arXiv:2301.11356v2 [cs.SC] 2 Nov 2023

动力学速率模型的自动发现——方法学框架

Miguel Ángel de Carvalho Servia
化学工程系 伦敦帝国理工学院 南肯辛顿，伦敦，SW7 2AZ，英国
m.de-carvalho-servia21@imperial.ac.uk

Ilya Orson Sandoval
化学工程系 伦敦帝国理工学院 南肯辛顿，伦敦，SW7 2AZ，英国
o.sandoval-cardenas20@imperial.ac.uk

King Kuok (Mimi) Hii
伦敦帝国理工学院化学系 白城，伦敦，W12 0BZ，英国
mimi.hii@imperial.ac.uk

Klaus Hellgardt
伦敦帝国理工学院化学工程系 南肯辛顿，伦敦，SW7 2AZ，英国
k.hellgardt@imperial.ac.uk

Dongda Zhang *
曼彻斯特大学化学工程系 曼彻斯特，M13 9PL，英国
dongda.zhang@manchester.ac.uk

Ehecatl Antonio del Rio Chanona *
伦敦帝国理工学院化学工程系 南肯辛顿，伦敦，SW7 2AZ，英国
a.del-rio-chanona@imperial.ac.uk

2023年11月3日

摘要

催化过程的工业化需要可靠的动力学模型来进行设计、优化和控制。机理模型需要大量领域知识，而数据驱动模型和混合模型缺乏可解释性。自动化知识发现方法，如 ALAMO（用于优化的代数模型自动学习）、SINDy（非线性动力学的稀疏识别）和遗传编程，日益受到关注，但也存在一些局限，例如需要预先假设模型结构、可扩展性较差以及对噪声敏感。为克服这些挑战，我们提出了两种方法论框架：ADoK-S 和 ADoK-W（采用符号回归的强/弱形式自动发现动力学速率模型），通过稳健的模型选择准则，自动生成催化动力学模型。我们采用遗传编程生成模型，并通过序贯优化流程对模型进行改进。我们在复杂程度逐步提高的三个案例研究中测试了这些框架，结果表明，即使催化系统的数据有限且含有噪声，它们也能找出潜在的动力学速率模型，展现出其在化学反应工程领域的应用潜力。

关键词： 化学反应工程、动力学模型生成、自动化知识发现、信息准则、符号回归

1 引言

数学模型是对复杂现象的逻辑表征，广泛应用于物理学[1, 2], 医学 [3, 4], 和化学反应工程 [5, 6]等众多领域。数学模型使研究人员能够将复杂现象提炼为定量表达式，这对于研究化学系统的动力学至关重要，进而也是工业过程开发的关键。

模型在科学与工程领域发挥着至关重要的作用，但如何构建模型仍是一个根本性问题。模型构建有三种经典范式：机理建模、数据驱动建模和混合建模。机理模型由基本定律（例如守恒方程）推导而来，但也可能包含经验表达式 [7, 8], ，并具有可解释性强、外推能力好以及具备物理意义等优点。然而，构建机理模型耗时较长，且需要领域专业知识。此外，这类模型的非线性可能会增加参数估计所需的实验工作量。尽管存在这些挑战，机理模型仍广泛应用于工业领域，并在研究中不断发展 [9–11]。

数据驱动模型仅利用数据即可快速构建，不同于需要了解系统相关知识的机理模型。数据驱动模型的结构灵活，能够迅速适应不同的变量或过程。与机理模型相比，它们的计算速度更快 [12], ，因此适用于实时仿真 [13–16], 、优化 [17–20], 和软测量开发 [21–23]。然而，由于不涉及物理知识，其外推能力通常有限，且模型性能取决于可用数据的数量和质量，因此在某些场景下使用可能并不安全。

混合模型旨在结合机理建模和数据驱动建模各自的优势。这类模型以机理模型为骨架，并包含一个用于改善拟合效果的数据驱动组件。混合建模主要有两种方法：并行方法和串行方法。并行方法使用数据驱动模块来描述模型与数据失配，而串行方法则用它来描述机理模型骨架的参数。无论采用哪种方法，混合模型都保留了机理模型的外推能力，同时具备数据驱动模型的灵活性和易于构建的特点 [24–26]。

混合建模为机理建模和数据驱动建模的局限性提供了一种优雅的解决方案，但并非唯一方案。另一种有效的方法是利用先进的统计和机器学习方法，根据现有数据自动生成并选择机理模型。这一策略也称为自动化知识发现或符号回归，在保留机理模型优点的同时，消除了其中一些缺点，例如需要背景知识以及构建耗时 [27]。本文提出的方法遵循这一范式。

为解决符号回归问题，人们提出了多种方法，包括 ALAMO [28], SINDy [29],Taylor等人 [30], Neumann等人 [31], Forster等人 [32], 以及遗传编程 [33]。然而，这些自动化知识发现框架面临着若干挑战，限制了它们在现实中的适用性。首先，它们通常需要对底层数据生成模型作出结构假设，尤其是需要设计矩阵（即模型库）的非进化策略。其次，它们在处理可用状态变量数量增加时，往往可扩展性较差，非进化策略尤为如此。第三，它们缺乏有充分依据且严谨的模型选择流程，而且所采用的模型选择流程可能不透明，也未经检验。最后，在发现非线性动力学时，如果无法直接获取速率测量值，它们可能会对噪声数据十分敏感。

本节介绍了数学建模在化学工程中的重要性、经典建模范式面临的挑战，以及现代自动化知识发现方法的不足。本研究旨在构建并评估两种具有普适性和稳健性的方法论框架，将严谨的模型选择流程融入动力学速率模型的自动发现。所提出的方法为催化系统的动力学模型发现带来了两项值得关注的创新。首先，我们提出了一种独特方法，将遗传编程与参数估计和信息准则相结合，以发现最优状态空间模型，从而在强形式下实现准确的速率近似。这种方法不同于传统方法：传统方法通常任意选取一个多项式进行插值和速率测量估计 [32]。此外，本文率先将符号回归的弱形式应用于遗传编程；据我们所知，这种方法此前尚未在该领域得到应用或实现。这些创新彰显了我们为该领域带来的重大进展。

