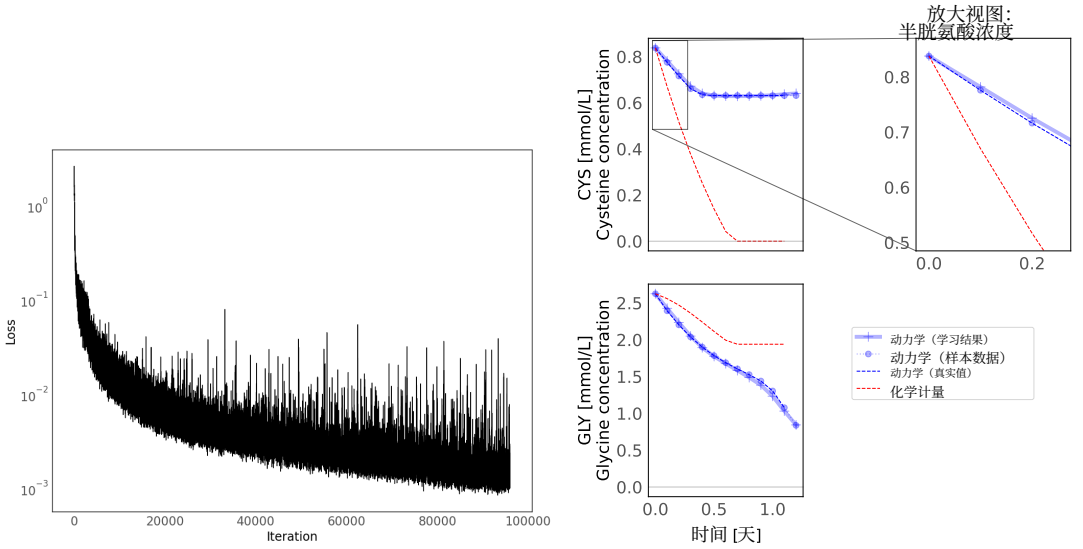
图 3： **化学计量情形下样本运行过程中活跃约束的变化**，观察浓度演变二阶导数（3b和 3d）中的不连续性。绘图所用的时间导数通过局部前向有限差分（FD）估算。使用颜色区分图例中描述的不同曲线。

3.2 黑箱神经网络辨识

为展示在不预设系统机制知识的情况下进行系统辨识，我们进行了黑箱右端项学习，将整个常微分方程系统表示为端到端神经网络。

鉴于如第 2.1 节所述，评估方程（8）有两种方法，我们也进行了两项实验：其中一项使用由真实值动力学系统生成的数据（每隔 $\Delta t = 0.1$ 小时采样一次，时间范围为 $t_{\max}$ =1。2小时，每条数据轨迹生成 13 个步长，共有 768 条数据轨迹），另一项使用由真实值化学计量系统生成的数据（其中 $\Delta t = 0.1$， $t_{\max} = 1.2$，包含 13 个步长，以及 768 条数据轨迹）。

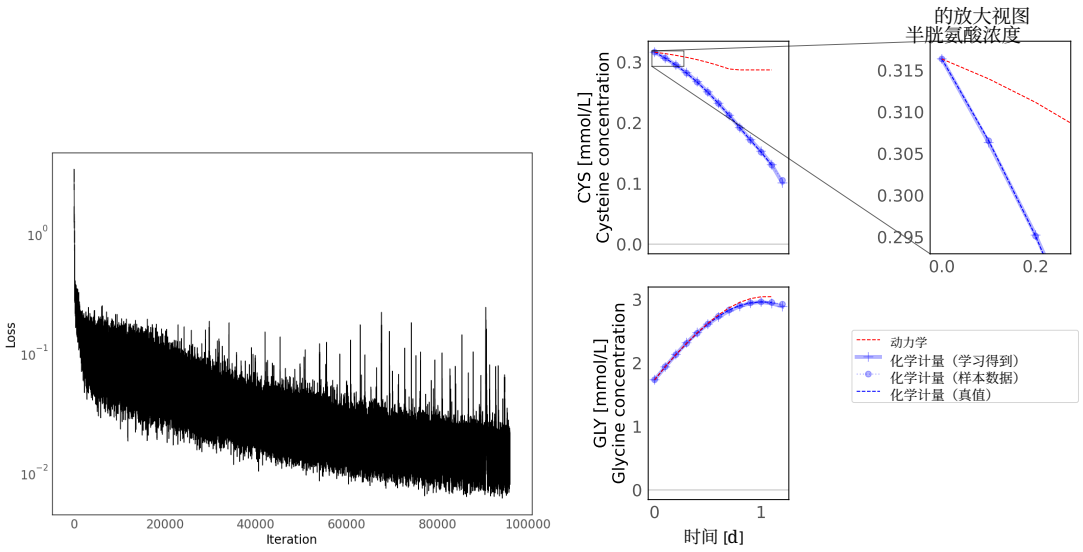
在这两种情况下，黑箱 ODE 都是通过在数据样本之间以固定步长 0.01 取点，并使用四阶龙格-库塔积分器进行训练的（黑盒辨识“不知道”模型中存在不连续性——它会在时间上的数据点之间进行平滑插值）。



(a) (b)

图4：黑箱训练结果（动力学）。另请参见图 16，查看更详细的结果。不同曲线以不同颜色区分，详见图例。

如图 4 和图5所示，我们的黑箱神经常微分方程能够非常紧密地拟合数据轨迹。这验证了底层方法，并表明，带有内部优化步骤的此类常微分方程可以成功近似为（可能略微“平滑化”的）闭式函数。



(a) (b)

图5：黑箱训练结果（化学计量）。另请参见图 17 ，了解更详细的结果。使用颜色区分图例中所述的不同曲线。

3.3 白箱神经网络：参数估计

为了展示我们称为“白箱”学习的全结构物理参数估计场景，我们尝试恢复（a）两个或（b）五个名义参数值。具体而言，我们在名义参数值下进行模拟，收集瞬态数据，选取其中四十三个（分别为四十个）作为已知数据，然后使用基于梯度的训练方法，根据这些数据估计其余两个（分别为五个）参数的数值。我们的初始猜测（对真实值进行扰动所得）标记在图 7中。对于这四次数值实验，数据集由 10 条短的单欧拉步长“轨迹”组成，每条轨迹长度为 0.05 小时（龙格–库塔积分得到了相近的结果，未显示）；网络设定同样采用欧拉方法，步长为 0.05。训练采用 RMSProp，共进行 4000 个训练轮次，每个训练轮次包含 1 个批次。

这展示了如何使用 [31] 对计算方程右端项这一内层优化问题进行微分；该问题在第 2.1节中讨论。这使得本实验能够进行基于梯度的学习，并在下文第3.4 节介绍灰箱方法之前，对这些算法进行初步验证。

3.3.1 已知模型，两个未知参数

在我们的首次白盒学习实验中，我们仅使用表4中的两个未知参数进行训练；我们发现（图6），能够较为准确地恢复这两个参数值。开展这一初步实验的一个动机，是帮助直观呈现图7中的梯度景观。我们从图8中看到，学习问题所诱导的梯度动力学具有很强的刚性，因此 Adam 等自适应训练方法 [41] 绝对不可或缺。训练过程呈现出两个阶段的下降：(a) 首先快速趋近参数空间中的深谷；随后 (b) 在谷底内缓慢移动，并因学习步长有限而产生一些振荡。此外，在化学计量（类型 2）情形下（图7b 和图8b），我们观察到最终梯度极其平缓，以至于即使使用 Adam，在损失谷底内移动任何显著距离也需要长得难以接受的时间。请注意，尽管真 $\boldsymbol{\alpha}$ 值确实是所设优化问题的一个极小值，如图7b所示，但这个极小值极其平缓，因此无法完美地发现真值 $\boldsymbol{\alpha}_2$ ——不过在实际应用中，整个“谷底”区域都能得到良好的拟合。这是所谓“模型松弛性”[42, 43]的一个例子；沿着谷底，所设损失函数对参数并不敏感，因此参数不可辨识。

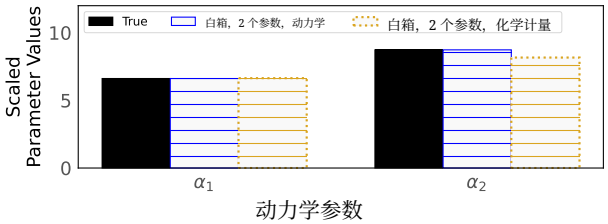
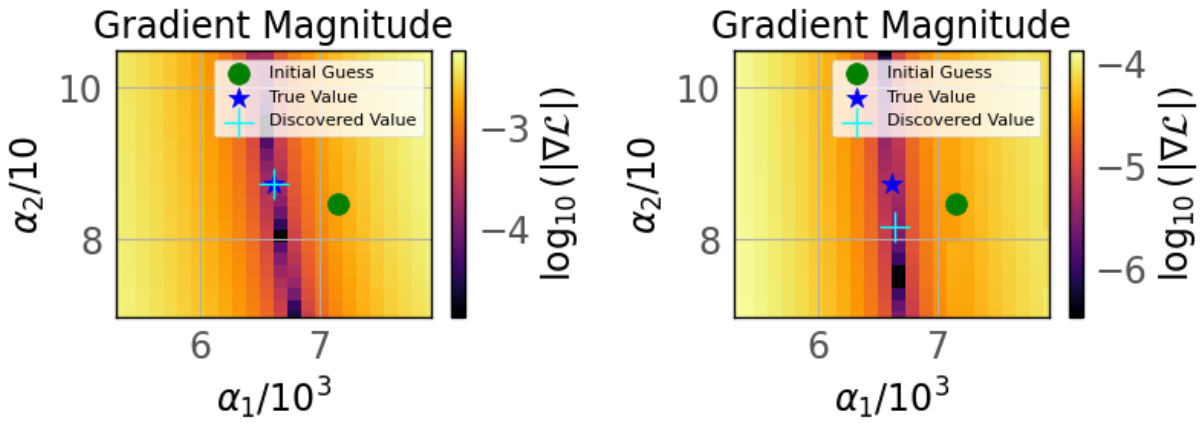
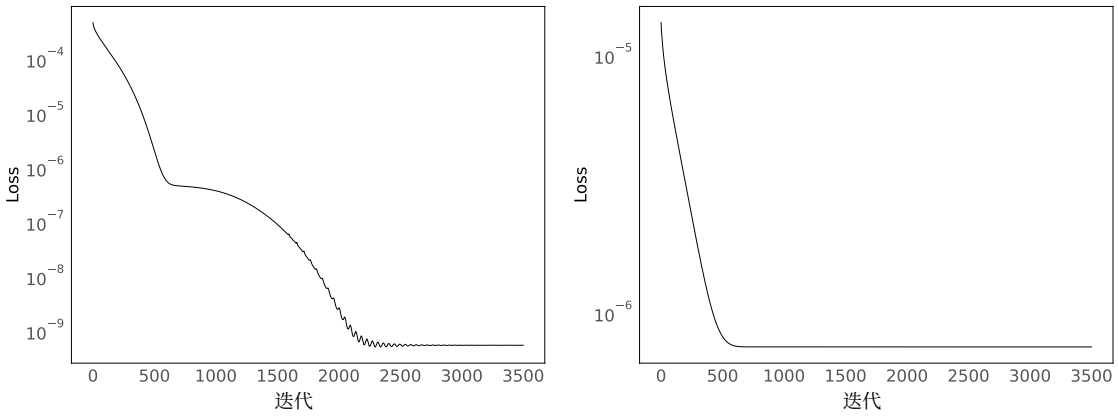