本文其余部分安排如下：第 2 节阐述提出这些方法的动机，并对其进行详细介绍；第 3 节介绍三个案例研究，用于分析所提出方法论框架的性能；第 4 节呈现并充分讨论研究结果，同时探讨所提方法的不足；第 5 节总结主要发现，并简要展望未来研究。

2 方法论框架

我们首先简要介绍所采用的方法 ADoK-S 和 ADoK-W（基于符号回归的强/弱形式自动发现动力学模型）。这两个框架均由三个主要步骤组成：（I）利用遗传编程（GP）算法生成候选模型；（II）通过序贯优化算法估计有希望的模型的参数；（III）采用赤池信息准则（AIC），以合理且透明的模型选择流程进行模型选择。[1]

ADoK-S 采用符号回归的传统实现方式，即强形式。这种方法需要速率测量值来推导速率模型。然而，这些测量值无法通过实验直接获得，因此需要进行估算。按照所述的三步流程，ADoK-S 找出最优浓度曲线，以描述观测物种浓度随时间的演化。随后对这些曲线进行数值微分，以估算反应系统的速率测量值。完成速率近似后，再执行相同的三步流程，以发现最适合这些速率的动力学速率模型。然后对所得速率模型进行积分，并与原始浓度数据进行比较。

相比之下，ADoK-W 采用符号回归的弱形式。这种模型提案策略绕过速率估算，直接根据测得的浓度数据构建速率模型。它通过实施三步流程来实现这一点，但遗传编程算法中加入了积分步骤。因此，可以对最优速率模型进行积分，并像 ADoK-S 一样与原始浓度数据进行比较。

如果模型输出不令人满意，无论是由于违反先验知识还是模型拟合不充分，这两种方法都能提供一种闭环方法。建模者可以选择执行一个针对发现任务定制的最优实验——由基于模型的实验设计（MBDoE）确定——随后将其与初始数据集拼接起来。利用新的实验数据，可以对这些方法进行迭代，并检查后续的模型输出。迭代次数可以根据建模者的需要确定，也可以持续到实验预算用尽为止。图 1 直观展示了 ADoK-S 和 ADoK-W 的工作流程，突出了其中最重要的步骤以及两种方法之间的差异。图 2 和图 3 提供了更详细的版本。

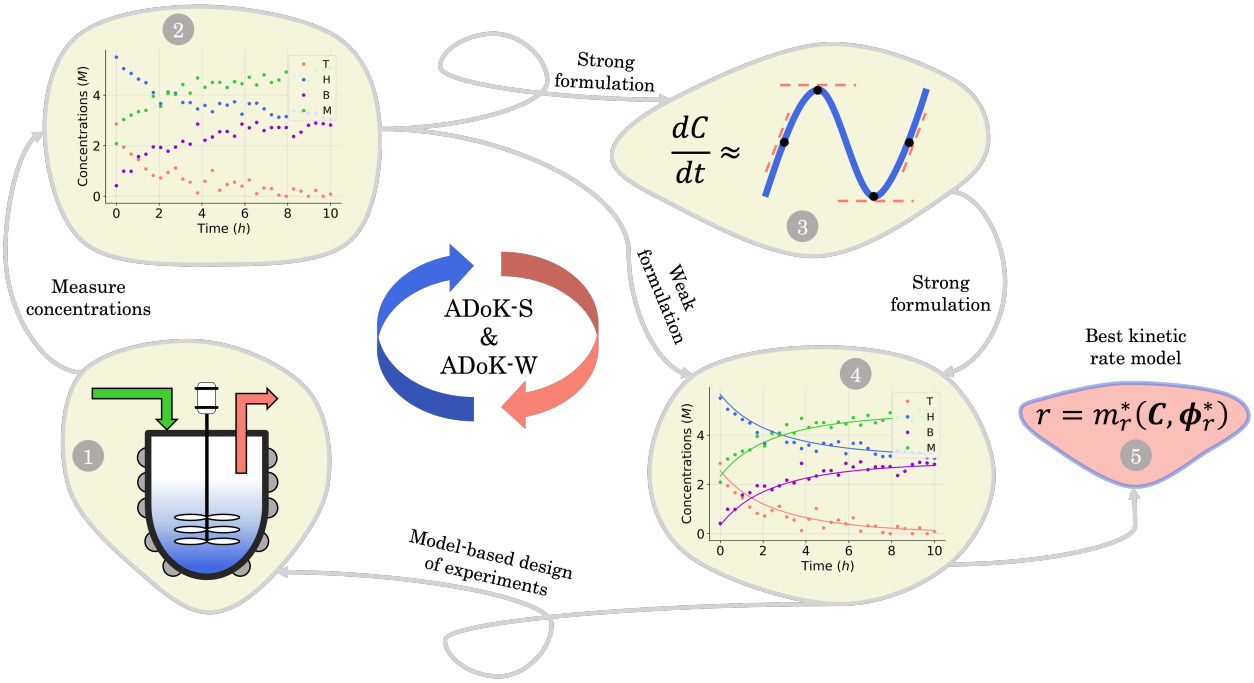


图 1：ADoK-S 和 ADoK-W 方法的流程图表示。ADoK-S 需要对浓度曲线进行数值微分，以估算速率测量值；而 ADoK-W 则利用内嵌的积分步骤，能够直接从浓度数据中提取速率模型。当模型输出不理想时，两种方法都支持通过最优实验（由基于模型的实验设计确定）进行迭代优化，直至达到所需精度或满足实验预算限制。

[1]AIC 的选择 以及其他准则的说明见“附 录”。

在此，我们还设定必要的数学符号，以便准确描述我们的方法。我们从标准的符号回归表述 [34], 开始，之后再介绍我们框架的弱形式和强形式。
定义域集合 $\mathcal{Z}$ 是任意数量的常数 Γ 与固定数量的变量 $\mathcal{X}$ 的并集。算子集 $\mathcal{P}$ 是算术运算（$\diamond$： $\mathbb{R}^2 \to \mathbb{R}$）与有限个特殊一元函数集合（$\Lambda$： $\mathbb{R} \to \mathbb{R}$）的并集。模型搜索空间 $\mathcal{M}$ 是通过在定义域集合 $\mathcal{Z}$ 上对算子集 $\mathcal{P}$ 进行迭代函数复合所能得到的所有可能表达式构成的空间。