图 6：白箱两参数的参数比较。



(a) **动力学（类型1）：** 数据来自真实值动力学模拟；白箱假设同样是动力学模拟。(b) **化学计量（类型2）：** 数据来自真实值化学计量模拟；白箱假设同样是化学计量模拟。

图 7：**白箱双参数情况下的梯度景观。** 动力学（7a）与化学计量（7b）的实现，来自第 2.1 节。梯度向量场的刚性导致一定程度的参数不可辨识性。



(a) 动力学（类型1）　　(b) 化学计量（类型 2）

图 8：**白箱双参数情况下，训练收敛至最终参数估计值的过程。**动力学（8a）与化学计量（8b）实现，来自第 2.1 节。另见图 18。

3.3.2 已知模型，五个未知参数

接下来，我们重复了上一节的参数估计实验，不过这次使用五个而非两个未知 $\boldsymbol{\alpha}$ 值。我们发现，数值再次得到了较为准确的恢复（图9），且损失动态（图10）与双参数情形相似。

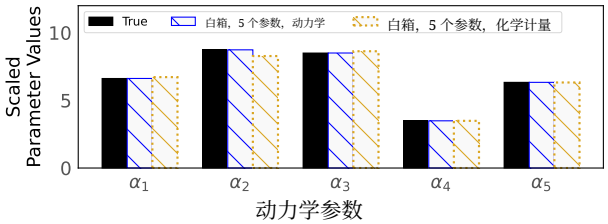
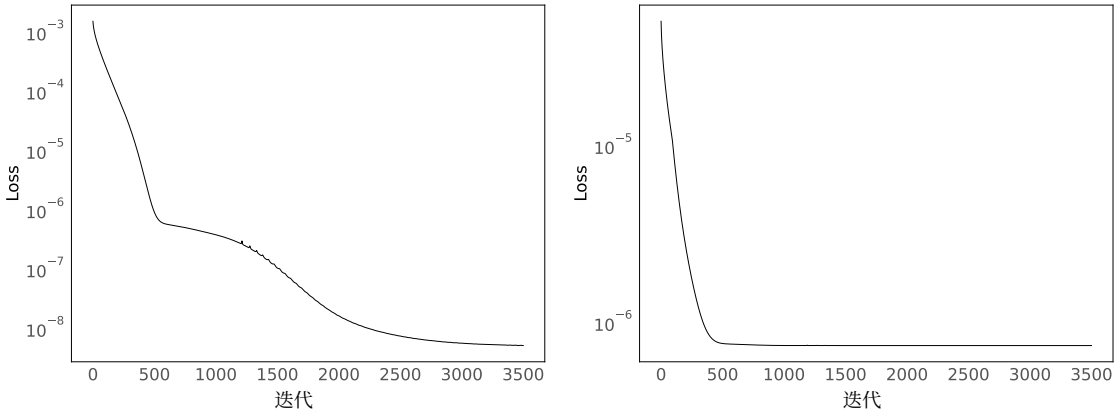


图9：白箱五参数情形的参数比较。



(a) 动力学（类型1）　　(b) 化学计量（类型 2）

图10：**白箱五参数情形下训练收敛至最终参数估计值的过程**。动力学（10a）与化学计量（10b）实现，见章节2.1。另见图19。

3.4 部分已知模型：灰箱辨识

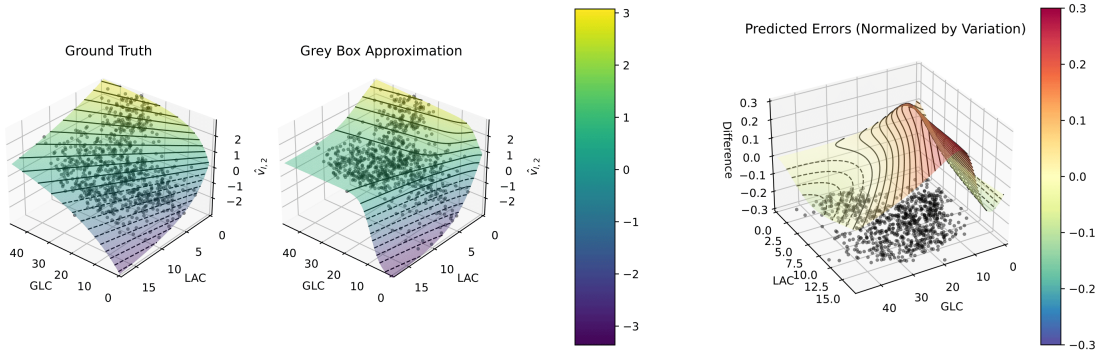
本节的工作中，我们假设 $\hat{v}_{I,2}$ in $\boldsymbol{f}_{\mathbf{kin}}$ 的表达式未知：我们只知道它是 GLC 和 LAC 的函数，因此用一个 2-8-8-1 多层感知机（MLP）替代它，该感知机具有可训练权重和偏置。我们将这个 MLP 嵌入灰箱计算图中（如图 1 所示），以生成用于损失评估的预测。

对于这四项实验，数据集由 800 条单欧拉步“轨迹”组成，每条轨迹长度为 0.05 小时。网络设定同样采用欧拉法，步长为 0.05。训练使用 RMSProp，共 500 个训练轮次，每个训练轮次包含 20 个批次。

3.4.1 部分已知模型，所有参数已知

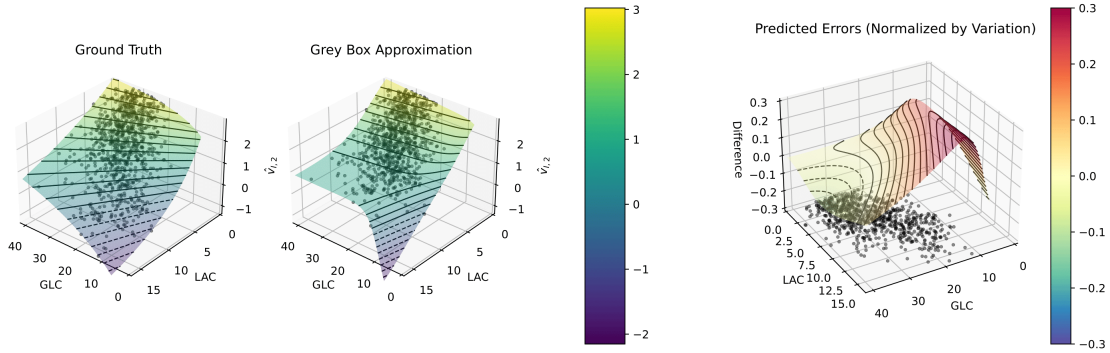
在我们的首批灰箱实验中，我们进一步假设，除用于 $\hat{v}_{I,\ 2}$ 的表达式外，所有表达式中使用的动力学参数数值均正确，无需校准。我们进行了两次实验：一次使用基于动力学的数据（来自类型1模拟的数据），且灰箱中的白箱部分也基于动力学表述；另一次则分别使用化学计量数据和相应的版本。

对于每项实验，我们发现（分别见图 11a 和图 12a），学到的通量函数大体上都得到了正确复现，但在其定义域的某些部分存在差异（网络预测相对平坦）。最大百分比差异（见图 11b 和图 12b；以训练数据中真值的变化范围为基准进行缩放）出现在通量近似为零的位置，这与预期相符。对此误差最显而易见的解释是，训练数据所依据的真实动力学很少经过该区域（低葡萄糖浓度），在化学计量情形下尤其如此。最后，小通量中的误差对我们实际在网络中最小化的量——浓度演化的预测误差——影响较小。



(a) (b)

图11：**通量比较（真实值与动力学灰箱模型）**。请注意，训练中访问的（葡萄糖、乳酸）点分散在曲面上（11a）或基准平面上（11b）。**11a：** 给定神经网络底物（葡萄糖和乳酸）输入时的真实值通量与神经网络近似通量。**11b：** 归一化预测误差（真实函数在数据范围内的最大值与最小值之差的分数）。请注意，（11a）与（11b）之间存在相对旋转（以便于观察）。



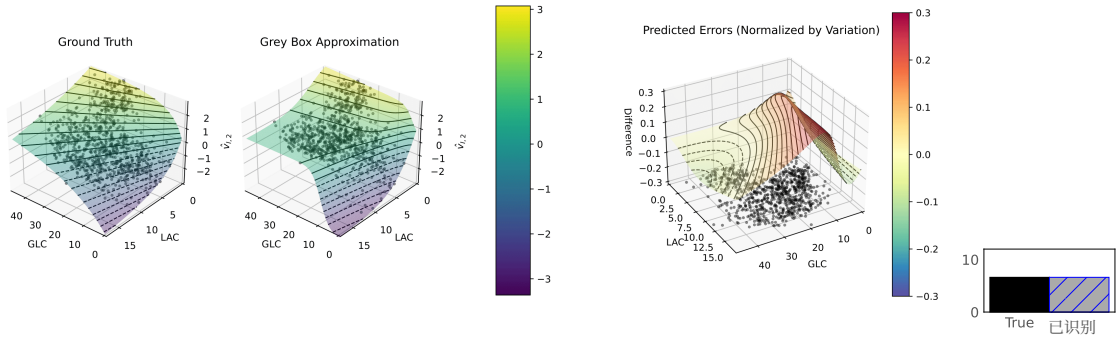
(a) (b)

图12：**通量对比（真实值与化学计量灰箱模型）**。12a：给定神经网络底物 GLC 和 LAC 时，通量的真实值与神经网络近似。12b：归一化预测误差（真实函数在数据范围内的最大值与最小值之差的比例）。请注意，训练中访问的 (GLC, LAC) 点分布在曲面上（12a）或底平面上（12b）。再次注意（12a）与（12b）之间的相对旋转（为便于观察）。

3.4.2 部分已知模型、部分已知参数

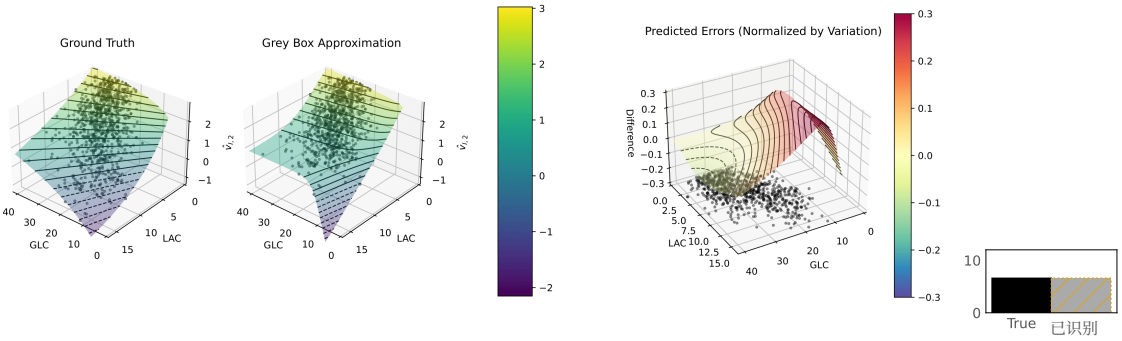
最后，我们将第3.3 节对白箱物理参数的拟合与第3.4.1节的灰箱拟合相结合，构建一个同时联合训练神经网络和物理组件的模型。在这项实验中，我们仍假设 $\boldsymbol{f}_{\mathbf{kin}}$ 中的 $\hat{}v_{I,\,2}$ 表达式未知；此外，我们还假设一个动力学参数 α_1 的值也需要校准。与之前一样，我们也研究了该实验的动力学版本和化学计量版本。

正如图 13、图 14 和表 4 所示，通量函数形状的恢复与第 3.4.1节具有相似的特征；然而，我们还能够准确地重新发现参数值，这展示了该方法在结合此类混合物理先验信息进行联合学习方面的潜力。



(a) (b) (c)

图13: **通量比较（动力学真实值与灰箱模型）。 13a：** 通量真实值与神经网络近似值，作为所学习动力学表达式输入——葡萄糖和乳酸——的函数。 **13b：** 归一化（如图 11b 和 12b所示）的预测误差。黑色数据点的绘制方式与图11 和12相同。 **13c：** 动力学参数的真实值与恢复值 α_1。



(a) (b) (c)

图14: **通量比较（化学计量真实值与灰箱模型）。 14a：** 通量真实值与神经网络近似值，作为所学习动力学表达式输入葡萄糖和乳酸的函数。**14b：** 归一化（如图 11b 和 12b所示）的预测误差。黑色数据点的绘制方式与图 11 和 12相同。 **14c：** 动力学参数的真实值与恢复值 α_1。

4 结论与未来展望

本文重新考察了一个描述中国仓鼠卵巢（CHO）细胞培养物中产生的生化反应的机理模型。在模拟该模型的动态过程时，对该方程组时间导数的计算实际上要求在每个时间步求解一个约束凸问题。这种“内部优化”可能导致：（a）演化浓度的二阶时间导数出现不连续性（即解本身为 C1，如图 3所示）；以及（b）难以计算系统雅可比矩阵，或难以计算演化状态相对于系统参数的敏感度梯度。

随后，我们展示了如何将模型的这类机理性物理知识与数据驱动方法相结合，以识别或校准此类系统。根据我们对物理定律的先验认知程度，我们的混合模型可以是黑箱、灰箱或白箱模型。重要的是，我们基于 [31] 对白箱和灰箱模型中的传统神经网络 /ODE-net 架构进行了改进：该方法可以将可微凸优化层嵌入数值积分器中（数值积分器可视为展开的循环神经网络），从而克服计算模型梯度时遇到的障碍。这类

数据驱动模型在从数据中识别代谢网络动力学并执行回归任务方面的潜力得到了展示。

我们在本文中设计和实现的方法及模型架构，应能广泛应用于演化方程右端本身包含优化问题的工程领域；机器人控制或可微模型预测控制 [44]便是其中的例子。在代谢工程领域，此类算法可与下游优化问题有效结合，用于实验设计、培养基组成优化，或生物反应器运行中最优补料／收获策略的设计。

致谢： 本研究部分得到了 AMBIC 的资助。T.B.、T.C. 和 I.G.K. 的研究也部分得到了 AFOSR 多学科大学研究计划的资助。C. M. 的研究部分得益于美国能源部先进生物能源与生物制品科学中心（美国能源部科学办公室生物与环境研究办公室，资助编号 DE-SC0018420）以及美国能源部科学办公室生物与环境研究办公室（资助编号 DE-SC0018260）的支持。

参考文献

[1] Michael Butler。《动物细胞培养：生物药品生产的近期成果与展望》。《应用微生物学与生物技术》，68:283–291，2005 年。[2] Ningning Ma、JoAnn Ellet、Centy Okediadi、Paul Hermes、Ellen McCormick 和 Susan Casnocha。单一营养物补料同时支持化学成分限定的 NS0 和 CHO 补料分批过程：提高生产率并改善乳酸代谢。《生物技术进展》，25(5):1353–1363，2009 年。[3] Costas D Maranas 和 Ali R Zomorrodi。《代谢网络优化方法》。John Wiley & Sons，2016 年。[4] Gregory N Stephanopoulos、Aristos A Aristidou 和 Jens Nielsen。《代谢工程：原理与方法》。Academic Press，1998 年 10 月。[5] Brett A Boghigian、Gargi Seth、Robert Kiss 和 Blaine A Pfeifer。代谢通量分析与药物生产。《代谢工程》，12(2):81–95，2010 年。[6] Chetan Goudar、Richard Biener、C Boisart、R¨udiger Heidemann、James Piret、Albert de Graaf 和 Konstantin Konstantinov。通过代谢物平衡以及 2d [13c、1h]cosy 核磁共振波谱，对灌流培养中的 CHO 细胞进行代谢通量分析。《代谢工程》，12(2):138–149，2010 年。[7] Lake-Ee Quek、Stefanie Dietmair、Jens O Kr¨omer 和 Lars K Nielsen。哺乳动物细胞培养中的代谢通量分析。《代谢工程》，12(2):161–171，2010 年。[8] Christophe Chassagnole、Naruemol Noisommit-Rizzi、Joachim W Schmid、Klaus Mauch 和 Matthias Reuss。大肠杆菌中心碳代谢的动态建模。《生物技术与生物工程》，79(1):53–73，2002 年。[9] Radhakrishnan Mahadevan、Jeremy S Edwards 和 Francis J Doyle。大肠杆菌二相生长的动态通量平衡分析。《生物物理学杂志》，83(3):1331–1340，2002 年。[10] Ryan P Nolan 和 Kyongbum Lee。CHO 细胞代谢的动态模型。《代谢工程》，13(1):108–124，2011 年。[11] Patrick F Suthers、Charles J Foster、Debolina Sarkar、Lin Wang 和 Costas D Maranas。利用化学计量平衡、热力学可行性及动力学定律形式体系，推进基于约束和机器学习的代谢建模：近期进展。《代谢工程》，63:13–33，2021 年。[12] RA Adomaitis、IG Kevrekidis、RM Farber、AS Lapedes、JL Hudson 和 M Kube。神经网络在真实实验数据的系统辨识与分岔分析中的应用。技术报告，洛斯阿拉莫斯国家实验室（LANL），美国新墨西哥州洛斯阿拉莫斯，1990 年。[13] JL Hudson、M Kube、RA Adomaitis、IG Kevrekidis、AS Lapedes 和 RM Farber。非线性信号处理与系统辨识：在电化学反应时间序列中的应用。《化学工程科学》，45(8):2075–2081，1990 年。[14] R Rico-Martines、IG Kevrekidis、MC Kube 和 JL Hudson。离散时间与连续时间非线性信号处理：吸引子、转迁及并行实现问题。载于*1993* 年美国控制会议，第 1475–1479 页。IEEE，1993 年。[15] J Nathan Kutz。《数据驱动建模与科学计算：复杂系统与大数据的方法》。牛津大学出版社，美国，2013 年。