变量可以表示为状态向量 $x \in \mathbb{R}^{n_x}$。数据点是由一对特定状态 x 及未知函数 f 的相应目标值 $y \in \mathbb{R}$ 构成的二元组： $\mathbb{R}^{n_x} \to \mathbb{R}$： $y = f(x)$。数据集 $\mathcal{D}$ 由 n_t 个数据点组成：$\mathcal{D} = \left\{ (x^{(i)},\ y^{(i)}) \mid i = 1, \ldots\ldots, n_t \right\}$。为了量化预测值与目标值之间的偏差，我们可以采用任意适当的正测度函数 ℓ： $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$。

符号模型 $m \in \mathcal{M}$具有有限个参数 θ_m，其维度 d_m 取决于模型。我们将模型在特定参数值下的预测以函数形式表示为 $m(\cdot \mid \theta_m)$。我们用 $\hat{y}_m$ 表示来自候选模型 m 的值的预测（即 $\hat{y}_m = m(\cdot \mid \theta_m)$）。对于我们的研究而言，将模型生成步骤与每个模型的参数优化分离开来非常重要。最优模型 m^* 定义为

$$m^* = \arg\min_{m \in \mathcal{M}} \sum_{i=1}^{n_t} \ell\left(\hat{y}_m^{(i)}, y^{(i)}\right), \tag{1}$$

且其最优参数满足

$$\theta_{m^*}^* = \arg\min_{\theta_{m^*}^*} \sum_{i=1}^{n_t} \ell\left(\hat{y}_{m^*}^{(i)}, y^{(i)}\right). \tag{2}$$

在动力系统中，状态变量是时间的函数， $x(t) \in \mathbb{R}^{n_x}$，表示真实动力系统在固定时间区间 $\Delta t = [t_0, t_f]$内的演化。动力学由变化率$\dot{x}(t) \in \mathbb{R}^{n_x}$ 和初始条件 $x_0 = x(t = t_0)$定义。

对于我们的动力学速率模型，我们假设 n_t 采样时间设定在固定时间区间 $t^{(i)} \in \Delta t$内。每个时间点 $t^{(i)}$ 的浓度测量值 C 是系统真实演化 $C^{(i)} \approx x(t^{(i)})$的样本，而速率估计值 r 则是变化率 $r^{(i)} \approx \dot{x}(t^{(i)})$的近似值。

这里，可用数据集 $\mathcal{D}$ 由时间和状态测量值的有序对 $\mathcal{D} = \{(t^{(i)}, C^{(i)}) \mid i = 1, \ldots, n_t\}$构成。与之前一样，我们用帽号表示由所提出模型 m给出的状态 C_m 或速率$\hat{r}_m$ 的预测。具有特定参数 θ_m 的模型输出分别表示为$\hat{C}_m(\cdot \mid \theta_m)$和$\hat{r}_m(\cdot \mid \theta_m)$。
模型的复杂度表示为 $\mathcal{C}(m)$。[2] 我们区分具有不同复杂度等级 $\kappa \in \mathbb{N}$ 的表达式族 $\mathcal{M}^\kappa = \{m \in \mathcal{M} \mid \mathcal{C}(m) = \kappa\}$。

2.1 ADoK-S 简介

对于 ADoK-S，目标是找到能够最佳地将状态映射到速率的模型 m：

$$\hat{r}_m(t \mid \theta_m) = m(x(t) \mid \theta_m). \tag{3}$$

为直接实现这一点，必须根据可用的浓度测量值 $C^{(i)}$推导出变化率 $r^{(i)}$ 的估计值。为此，我们的方法构建一个中间符号模型 η，使得 $\eta(t^{(i)}) \approx C^{(i)}$，并遵循标准符号回归流程（见（1）和（2））；模型选择过程则如第 2.3 节所述。
由于该模型可微，其导数可用于近似所需的速率： $\dot{\eta}\left(t^{(i)}\right) \approx r^{(i)}$。有了这些估算值后，优化问题可写为如下形式。外层针对固定的复杂度 κ对模型提案进行优化，

$$m^\star = \arg\min_{m \in \mathcal{M}^\kappa} \sum_{i=1}^{n_t} \ell\left(\hat{r}_m(t^{(i)} \mid \theta_m), r^{(i)}\right), \tag{4}$$

[2]在这里，我们使用.....的数量 表达式树中的节点作为一个符号表达式的复杂度 bolic 表达式 [35]。

而内层则针对最佳模型的参数进行优化，

$$\theta_{m^\star}^\star = \arg\min_{\theta_{m^\star}} \sum_{i=1}^{n_t} \ell\left(\hat{r}_{m^\star}(t^{(i)} \mid \theta_{m^\star}), r^{(i)}\right). \tag{5}$$

在公式 (4) 和 (5) 中，ℓ 表示残差平方和（RSS）。图 2 展示了该方法的完整流程。

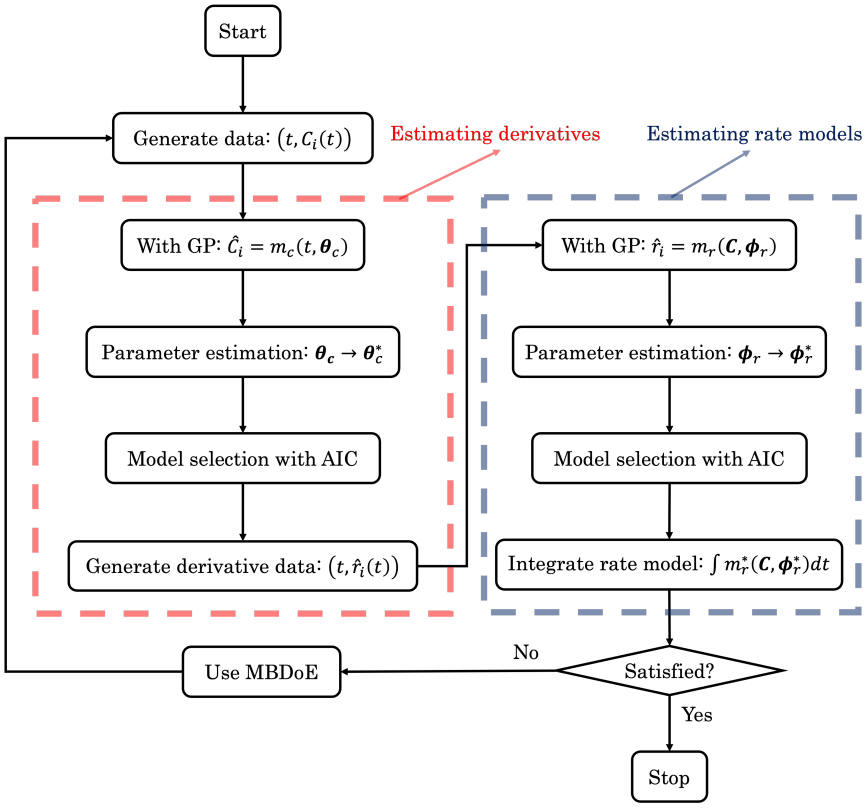


图 2：ADoK-S（**基于 强**形式的 符号**回**归进行动力学**自动**发现）的流程图；红色和蓝色虚线框分别表示估算速率测量值和动力学速率模型的步骤。

2.2 ADoK-W 简介

对于 ADoK-W，我们旨在找到模型 m，使其能够最佳地将状态变量映射到定义状态动力学的微分方程组，进而预测浓度演化：

$$\dot{x}_m(t \mid \theta_m) = m(x(t) \mid \theta_m), \tag{6a}$$

$$\hat{C}_m(t \mid \theta_m) = C_0 + \int_{t_0}^{t} \dot{x}_m(\tau \mid \theta_m) d\tau, \tag{6b}$$

其中，初始条件 C_0 是第一个浓度测量值。对于这一表述，外层同样针对特定复杂度级别 κ对模型提案进行优化，

$$m^\star = \arg\min_{m \in \mathcal{M}^\kappa} \sum_{i=1}^{n_t} \ell\left(\hat{C}_m(t^{(i)} \mid \theta_m), C^{(i)}\right), \tag{7}$$

而内层则对最佳模型的参数进行优化，

$$\theta_{m^\star}^\star = \arg\min_{\theta_{m^\star}} \sum_{i=1}^{n_t} \ell\left(\hat{C}_{m^\star}(t^{(i)} \mid \theta_{m^\star}), C^{(i)}\right). \tag{8}$$

在式（7）和（8）中，ℓ 表示残差平方和。该变体的整个过程如图3所示。

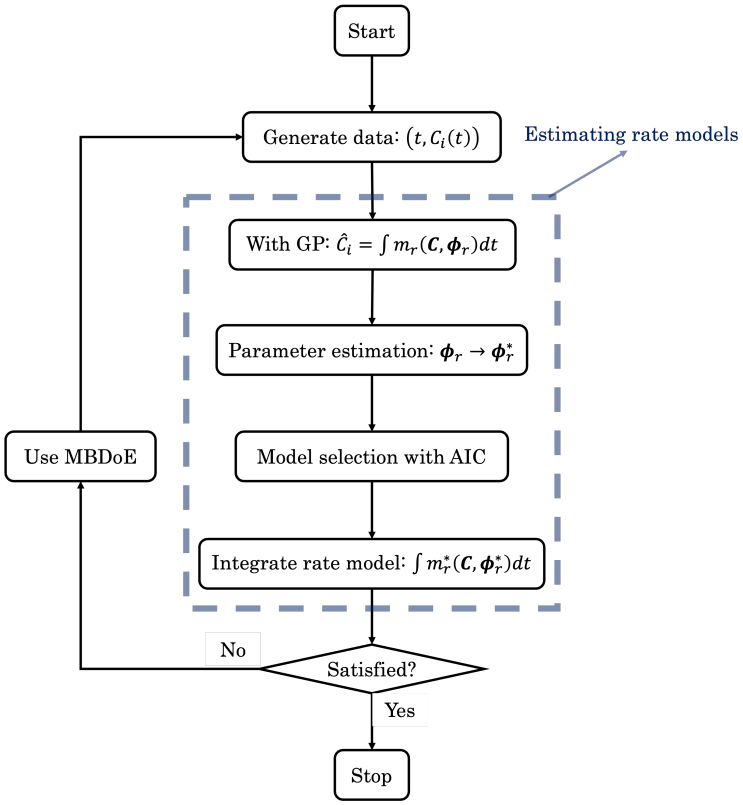


图3：ADoK-W（**A**utomated **D**iscovery **o**f **K**inetics using a **W**eak formulation of symbolic regression）的流程图；蓝色虚线框表示估计速率模型的步骤。

2.3 模型选择

给定一个维度为 d_m、参数为 θ_m 的模型 m，赤池信息准则定义为

$$\mathrm{AIC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D}) + 2d_m, \tag{9}$$

其中，$\mathcal{L}$ 具体表示负对数似然（NLL）。给定两个相互竞争的模型 m_1 和 m_2，根据式（9）计算出的赤池信息准则值最低的模型为首选模型。在ADoK-S和ADoK-W框架中选择赤池信息准则作为模型选择准则的原因详见“附录”。

2.4 基于模型的实验设计

用于回归的数据集可能不足以提出合适的模型。在这种情况下，若假设实验预算尚未用尽，则可以利用优化模型中隐含的信息，提取出具有信息价值的新实验方案。为此，我们可以搜索一个初始条件，使得使用现有数据集得到的最佳两个模型提案 η 和 μ 的状态预测值 $\hat{x}(t)$ 之间的偏差最大。此基于模型的实验设计方法发表于 [36]:

$$x_0^{(new)} = \arg\max_{x_0} \int_{t_0}^{t_f} \ell\left(\hat{x}_\eta\left(\tau \mid \theta_\eta^\star\right), \hat{x}_\mu\left(\tau \mid \theta_\mu^\star\right)\right) \, d\tau. \tag{10}$$

在公式 (10) 中， ℓ 表示残差平方和。从提出的初始条件出发，可以开展实验，获取一批新的数据点并将其加入原始数据集。最后，可以利用扩充后的数据集重新执行整个模型提案与选择流程，从而形成信息丰富的实验与最优模型之间的闭环。

3 催化动力学案例研究

为评估我们方法的有效性，我们分析了文献中的三个催化反应案例研究：一个理论异构化反应 [37],一氧化二氮的分解 [38],以及甲苯加氢脱烷基化 [39]。为简洁起见，我们的讨论主要聚焦于最复杂的实例——甲苯加氢脱烷基化。关于该案例及其他案例的全面讨论，详见“附录”。