[16] Steven L Brunton 和 J Nathan Kutz。《数据驱动科学与工程：机器学习、动力系统与控制》。剑桥大学出版社，第 2 版，2022 年。[17] K Krischer、IG Kevrekidis、MC Kube 和 JL Hudson。基于铜电溶解数据的离散时间与连续时间非线性信号处理。《化学工程通讯》，118(1):25–48，1992 年。[18] R Rico-Martinez、J S Anderson 和 I G Kevrekidis。连续时间非线性信号处理：一种基于神经网络的灰箱辨识方法。载于《IEEE 信号处理神经网络研讨会论文集》，第 596–605 页，1994 年。[19] Cristina P Martin-Linares、Yorgos M Psarellis、Georgios Karapetsas、Eleni D Koronaki 和 Ioannis G Kevrekidis。用于薄膜流动的物理无关与物理融入机器学习：基于小规模数据的建模与预测。arXiv 预印本 arXiv:2301.12508，2023 年。[20] Yorgos M Psarellis 、 Seungjoon Lee 、 Tapomoy Bhattacharjee 、 Sujit S Datta、Juan M Bello-Rivas 和 Ioannis G Kevrekidis。通过机器学习数据驱动地发现细菌的趋化迁移。arXiv 预印本 arXiv:2208.11853，2022 年。[21] Felix P Kemeth、Sergio Alonso、Blas Echebarria、Ted Moldenhawer、Carsten Beta 和 Ioannis G Kevrekidis。振幅方程的黑箱与灰箱学习：在相场系统中的应用。arXiv 电子预印本，页码 arXiv–2207，2022 年。[22] Felix P Kemeth、Tom Bertalan、Thomas Thiem、Felix Dietrich、Sung Joon Moon、Carlo R Laing 和 Ioannis G Kevrekidis。在学习得到的涌现空间中学习涌现偏微分方程。《自然·通讯》，13(1):3318，2022 年。[23] Seungjoon Lee、Yorgos M Psarellis、Constantinos I Siettos 和 Ioannis G Kevrekidis。根据基于智能体的蒙特卡洛模拟数据，学习趋化偏微分方程/闭合关系的黑箱与灰箱模型。arXiv 电子预印本，页码 arXiv–2205，2022 年。[24] Maziar Raissi、Paris Perdikaris 和 George E Karniadakis。物理信息神经网络：用于求解涉及非线性偏微分方程的正向与逆向问题的深度学习框架。《计算物理学杂志》，378:686–707，2019 年。[25] Alireza Yazdani、Lu Lu、Maziar Raissi 和 George Em Karniadakis。用于推断参数和隐含动态的系统生物学信息深度学习。《PLOS 计算生物学》，16(11):1–19，2020 年 11 月。[26] Mitchell Daneker、Zhen Zhang、George Em Karniadakis 和 Lu Lu。系统生物学：通过系统生物学信息神经网络进行可辨识性分析与参数辨识。arXiv 预印本 arXiv:2202.01723，2022 年。[27] Lu Lu、Xuhui Meng、Zhiping Mao 和 George Em Karniadakis。Deepxde：用于求解微分方程的深度学习库。《SIAM 评论》，63(1):208–228，2021 年。[28] George Em Karniadakis、Ioannis G Kevrekidis、Lu Lu、Paris Perdikaris、Sifan Wang 和 Liu Yang。物理信息机器学习。《自然·物理评论》，3(6):422–440，2021 年。[29] Paul I Barton 和 Cha Kun Lee。混合系统的建模、仿真、灵敏度分析与优化。《ACM 建模与计算机仿真汇刊》，12(4):256–289，2002 年 10 月。[30] J A Gomez、K H¨offner 和 P I Barton。Dfbalab：用于动态通量平衡分析的快速可靠 MATLAB 代码。《BMC 生物信息学》，15(409)，2014 年。[31] Brandon Amos 和 J Zico Kolter。Optnet：作为神经网络层的可微优化。载于《国际机器学习会议》，2017 年。[32] Steven Diamond 和 Stephen Boyd。CVXPY：一种嵌入 Python 的凸优化建模语言。《机器学习研究杂志》，17(83):1–5，2016 年。[33] Akshay Agrawal、Robin Verschueren、Steven Diamond 和 Stephen Boyd。凸优化问题的重写系统。《控制与决策杂志》，5(1):42–60，2018 年。[34] Akshay Agrawal、Brandon Amos、Shane Barratt、Stephen Boyd、Steven Diamond 和 J Zico Kolter。可微凸优化层。载于《神经信息处理系统进展》，第 9558–9570 页，2019 年。[35] A Agrawal、S Barratt、S Boyd、E Busseti 和 W Moursi。通过锥规划进行微分。《应用与数值优化杂志》，1(2):107–115，2019 年。[36] A Agrawal、S Barratt、S Boyd、E Busseti 和 W Moursi。diffcp：通过锥规划进行微分，版本 1.0.19。https://github.com/cvxgrp/diffcp，2019 年。

[37] Brandon Amos. 面向机器学习的基于可微优化的建模。博士论文，卡内基梅隆大学，2019年5月。[38] Anirudh Goyal、Alex Lamb、Ying Zhang、Saizheng Zhang、Aaron Courville 和 Yoshua Bengio。Professor forcing：一种训练循环网络的新算法。载于神经信息处理系统第*30*届国际会议论文集，NIPS'16，第4608–4616页，美国纽约州雷德胡克，2016年。Curran Associates Inc. [39] E Hairer、S P Nørsett 和 G Wanner。常微分方程求解 *I*：非刚性问题。Springer，德国柏林、海德堡，1993年。[40] P.I. Barton。混合系统的建模、仿真与灵敏度分析。载于*CACSD* 会议论文集：*IEEE*计算机辅助控制系统设计国际研讨会（编号*00TH8537*），第117–122页，2000年。[41] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*CoRR*，abs/1412.6980，2014年。[42] Bryan C Daniels、Yan-Jiun Chen、James P Sethna、Ryan N Gutenkunst 和 Christopher R Myers。系统生物学中的松弛性、鲁棒性与可进化性。生物技术当前观点，19(4):389–395，2008年。蛋白质技术／系统生物学。[43] Alexander Holiday、Mahdi Kooshkbaghi、Juan M. Bello-Rivas、C. William Gear、Antonios Zagaris 和 Ioannis G. Kevrekidis。用于参数降维的流形学习。计算物理学杂志，392:419–431，2019年。[44] Brandon Amos、Ivan Jimenez、Jacob Sacks、Byron Boots 和 J. Zico Kolter。用于端到端规划与控制的可微MPC。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编，神经信息处理系统进展，第31卷。Curran Associates, Inc.，2018年。

物理变量和动力学参数

变量	C_1	C_2	C_3	C_4	C_5	C_6	C_7
缩略语	BIOM	ANTI	GLC	LAC	ALA	ASN	ASP
名称	生物量	抗体	葡萄糖	乳酸	丙氨酸	天冬酰胺	天冬氨酸
单位	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升
样本数值	1.18	0.0801	30.9	1.36	0.354	5.23	1.21
动力学 $\mathrm{d}C_i/\mathrm{d}t$	2.02	0.0361	−1.76	1.23	0.179	−0.152	−0.097
化学计量 $\mathrm{d}C_i/\mathrm{d}t$	2.04	0.0668	−1.83	1.2	0.142	−0.153	−0.151
动力学 - 化学计量	−0.0116	−0.0307	0.07	0.0336	0.0372	0.00153	0.0536
样本数值	0.332	4.34	1.14	7.74	3.17	2.68	0.483
动力学 $\mathrm{d}C_i/\mathrm{d}t$	−0.00657	−0.997	1.3	−0.904	0.181	0.482	0.875
化学计量 $\mathrm{d}C_i/\mathrm{d}t$	−0.0274	−0.999	1.05	−0.764	0.163	0.679	0.88
动力学 - 化学计量	0.0208	0.00239	0.242	−0.14	0.0179	−0.197	−0.00501

表3：所有变量的名称和单位 $\boldsymbol{C} = (C_1, C_2, \cdots, C_{14})^T$。

	编号符号	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9
	物理符号	v_{max1}	K_{i1}	K_{m1}	v_{max2f}	K_{m2glc}	v_{max3f}	v_{max3r}	K_{m3ala}	K_{m3glc}
	真实值	6617.8	87.349	84.982	3490.4	6.3331	950.80	949.28	0.2165	2.0026
§3.3.1	值（白箱 2P 动能）	6617.2	87.354							
	值（白箱 2P 化学计量）	6643.8	81.666							
§3.3.2	值（白箱 5P 动能）	6621	87.371	85.025	3489.9	6.3337				
	值（白箱5P 化学计量）	6720.2	82.774	86.327	3487	6.3236				
§3.4.2	值（GB 动能）	6617.9								
	值（GB 化学计量）	6594.1								
	单位	毫摩尔/天	毫摩尔/升	毫摩尔/升	毫摩尔/天	毫摩尔/升	毫摩尔/天	毫摩尔/天	毫摩尔/升	毫摩尔/升

	编号符号	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}
	物理符号	v_{max8f}	v_{max8r}	K_{m8gln}	K_{m8glu}	K_{m8nh3}	v_{max9f}	v_{max9r}	K_{m9nh3}	v_{max10f}
	真实值	9568.5	415.83	5.9198	2.0582	7.5053	3.3291	7.8132	0.6866	143.32
	单位	毫摩尔/天	毫摩尔/天	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/天	毫摩尔/天	毫摩尔/升	毫摩尔/天

	编号符号	α_{19}	α_{20}	α_{21}	α_{22}	α_{23}	α_{24}	α_{25}	α_{26}	α_{27}
	物理符号	v_{max10r}	K_{m10asn}	K_{m10asp}	K_{m10nh3}	v_{max11}	v_{max12f}	v_{max12r}	K_{m12ser}	K_{m12gly}
	真实值	95.194	0.0157	3.5060	0.6301	0.5465	0.6330	92.978	3.0862	0.2020
	单位	毫摩尔/天	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/天	1	毫摩尔/天	毫摩尔/升	毫摩尔/升

	编号符号	α_{28}	α_{29}	α_{30}	α_{31}	α_{32}	α_{33}	α_{34}	α_{35}	α_{36}
	物理符号	v_{max13}	K_{m13}	v_{max16}	K_{m16a}	K_{m16b}	v_{max17}	K_{i17}	v_{max2r}	K_{m2lac}
	真实值	72.593	1.4396	6316.7	0.9967	0.9901	76.88	46.045	3996.2	4.3040
	单位	毫摩尔/天	毫摩尔/升	毫摩尔/天	毫摩尔/升	毫摩尔/升	毫摩尔/天	毫摩尔/升	毫摩尔/天	毫摩尔/升

	编号符号	α_{37}	α_{38}	α_{39}	α_{40}	α_{41}	α_{42}	α_{43}	α_{44}	α_{45}
	物理符号	K_{m9glu}	K_{m9gln}	K_{m11asp}	K_{m11asn}	K_{m12nh3}	K_{m16c}	v_{max35}	K_{m35a}	K_{m35b}
	真实值	0.7519	4.1330	6.4439	6.5433	0.2607	0.0108	2.371	6.8779	9.1087
	单位	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/升	毫摩尔/天	毫摩尔/升	毫摩尔/升

表4：所有参数的名称和真实值 $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_{45})^T$，以及白箱模型和灰箱模型发现的若干参数值。此处，“2PK”等术语的含义可在表2中查阅。

B 常微分方程表达式 $\mathrm{d}\boldsymbol{C}\,\mathrm{d}t \equiv \boldsymbol{f}_{\mathbf{eqn:ode}}(\boldsymbol{C};\ \boldsymbol{v}(\boldsymbol{C};\boldsymbol{\alpha}))$

$$
\begin{aligned}
\frac{\mathrm{d}C_1}{\mathrm{d}t} &= v_{18}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_2}{\mathrm{d}t} &= v_{19}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_3}{\mathrm{d}t} &= -v_{20}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_4}{\mathrm{d}t} &= v_{21}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_5}{\mathrm{d}t} &= v_{22}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_6}{\mathrm{d}t} &= -v_{23}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_7}{\mathrm{d}t} &= v_{24}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_8}{\mathrm{d}t} &= -v_{25}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_9}{\mathrm{d}t} &= -v_{26}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_{10}}{\mathrm{d}t} &= v_{27}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_{11}}{\mathrm{d}t} &= -v_{28}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_{12}}{\mathrm{d}t} &= v_{29}\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_{13}}{\mathrm{d}t} &= (v_{25}+v_{33})\frac{C_{14}}{1000}\\
\frac{\mathrm{d}C_{14}}{\mathrm{d}t} &= \frac{\frac{\mathrm{d}C_1}{\mathrm{d}t}}{2.31} - 0.003C_{14}
\end{aligned}
$$