3.1 甲苯加氢脱烷基化

甲苯加氢脱烷基化反应可用式（11）表示，而式（12）则描述了反应速率 [39], 其中动力学参数定义如下： $K_A = 2$ $\mathrm{M}^{-1}\,\mathrm{h}^{-1}$, $K_B = 9\,\mathrm{M}^{-1}$ 和 $K_C = 5\,\mathrm{M}^{-1}$。

$$C_6H_5CH_3 + H_2 \rightleftharpoons C_6H_6 + CH_4 \tag{11}$$

$$r = -\frac{dC_T}{dt} = -\frac{dC_H}{dt} = \frac{dC_B}{dt} = \frac{dC_M}{dt} = \frac{K_A C_T C_H}{1 + K_B C_B + K_C C_T} \tag{12}$$

在公式 (12) 中， *T*、*H*、*B* 和 *M* 分别对应甲苯、氢气、苯和甲烷。从公式 12 出发，建立了一个计算机模拟数据集，其中 $\Delta t = [0, 10]$ h 和 $n_t = 30$。该数据集由五个不同的实验组成，每个实验均在不同的初始条件下进行；这些实验从一个 2^k析因设计 [40] 中随机选取。

对于所有实验，假设系统同时处于等容和等温状态，并加入高斯噪声以模拟实验数据。第二个和第四个实验生成的数据集分别如图 4 a) 和 e) 所示。该数据集包含 150 个数据点，对于动力学研究而言规模合理, [30, 41, 42], 尤其是考虑到近年来高通量装置的发展。

4 结果与讨论

4.1 ADoK-S 性能

如图 2 所示，在 ADoK-S 中从动态浓度轨迹推导动力学模型的第一阶段是提出浓度剖面模型。这需要应用遗传编程算法（在 ADoK-S 中，我们使用 Cranmer [35] 提供的实现），其表达式构建规则如下: $\mathcal{P} = \{+, -, \div, \times, \exp\}$ and $\mathcal{X} = \{t\}$，其中 t 表示时间变量。基于我们对动力学建模的物理理解，这种选择被认为是合理的——它清晰地展示了如何将专家知识引入符号搜索。
值得注意的是，有时描述反应系统动力学的基本常微分方程（ODE）系统可能不存在闭式表达式解。在这种情况下，无论采用何种构造规则，提出的浓度模型都存在缺陷。尽管如此，在大多数经过测试的案例中（见“附录”），所选构造规则均显示出能够成功近似浓度轨迹的变化，更重要的是，能够近似速率测量值，而不受是否存在闭式表达式的影响。
图 4 b) 和 c) 展示了 ADoK-S 近似反应系统浓度分布及速率测量值的能力。然而，图 4 f) 和 g) 则显示了相反的情况：ADoK-S 显然难以捕捉浓度动态演化的适当行为，导致后续的速率

近似效果不佳。这些结果进一步推动了 ADoK-W 的开发，并促使人们尽管其计算时间较长，仍在复杂案例研究中使用它。

ˆ ˆ 在这个特定案例研究中，我们为每项实验构建四个浓度模型，具体而言，($C_{T,\ i}$, $C_{H,\ i}$, ˆ$C_{B,\ i}$, ˆ$C_{M,\ i}$)，对应于 $i \in [1, 2, ..., 5]$，其中 i 表示实验编号。需要强调的是，这些模型均独立开发。有人可能会认为，这种方法可能导致模型违反质量守恒等基本物理原理。然而，我们认为，在这一初始阶段，首要目标是准确拟合系统的速率测量值，从而有助于构建精确的动力学模型。因此，在此情境下，一定程度的物理不一致或许可以接受。

本节重点介绍第二项实验的结果，因为所有其他实验均采用了相同的方法。遗传编程算法针对每个由用户设定上限的复杂度级别，为*T*、*H*、*B* 和*M* 的浓度提出模型结构。下文展示第二项实验中*T* 的浓度曲线。此处，p_i 表示可从特定模型的时变浓度数据集中估算的 i^{th} 参数。此外，ˆ C_i 表示 ADoK-S 在第二项实验中为物种 *T* 提出的 i^{th} 浓度模型。

$$\hat{C}_1(t) = p_1 \tag{13a}$$

$$\hat{C}_2(t) = \frac{p_1}{\exp(t)} \tag{13b}$$

$$\hat{C}_3(t) = \frac{p_1}{p_2 + t} \tag{13c}$$

$$\hat{C}_4(t) = \frac{p_1}{\exp(p_2) + t} \tag{13d}$$

$$\hat{C}_5(t) = \frac{p_1}{t + p_2} - p_3 \tag{13e}$$

$$\hat{C}_6(t) = \frac{p_1}{\exp(p_2 t) + t} \tag{13f}$$

$$\hat{C}_7(t) = \frac{p_1}{p_2 + \frac{t}{p_3}} - p_4 \tag{13g}$$

生成浓度模型结构后，下一步是进行参数估计，旨在找到使浓度模型的响应与测得浓度之间的误差最小化的最优参数值。

确定参数值后，针对每个模型计算负对数似然（NLL）函数和赤池信息准则（AIC）（式（9））， 以便进行模型选择。在提出的模型中， m_7 是所选模型，用于近似第二个实验中物种*T*的消耗量测量值的速率。

图 4 b) 和 f) 展示了 ADoK-S 为第二个和第四个实验中的所有物种构建、优化并选定的浓度模型。选定浓度模型后，通过对这些模型进行数值微分，估算生成速率（针对产物）和消耗速率（针对反应物）。图 4 c) 和 g) 展示了这些估算速率随时间的变化，并与真实系统 $\dot{x}(t)$ 的速率测量值进行比较；这类测量在实践中通常难以获取。正如前文所述，该方法在实验二的速率估算方面表现出色，但在实验四中则表现不佳。较早达到平衡（$\sim$ 2 小时）且加性噪声较高，导致实验四提供的动力学信息较少，使 ADoK-S 等框架难以提取有意义的见解并近似速率测量值。

与图 2 所示的流程图一致，ADoK-S 的下一阶段采用相同的遗传编程算法（用于推导浓度分布）来提出速率模型。该过程以迭代方式展开，不断优化速率模型群体，旨在满足式 (4)。表达式的构建规则为 $\mathcal{P} = \{+, -, \div, \times\}$ and$\mathcal{X} = \{C_T, C_H, C_B, C_M\}$，这些规则是基于我们对动力学模型的既有认识所作的选择，也是将专家知识融入该方法的另一种途径。基于这些构建规则，遗传编程算法提出了 13 种速率模型结构；为简洁起见，我们仅列出其中几种:

$$\hat{r}_1 = K_1 \tag{14a}$$
$$\hat{r}_2 = K_1 C_T \tag{14b}$$
$$\hat{r}_3 = K_1 C_T C_H \tag{14c}$$
$$\hat{r}_4 = K_1 C_H - K_2 C_T \tag{14d}$$
$$\hat{r}_5 = K_1 C_T C_H - K_2 C_T \tag{14e}$$
$$\hat{r}_6 = \frac{K_1 C_T((C_H - K_2)(K_3 - C_B) + K_4)}{C_B - K_5}. \tag{14f}$$

根据浓度数据估计动力学模型的参数 K_i for $i \in [1, 2, ..., 5]$，这一过程称为动态参数估计。通过满足式 (5) 来求解参数估计问题，并利用人工蜂群算法和有限内存BFGS算法进行优化，以确定最优解。计算所有候选模型的负对数似然和赤池信息准则值后，选出的模型为 $\hat{r}_3$，其响应见图 4 d) 和 h)。

式 (14a) 中展示的方程，包括 $\hat{r}_3$，均与式 (12) 中展示的数据生成速率模型不符。此外，如图 4 h) 所示，模型的响应与第四次实验的浓度数据吻合较差。因此，ADoK-S 必须使用基于模型的实验设计再进行一次迭代。在基于模型的实验设计中，使用 ADoK-S 得出的排名前两位的模型，即 $\hat{r}_3$ 和 $\hat{r}_5$，通过求解式 (10) 来提出一项判别性实验。

基于模型的实验设计流程建议在初始条件 ($C_{T,\ 0}$，$C_{H,\ 0}$，$C_{B,\ 0}$，$C_{M,\ 0}$) =(1。948， 7。503， 1。232，2。504) M 下进行第六次实验。新实验经历与最初五次实验相同的一系列操作：生成、优化，并选择最佳浓度模型来近似速率。计算出新实验的速率后，将其与先前的近似值拼接起来，并据此生成、优化和筛选新的速率模型。为简洁起见，此处不展示提出的浓度模型和速率模型，但以下列出了增加额外实验后由 ADoK-S 选出的最佳（$\hat{r}_1$）和次佳（$\hat{r}_2$）动力学模型:

$$\hat{r}_1 = \frac{C_T C_H}{C_T + K_1} \tag{15a}$$
$$\hat{r}_2 = K_1 C_T(-K_2 C_B^2 + K_3 C_B + K_4 C_H - C_T + K_5) + K_6. \tag{15b}$$

尽管新模型的拟合优度较最初选定的速率模型有所提升，但仍不尽如人意。因此，ADoK-S 又进行了一次迭代，基于模型的实验设计程序建议开展第七次实验，初始条件为 ($C_{T,0}, C_{H,0}, C_{B,0}, C_{M,0}$) = (2.560,5.654,0.341,2.337) M。ADoK-S 在第七次实验后选定的动力学模型（记作 r^*）如下所示:

$$r^* = \frac{K_1 C_T C_H}{K_2 + K_3 C_B + K_4 C_T}. \tag{16}$$

如图所示，经过 ADoK-S 的两次迭代，该方法能够发现一个与数据生成模型（式 (12)）结构相同的动力学模型（式 (16)）。

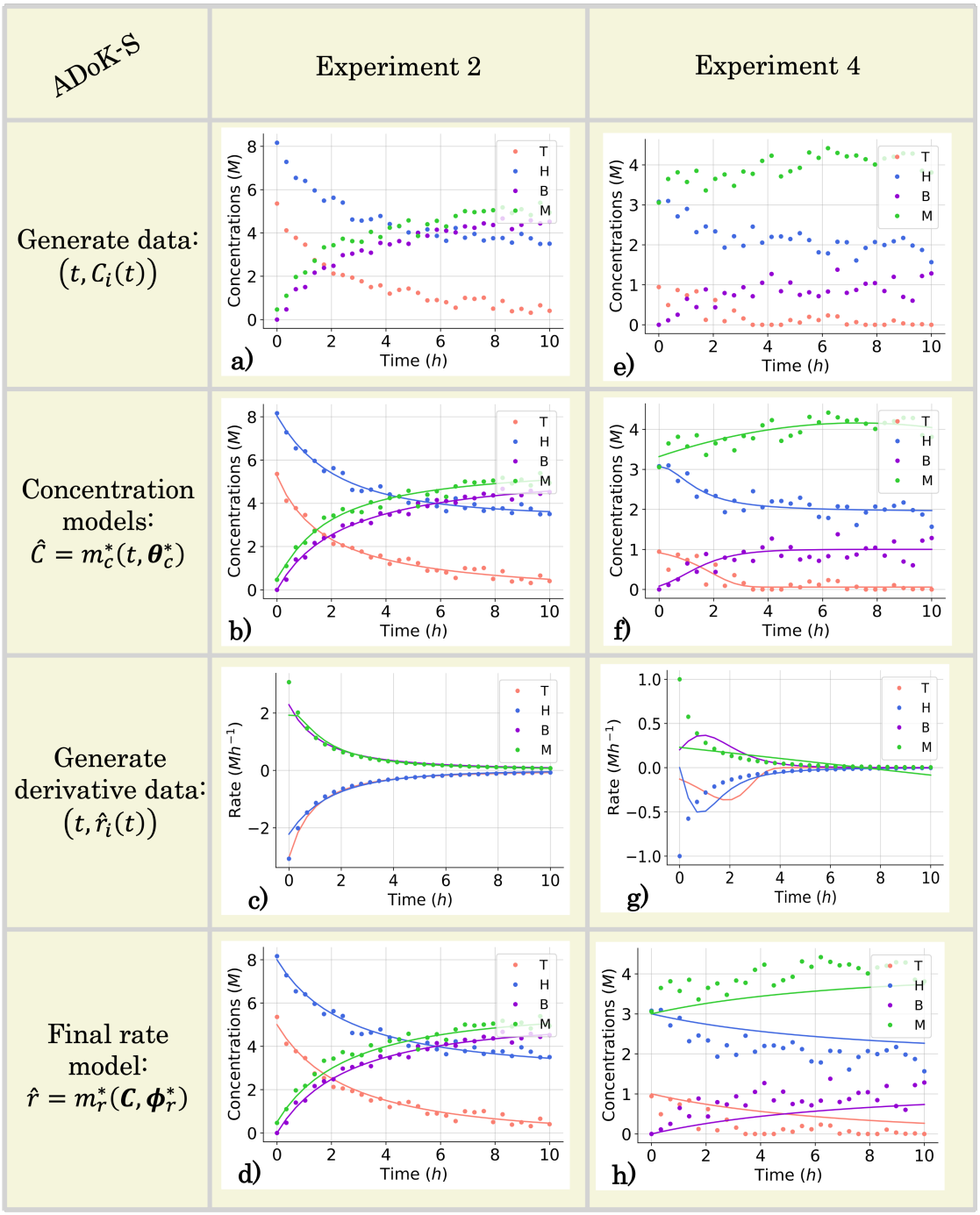


图4：第二次和第四次计算实验的条件分别为（$C_{T,\,0}$，$C_{H,\,0}$，$C_{B,\,0}$，$C_{M,\,0}$）∈（5，8，0，0.5），（1，3，0，3）M，其中T、H、B、M分别表示甲苯、氢气、苯和甲烷。**a)** 和**e)**：用于执行甲苯加氢脱烷基化 ADoK-S 的第二次和第四次实验所测得的浓度数据。**b)** 和**f)**：赤池信息准则（AIC）选出的浓度分布，用于对第二次和第四次实验中可观测物种的浓度动态轨迹随时间的变化进行建模。这些模型用于近似速率测量值。**c)** 和**g)**：所选浓度分布的数值导数以及真实速率测量值（实际情况下无法获取）。**d)** 和**h)**：针对第二次和第四次实验，采用初始实验集执行 ADoK-S 后，第一次迭代所选速率模型的响应。

4.2 ADoK-W 性能

如第 4.1 节所述，正如 [43]所预料的那样，在噪声严重的条件下，ADoK-S 在近似复杂系统的速率测量值方面表现出局限性。这促使人们希望在可能的情况下绕过速率估计，以进行知识发现。

ADoK-S 在理论上的不足，加上它在辨识甲苯加氢脱烷基化背后的真实模型时表现欠佳，促使人们开发了 ADoK-W。ADoK-W 利用符号回归的弱形式，在提出速率模型时减少了对速率近似的需求。这种新设计使 ADoK-W 能够直接从浓度数据提出速率模型，而不再局限于提出直接输入输出映射模型。这项创新在于将积分步骤纳入负责提出模型的遗传编程算法中。

不过，除这一差异外，ADoK-W 的运行方式与 ADoK-S 完全相同。首先，使用遗传编程构建模型。随后，对每个复杂度类别中最优的模型进行优化，方法是求解参数估计问题。在优化后的模型集中，选择赤池信息准则值最低的模型。如果建模者认为算法的输出不令人满意（且实验预算允许），可以开展由基于模型的实验设计指导的额外实验，将测量值与先前的数据集拼接起来，然后再次运行 ADoK-W 算法。

ADoK-S 中为速率模型构建所制定的规则，在执行 ADoK-W 时保持不变。通过在不同复杂度下求解式（7），遗传编程算法构建了六个速率模型，如下所示：

$$\hat{r}_1 = K_1 \tag{17a}$$
$$\hat{r}_2 = K_1 C_T \tag{17b}$$
$$\hat{r}_3 = K_1 C_T C_H \tag{17c}$$
$$\hat{r}_4 = \frac{K_1 C_T C_H}{C_B + K_2} \tag{17d}$$
$$\hat{r}_5 = \frac{K_1 C_T C_H}{C_T + K_2 C_B} \tag{17e}$$
$$\hat{r}_6 = \frac{K_1 C_T C_H}{K_2 + K_3 C_T + K_4 C_B}. \tag{17f}$$

按照图 3 所示的框架，提出最佳速率模型后，通过确定满足式（8）的参数对这些模型进行优化。优化完成后，计算赤池信息准则（AIC）值，并选择值最低的模型。在本案例研究中，选定的模型 r_6 与式（12）中的真实模型完全一致，且整个方法未进行任何基于模型的实验设计迭代。图 5 展示了第二次和第四次实验的实测浓度数据，以及 ADoK-W 所选动力学模型的响应。

5 结论

在本研究中，我们提出了两个数据驱动框架 ADoK-S 和 ADoK-W，用于解决符号回归问题，从含噪浓度测量中发现动力学速率模型。这些方法将遗传编程算法与参数估计和信息准则相结合，无需施加过多限制，即可生成、优化并选择模型。与可能模糊模型可解释性的黑箱模型和混合模型不同，也不同于构建起来可能耗时耗力的传统机理模型，我们的方法为可解释模型的开发提供了透明、高效的流程。

ADoK-S 需要速率测量值来提出速率模型，这符合强形式；而 ADoK-W 无需进行这类测量，能够直接根据浓度数据构建速率模型，这正是弱形式的特点。在甲苯加氢脱烷基化案例研究中，两种方法均成功识别出了该反应的潜在速率模型。

然而，由于速率近似误差和系统复杂性，ADoK-S 需要额外迭代两次才能发现真实动力学模型。另一方面，ADoK-W 仅凭最初五次实验便找到了数据生成模型，展现出在复杂空间中更出色的表现，尽管并非没有“免费午餐”。ADoK-S 可以在几分钟内提出速率模型，而 ADoK-W 则需要数小时才能做到同样的事。