C 所有反应列表

ID	Kin.	化学计量。	动力学 - 化学计量。	反应
1	1.18	2.22	−1.04	葡萄糖-6-磷酸 → 2丙酮酸 + 3三磷酸腺苷 + 2细胞质 NADH
2	0.0801	−0.901	0.981	丙酮酸 + 细胞质 NADH ↔ 乳酸
3	30.9	28.6	2.32	丙酮酸 + 谷氨酸 ↔ 丙氨酸 +α-酮戊二酸
4	1.024×10^{-12}	1.153×10^{-15}	1.023×10^{-12}	丙酮酸 + 草酰乙酸 → α-酮戊二酸 + $2CO_2$ + 2线粒体 NADH
5	22.5	20.4	2.06	α-酮戊二酸 → 苹果酸 + CO_2 + 线粒体 NADH + $FADH_2$ + 三磷酸腺苷
6	1.151×10^{-13}	7.718×10^{-16}	1.144×10^{-13}	苹果酸 → 草酰乙酸 + 线粒体 NADH
7	25.7	22.3	3.35	苹果酸 → 丙酮酸 + CO_2
8	1.36	1.36	-2.852×10^{-5}	谷氨酰胺 ↔ 谷氨酸 + NH_3
9	0.354	3.78	−3.42	α-酮戊二酸 + NH_3 + 线粒体 NADH ↔ 谷氨酸
10	5.23	5.23	-3.530×10^{-5}	天冬酰胺 ↔ 天冬氨酸 + NH_3
11	1.21	1.82	−0.61	天冬氨酸 + α-酮戊二酸 ↔ 草酰乙酸 + 谷氨酸
12	0.332	−0.597	0.929	丝氨酸 + CO_2 + NH_3 + 细胞质 NADH ↔ 2甘氨酸
13	4.34	3.41	0.929	胱氨酸 + 细胞质 NADH → 2半胱氨酸
14	14.8	19.4	−4.55	线粒体 NADH + $0.5O_2$ → 2.5三磷酸腺苷
15	16.6	20.5	−3.89	$FADH_2$ + $0.5O_2$ → 1.5三磷酸腺苷
16	1.14	4.27	−3.13	0.0838丙氨酸 + 0.041天冬酰胺 + 0.0804天冬氨酸 + 8.6825三磷酸腺苷 + 0.0261半胱氨酸 + 0.452葡萄糖-6-磷酸 +0.0873谷氨酰胺 + 0.056甘氨酸 + 0.427草酰乙酸 + 0.096丝氨酸 → BIOM + 0.004$FADH_2$ +0.0082谷氨酸 + 0.4445苹果酸 + 0.6391线粒体 NADH + 0.2085丙酮酸
17	7.74	7.34	0.399	0.0614丙氨酸 + 0.0344天冬酰胺 + 0.0389天冬氨酸 + 9.2三磷酸腺苷 + 0.024半胱氨酸 +0.0479谷氨酸 + 0.0449谷氨酰胺 + 0.0719甘氨酸 + 0.1丝氨酸 → ANTI
18	3.17	4.27	−1.1	BIOM →BIOM(细胞外)
19	2.68	7.34	−4.66	ANTI →ANTI(细胞外)
20	4.28	2.89	1.39	葡萄糖(细胞外) + 三磷酸腺苷 → 葡萄糖-6-磷酸
21	0.0801	−0.901	0.981	乳酸 ↔乳酸(细胞外)
22	30.3	27.8	2.56	丙氨酸 ↔丙氨酸(细胞外)
23	5.54	5.65	−0.115	天冬酰胺(细胞外) → 天冬酰胺
24	3.62	2.78	0.846	天冬氨酸 ↔天冬氨酸(细胞外)
25	5.756×10^{-15}	3.41	−3.41	胱氨酸(细胞外) + 谷氨酸 → 胱氨酸 + 谷氨酸(细胞外)
26	1.8	2.06	−0.255	谷氨酰胺(细胞外) ↔谷氨酰胺
27	0.0431	−1.96	2.01	甘氨酸 ↔甘氨酸(细胞外)
28	0.249	3.07	−2.82	丝氨酸(细胞外) →丝氨酸
29	5.9	3.4	2.5	NH_3 ↔NH_3(细胞外)
30	19.9	23.2	−3.26	O_2(细胞外) ↔ O_2
31	47.8	43.4	4.47	CO_2 →CO_2(细胞外)
32	4.23	3.27	0.965	2半胱氨酸 + O_2 → 2胱氨酸
33	8.539×10^{-15}	−22.8	22.8	谷氨酸 →谷氨酸(细胞外)
34	−5.6	4.368×10^{-15}	−5.6	细胞质 NADH → 0.5线粒体 NADH + 0.5$FADH_2$
35	0.483	−1.26	1.75	葡萄糖-6-磷酸 + 三磷酸腺苷 + 2谷氨酸 → 2细胞质 NADH + 2丝氨酸 + 2α-酮戊二酸

表 5：代谢网络中所有反应的表达式。请注意，我们区分胞质和线粒体 NADH，而所有代谢物均位于细胞内，除非另有说明。带有“↔”的反应是可逆的，而“→”仅用于不可逆反应。编号的颜色如附录 F所述。通量（数值列）是在与表 3相同的状态下计算的。

D 动力学表达式 $\hat{\boldsymbol{v}}_I = \boldsymbol{f}_{\text{kin}}(C;\boldsymbol{\alpha})$

$$\hat{v}_{I,1} = \frac{v_{max1}\frac{\text{GLC}}{K_{m1}}}{1+\frac{\text{LAC}}{K_{i1}}+\frac{\text{GLC}}{K_{m1}}+\frac{\text{LAC}}{K_{i1}}\frac{\text{GLC}}{K_{m1}}}$$

$$\hat{v}_{I,2} = \frac{v_{max2f}\frac{\text{GLC}}{K_{m2glc}} - v_{max2r}\frac{\text{LAC}}{K_{m2lac}}}{1+\frac{\text{GLC}}{K_{m2glc}}+\frac{\text{LAC}}{K_{m2lac}}}$$

$$\hat{v}_{I,3} = \frac{v_{max3f}\frac{\text{GLC}}{K_{m3glc}} - v_{max3r}\frac{\text{ALA}}{K_{m3ala}}}{1+\frac{\text{GLC}}{K_{m3glc}}+\frac{\text{ALA}}{K_{m3ala}}}$$

$$\hat{v}_{I,4} = \frac{v_{max8f}\frac{\text{GLN}}{K_{m8gln}} - v_{max8r}\frac{\text{GLU}}{K_{m8glu}}\frac{\text{NH3}}{K_{m8nh3}}}{1+\frac{\text{GLN}}{K_{m8gln}}+\frac{\text{GLU}}{K_{m8glu}}+\frac{\text{NH3}}{K_{m8nh3}}+\frac{\text{GLU}}{K_{m8glu}}\frac{\text{NH3}}{K_{m8nh3}}}$$

$$\hat{v}_{I,5} = \frac{v_{max9f}\frac{\text{NH3}}{K_{m9nh3}} - v_{max9r}(\frac{\text{GLU}}{K_{m9glu}}+\frac{\text{GLN}}{K_{m9gln}})}{1+\frac{\text{NH3}}{K_{m9nh3}}+\frac{\text{GLU}}{K_{m9glu}}+\frac{\text{GLN}}{K_{m9gln}}+\frac{\text{GLU}}{K_{m9glu}}\frac{\text{GLN}}{K_{m9gln}}}$$

$$\hat{v}_{I,6} = \frac{v_{max10f}\frac{\text{ASN}}{K_{m10asn}} - v_{max10r}\frac{\text{ASP}}{K_{m10asp}}\frac{\text{NH3}}{K_{m10nh3}}}{1+\frac{\text{ASN}}{K_{m10asn}}+\frac{\text{ASP}}{K_{m10asp}}+\frac{\text{NH3}}{K_{m10nh3}}+\frac{\text{ASP}}{K_{m10asp}}\frac{\text{NH3}}{K_{m10nh3}}}$$

$$\hat{v}_{I,7} = \frac{v_{max11}(\frac{\text{ASP}}{K_{m11asp}}+\frac{\text{ASN}}{K_{m11asn}})}{1+\frac{\text{ASN}}{K_{m11asn}}+\frac{\text{ASP}}{K_{m11asp}}+\frac{\text{ASP}}{K_{m11asp}}\frac{\text{ASN}}{K_{m11asn}}}$$

$$\hat{v}_{I,8} = \frac{v_{max12f}\cdot\hat{v}_{I,10}\frac{\text{SER}}{K_{m12ser}}\frac{\text{NH3}}{K_{m12nh3}} - v_{max12r}(\frac{\text{GLY}}{K_{m12gly}})^2}{1+\frac{\text{SER}}{K_{m12ser}}+\frac{\text{GLY}}{K_{m12gly}}+\frac{\text{NH3}}{K_{m12nh3}}+(\frac{\text{GLY}}{K_{m12gly}})^2}$$

$$\hat{v}_{I,9} = \frac{v_{max13}\frac{\text{CYS}}{K_{m13}}}{1+\frac{\text{CYS}}{K_{m13}}}$$

$$\hat{v}_{I,10} = \frac{v_{max16}\frac{\text{GLN}}{K_{m16a}}\frac{\text{ASN}}{K_{m16b}}\frac{\text{ALA}}{K_{m16c}}}{1+\frac{\text{GLN}}{K_{m16a}}+\frac{\text{ASN}}{K_{m16b}}+\frac{\text{ALA}}{K_{m16c}}+\frac{\text{GLN}}{K_{m16a}}\frac{\text{ASN}}{K_{m16b}}+\frac{\text{GLN}}{K_{m16a}}\frac{\text{ALA}}{K_{m16c}}+\frac{\text{ASN}}{K_{m16b}}\frac{\text{ALA}}{K_{m16c}}+\frac{\text{GLN}}{K_{m16a}}\frac{\text{ASN}}{K_{m16b}}\frac{\text{ALA}}{K_{m16c}}}$$

$$\hat{v}_{I,11} = \frac{v_{max17}}{1+\frac{\text{LAC}}{K_{i17}}}$$

$$\hat{v}_{I,12} = \hat{v}_{I,10}$$

$$\hat{v}_{I,13} = \hat{v}_{I,11}$$

$$\hat{v}_{I,14} = \frac{v_{max35}\frac{\text{GLC}}{K_{m35a}}(\frac{\text{GLU}}{K_{m35b}})^2}{1+\frac{\text{GLC}}{K_{m35a}}+\frac{\text{GLU}}{K_{m35b}}+(\frac{\text{GLU}}{K_{m35b}})^2}$$

E 动力学表达式与 [10] 的偏差

所有由动力学定义的细胞内反应均采用 Michaelis-Menten 形式进行建模，依据的是 Nolan and Lee [10] 提出的动力学表达式。部分修改总结如下，详情将在另一篇论文中说明（Khare等人，即将发表）。

用于缩放最大正向反应速率 - v_{max} 的温度依赖常数 TC，以及抑制指数常数 exp_i，均已从动力学表达式中移除。此外，用于在反应动力学中反映细胞氧化还原状态的氧化还原变量 R 也已移除。取而代之的是，反应仅依据细胞外代谢物进行建模，其消耗速率和生成速率取决于相关细胞外代谢物的浓度。根据文献和 $_1$3C标记（细胞内和细胞外）代谢物示踪数据，修订了反应化学计量关系，并纳入了额外反应。修改原因如下：

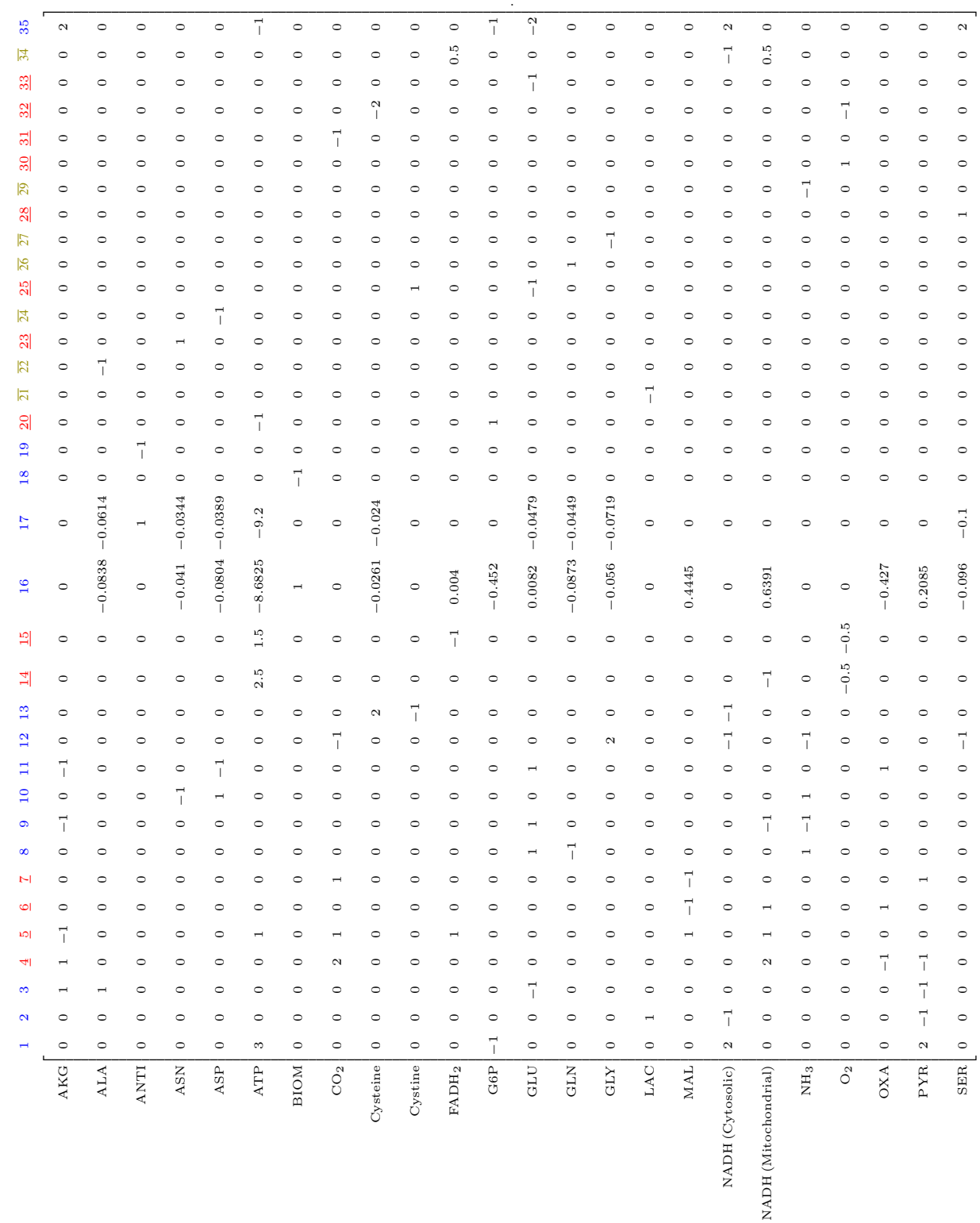
- 反应 1（对应于 $\hat{v}_{I,\ 1}$）：G6P $\rightarrow$ 2PYR $+$ 3ATP $+$ 2NADH 从动力学表达式中移除了温度相关常数 TC_1 和抑制指数常数 exp_1 （后者也与温度相关）。

- 反应 2（对应于 $\hat{v}_{I,\ 2}$）：PYR $+$NADH $\leftrightarrow$ LAC 移除了氧化还原变量 R，并将反应改为连续反应。
- 反应 3（对应于 $\hat{v}_{I,\ 3}$）：PYR $+$GLU $\leftrightarrow$ ALA $+$AKG 移除了温度常数 TC_{3b}。
- 反应 8（对应于 $\hat{v}_{I,\ 4}$）：GLN $\leftrightarrow$ GLU $+$$NH_3$移除了温度常数 TC_8。
- 反应 9（对应于 $\hat{v}_{I,\ 5}$）：AKG $+$ NH_3 $+$ NADH $\leftrightarrow$ GLU 移除了对细胞内代谢物 AKG（未进行动态追踪）的依赖，使反应表达式符合 Michaelis-Menten动力学。

- 反应10（对应于$\hat{v}_{I,\ 6}$）：天冬酰胺 $\leftrightarrow$ 天冬氨酸 $+$ NH_3动力学表达式中移除了温度相关常数 TC_{10}。
- 反应11（对应于$\hat{v}_{I,\ 7}$）：天冬氨酸 $+$ α-酮戊二酸 $\leftrightarrow$ 草酰乙酸 $+$ 谷氨酸。该反应表达式已改为不可逆形式。
- 反应12（对应于$\hat{v}_{I,\ 8}$）：丝氨酸 $+$$CO_2$ $+$ NH_3 $+$ NADH $\leftrightarrow$ 2甘氨酸。除丝氨酸外，正向反应还依赖于氨。
- 反应13（对应于$\hat{v}_{I,\ 9}$）：胱氨酸 $+$NADH $\rightarrow$ 2半胱氨酸。已移除温度相关变量 r。
- 反应16（对应于$\hat{v}_{I,\ 10}$）：生物量生成。已移除温度常数 TC_{16} 及温度依赖性。
- 反应17（对应于$\hat{v}_{I,\ 11}$）：抗体合成。已移除指数抑制常数 exp_{17}。
- 反应33：由于我们仅将细胞内反应定义为米氏-门腾表达式，而未将转运反应纳入其中，因此从我们的表达式中移除了该反应。所有转运反应均按化学计量计算。
- 反应35（对应于 $\hat{v}_{I,\ 14}$）：葡萄糖-6-磷酸 $+$ 三磷酸腺苷 $+$ 2谷氨酸 $\rightarrow$ 2细胞质 NADH $+$ 2丝氨酸 $+$ 2α-酮戊二酸。该反应根据糖酵解中存在丝氨酸合成途径（SSP）而添加。

F 化学计量矩阵S

蓝色：细胞内通量的索引橄榄色上划线：可逆细胞外通量的索引红色下划线：不可逆细胞外通量的索引



G 基于化学计量的方法的闭式解（无不等式界限）

方程 eq. (6) 的目标函数可改写为:

$$||\boldsymbol{v}_I - \hat{\boldsymbol{v}}_I||_2^2 = \boldsymbol{v}_I^T \cdot \boldsymbol{v}_I - 2\hat{\boldsymbol{v}}_I \cdot \boldsymbol{v}_I + \hat{\boldsymbol{v}}_I^T \cdot \hat{\boldsymbol{v}}_I. \tag{19}$$

请注意，eq. (19) 的最后一项不是 $\boldsymbol{v}_I$ 或 $\boldsymbol{v}_E$的函数，因此将进一步省略，并得到优化问题的修订形式:

$$\min_{\boldsymbol{v}_I, \boldsymbol{v}_E} \boldsymbol{v}_I^T \cdot \boldsymbol{v}_I - 2\hat{\boldsymbol{v}}_I \cdot \boldsymbol{v}_I, \text{ s.t. } \mathbf{S}_I \cdot \boldsymbol{v}_I + \mathbf{S}_E \cdot \boldsymbol{v}_E = 0. \tag{20}$$

为求解该优化问题，我们定义拉格朗日函数

$$\mathcal{L} = \boldsymbol{v}_I^T \cdot \boldsymbol{v}_I - 2\hat{\boldsymbol{v}}_I \cdot \boldsymbol{v}_I + \boldsymbol{\gamma}^T \cdot (\mathbf{S}_I \cdot \boldsymbol{v}_I + \mathbf{S}_E \cdot \boldsymbol{v}_E), \tag{21}$$

其中， $\boldsymbol{\gamma} \in \mathbb{R}^M$ 是拉格朗日乘子向量。 $\boldsymbol{v}_I$ 为极值的必要条件是

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_I} = 2\boldsymbol{v}_I - 2\hat{\boldsymbol{v}}_I + \mathbf{S}_I^T \boldsymbol{\gamma} = 0. \tag{22}$$

$\boldsymbol{v}_E$ 成为极值的必要条件是

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_E} = \mathbf{S}_E^T \boldsymbol{\gamma} = 0. \tag{23}$$

$\boldsymbol{\gamma}$ 成为极值的必要条件仅仅是约束

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = \mathbf{S}_I \cdot \boldsymbol{v}_I + \mathbf{S}_E \cdot \boldsymbol{v}_E = 0. \tag{24}$$

注意，式（22）、式（23）和式（24）是关于 $\boldsymbol{v}_E$、 $\boldsymbol{v}_I$ 和 $\boldsymbol{\gamma}$的线性方程，因此得到一个包含 $E + I + M$ =21 + 14 + 24 = 59 个方程和变量的线性系统:

$$\begin{bmatrix} 2\mathbf{I}_I & 0 & \mathbf{S}_I^T \\ 0 & 0 & \mathbf{S}_E^T \\ \mathbf{S}_I & \mathbf{S}_E & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_I \\ \boldsymbol{v}_E \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} 2\hat{\boldsymbol{v}}_I \\ 0 \\ 0 \end{bmatrix}, \tag{25}$$

其中$\mathbf{I}_I \in \mathbb{R}^{I \times I}$ 是单位矩阵。由于我们计算得出其行列式

$$\mathbf{A} = \begin{bmatrix} 2\mathbf{I}_I & 0 & \mathbf{S}_I^T \\ 0 & 0 & \mathbf{S}_E^T \\ \mathbf{S}_I & \mathbf{S}_E & 0 \end{bmatrix} \tag{26}$$

不为零（也就是说，$\mathbf{A}$可逆），因此我们知道式（25）有唯一解:

$$\begin{bmatrix} \boldsymbol{v}_I \\ \boldsymbol{v}_E \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} 2\mathbf{I}_I & 0 & \mathbf{S}_I^T \\ 0 & 0 & \mathbf{S}_E^T \\ \mathbf{S}_I & \mathbf{S}_E & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2\hat{\boldsymbol{v}}_I \\ 0 \\ 0 \end{bmatrix}. \tag{27}$$

如果将$\mathbf{I}_E \in \mathbb{R}^{E \times E}$ 记作单位矩阵，那么

$$\begin{bmatrix} \boldsymbol{v}_I \\ \boldsymbol{v}_E \end{bmatrix} = \begin{bmatrix} \mathbf{I}_I & 0 & 0 \\ 0 & \mathbf{I}_E & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_I \\ \boldsymbol{v}_E \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_I & 0 & 0 \\ 0 & \mathbf{I}_E & 0 \end{bmatrix} \begin{bmatrix} 2\mathbf{I}_I & 0 & \mathbf{S}_I^T \\ 0 & 0 & \mathbf{S}_E^T \\ \mathbf{S}_I & \mathbf{S}_E & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2\hat{\boldsymbol{v}}_I \\ 0 \\ 0 \end{bmatrix} \tag{28}$$

是基于化学计量的方法（无界限）的闭式表达式 .

H 附加图

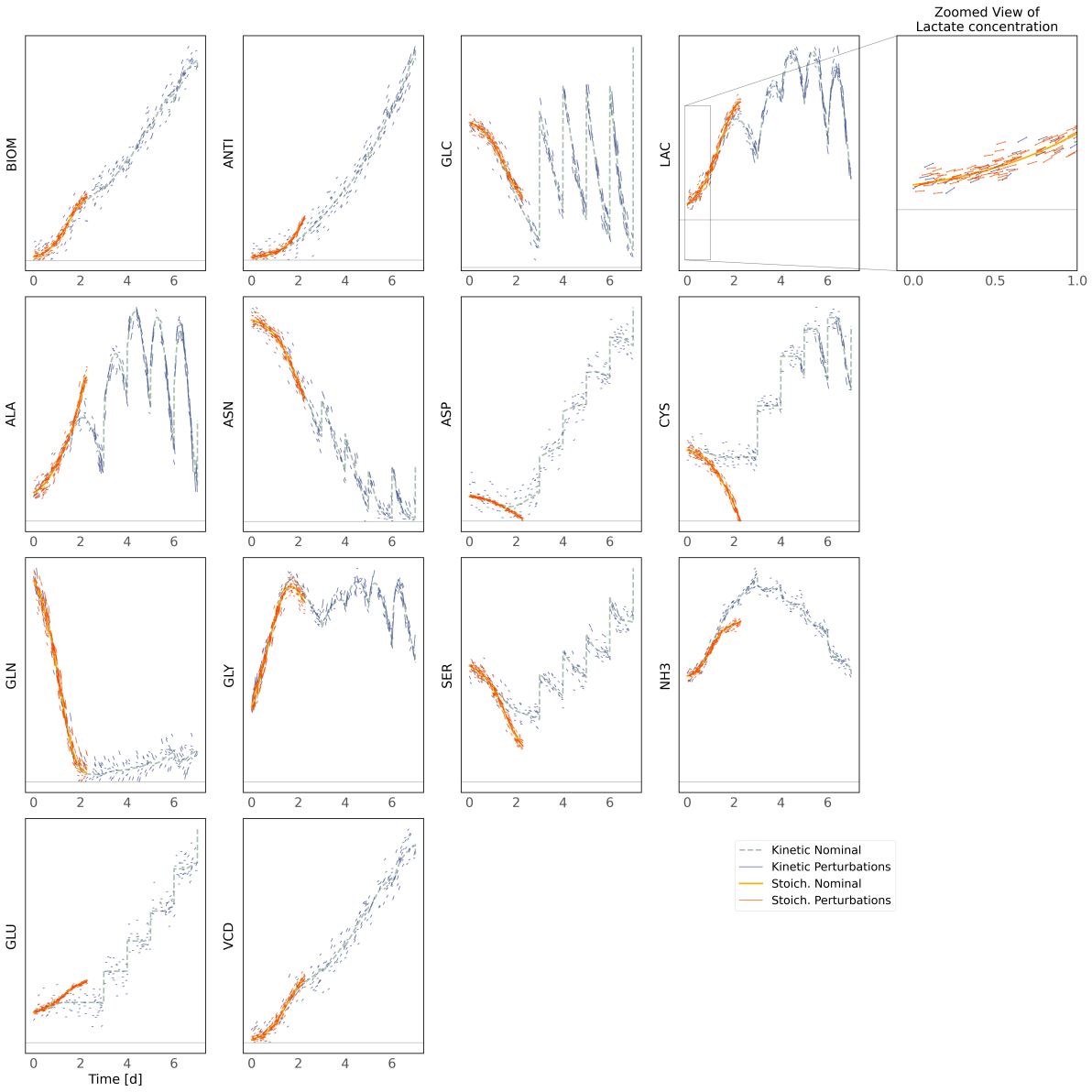


图15：标称轨迹叠加在从扰动初始条件采样得到的短时流上。颜色用于区分不同曲线，如图例所示。

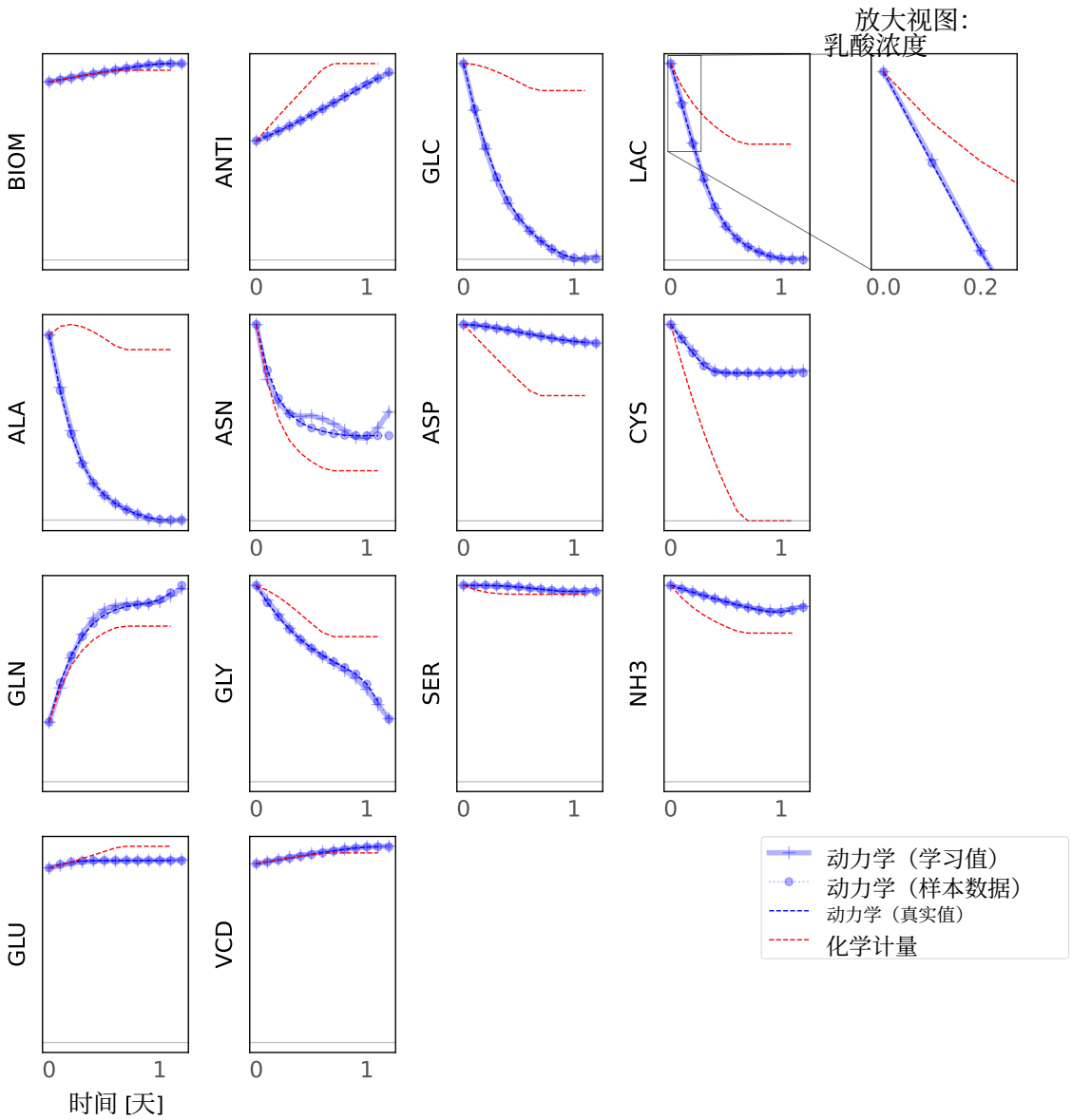


图16：黑箱预测所有变量（动力学（类型1））。另见图。4。颜色用于区分图例中描述的不同曲线。

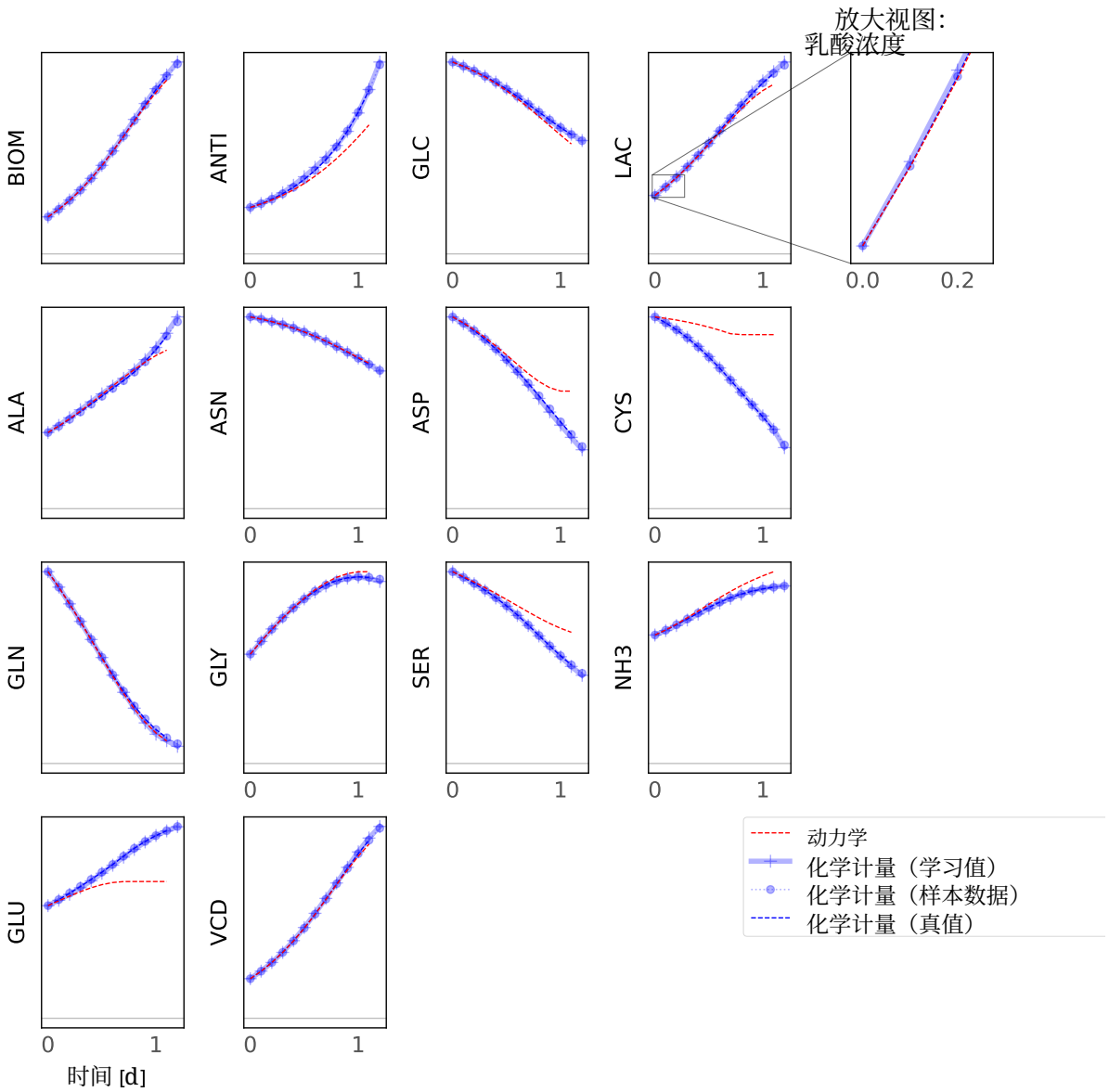
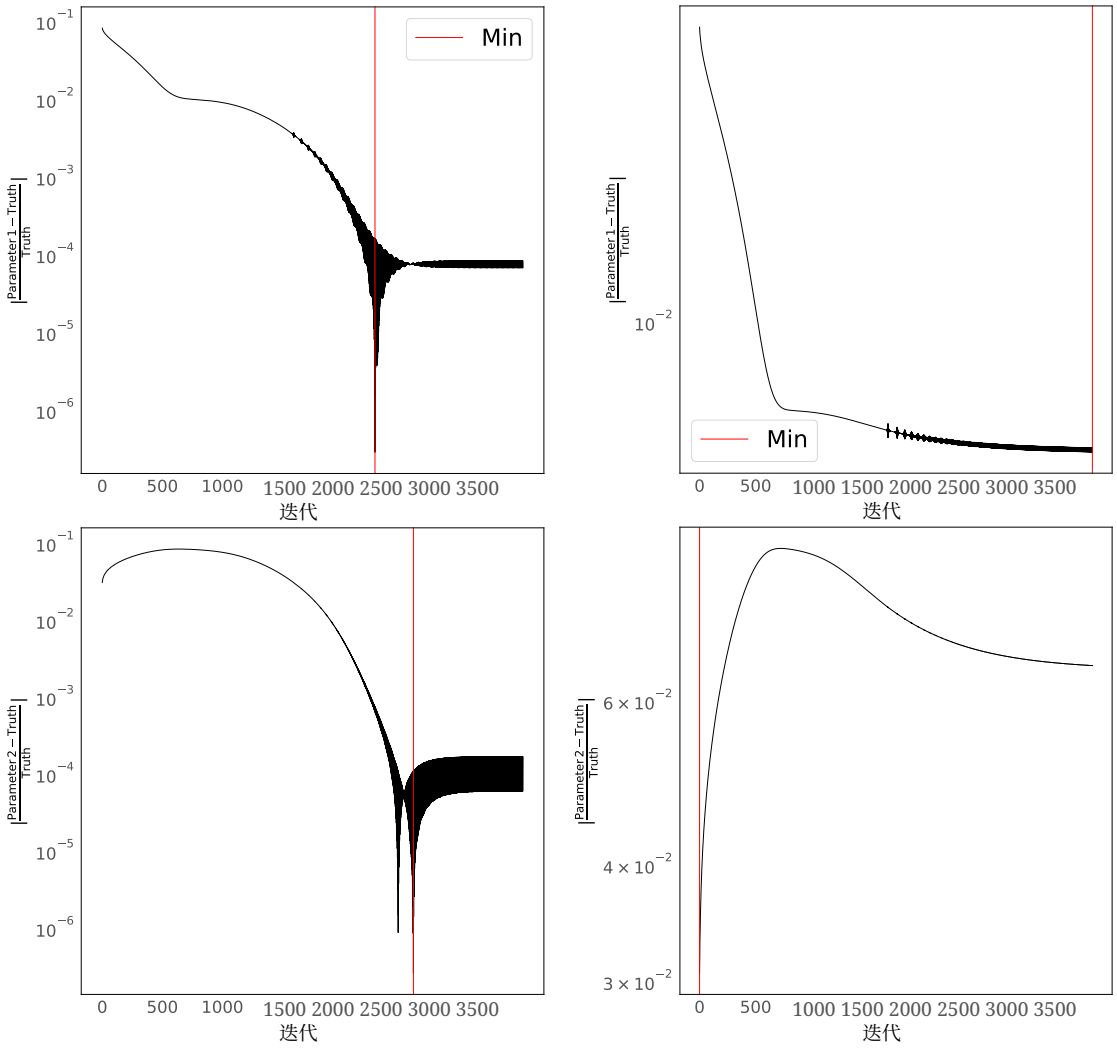
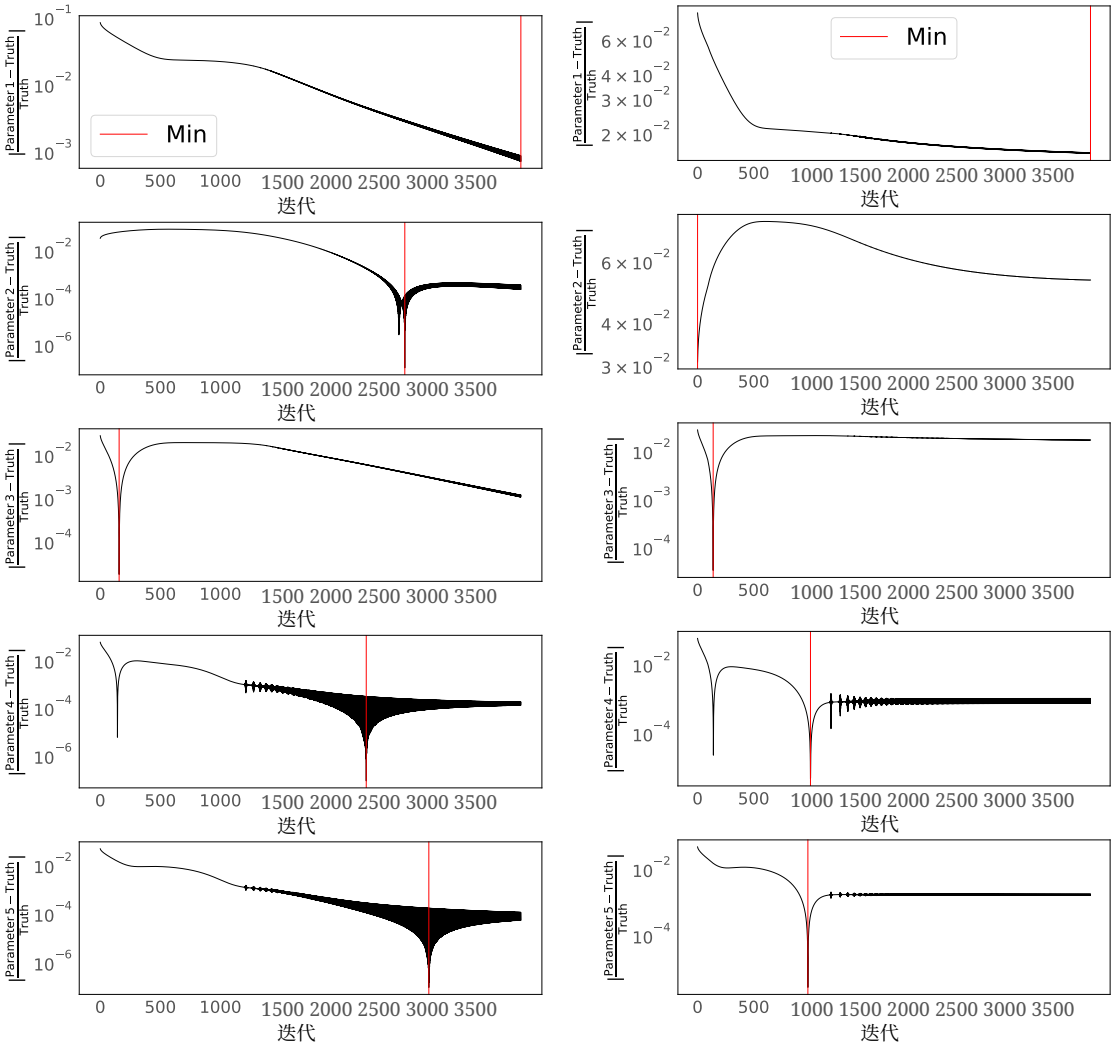


图17：黑箱预测；所有变量（化学计量（类型 2））。不同颜色用于区分图例中描述的不同曲线。



(a) 动力学　　(b) 化学计量

图18：绝对参数误差历史。（双参数白盒情况。）



(a) 动力学　　(b) 化学计量

图19：五参数白盒情况下的绝对参数误差历史。