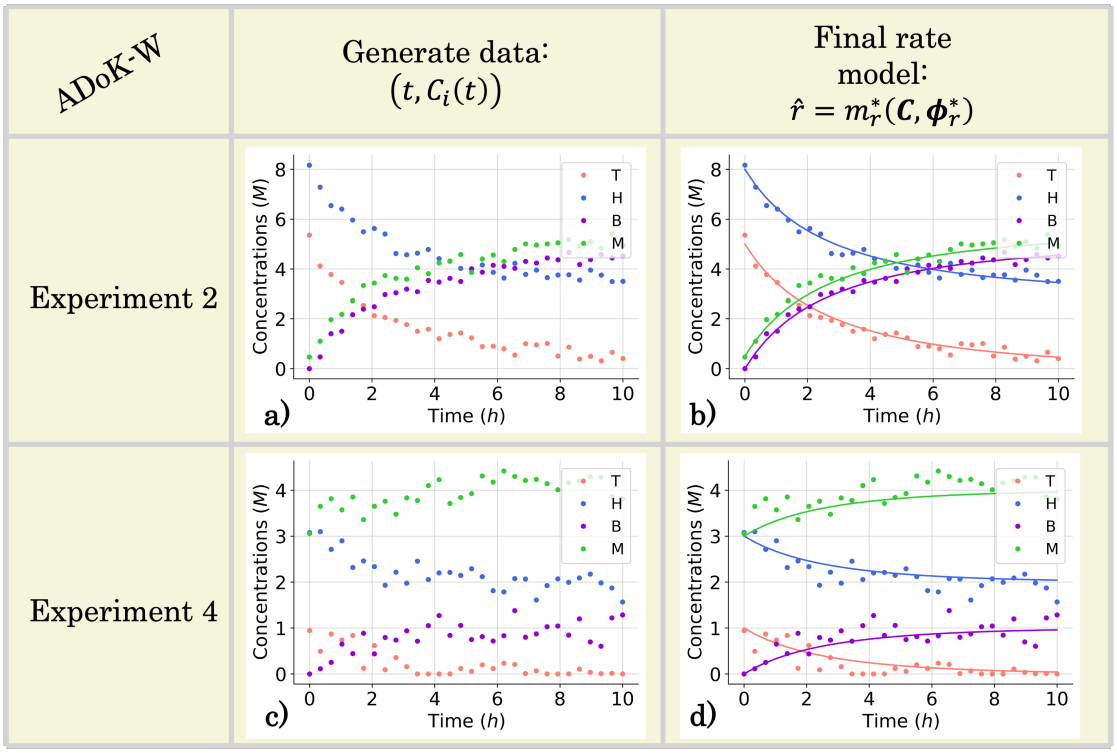


图5：第二次和第四次计算实验的条件分别为 ($C_{T,\,0}$，$C_{H,\,0}$，$C_{B,\,0}$，$C_{M,\,0}$) ∈(5，8，0，0.5)，(1，3，0，3) M，其中 T、H、B、M 分别表示甲苯、氢气、苯和甲烷。**a)** 和 **c)**：用于执行甲苯加氢脱烷基化 ADoK-W 的第二次和第四次实验测得的浓度数据。**b)** 和**d)**：针对第二次和第四次实验，ADoK-W 仅经过首次迭代后所选速率模型的响应。

本案例研究的结果，以及“附录”中展示的结果（这些结果同样表明能够从数据中成功发现动力学速率模型），凸显了在反应工程和催化领域的动力学模型开发中采用自动化知识发现方法的潜力。总结结果见表1。尽管我们展示了在先验知识极少的情况下实现这一潜力的可能性，但较长的计算时间表明，有必要纳入质量守恒定律、平衡行为等物理约束，以缩小搜索空间并提高计算效率。

表1：汇总了 ADoK-S 和 ADoK-S 在所研究的全部三个案例研究中的性能结果。

	假设的异构化反应反应	一氧化二氮分解	氢脱烷基化甲苯 of
迭代次数 – ADoK-S	2	2	3
迭代次数 – ADoK-W	2	3	1
数据生成动力学模型	$\frac{7C_A-3C_B}{4C_A+2C_B+6}$	$\frac{2C_{N_2O}^2}{1+5C_{N_2O}}$	$\frac{2C_TC_H}{1+9C_B+5C_T}$
发现的速率模型 – ADoK-S	$\frac{8.666C_A-3.642C_B}{4.976C_A+2.525C_B+7.003}$	$\frac{1.937C_{N_2O}^2}{1+4.803C_{N_2O}}$	$\frac{1.669C_TC_H}{1+7.347C_B+4.439C_T}$
发现的速率模型 – ADoK-W	$\frac{9.998C_A-4.496C_B+0.386}{6.038C_A+2.137C_B+7.892}$	$\frac{1.584C_{N_2O}^2}{1+3.798C_{N_2O}}$	$\frac{1.124C_TC_H}{1+4.932C_B+2.928C_T}$

需要指出的是，任何数据驱动方法（包括本文介绍的方法）的成功，都在很大程度上取决于所使用的数据。本案例研究中作出的数据假设未必总是成立。对于快速反应或

涉及不同物种的反应，采样速率以及假设所有物种都可测量，可能都不切实际。假设设备经过完美校准且不存在系统误差，虽然较为乐观，但未必能准确反映真实的实验设置。

最后，尽管 ADoK-S 和 ADoK-W 是为发现催化系统的动力学模型而设计的，但无需进行重大修改，便可将其扩展用于探索非反应系统的动力学。这拓宽了它们在其他领域和学科中的潜在应用。

利益冲突声明

作者声明，他们不存在已知的竞争性经济利益或个人关系，可能会对本文所报告的工作产生影响。

致谢与经费支持

本研究得到工程与物理科学研究理事会（EPSRC）资助，项目编号为 EP/S023232/1。

代码和数据可用性

本研究中用于生成所有结果和图表的代码可通过 https://github.com/ MACServia/auto_discov_kin_rate_models 获取。

A 案例研究

A.1 假设的异构化反应

用于评估所提出方法 ADoK-S 和 ADoK-W 性能的最简单案例研究，是一个催化异构化反应：*A* 在催化活性位点的作用下可逆地转化为*B* 。该反应如下所示：

$$A \rightleftharpoons B. \tag{18}$$

描述*A* 和*B* 浓度随时间演化的动力学速率模型如下所示。该表达式直接引自 Marin 等人 [37]。

$$r = -\frac{dC_A}{dt} = \frac{dC_B}{dt} = \frac{K_A C_A - K_B C_B}{K_C C_A + K_D C_B + K_E} \tag{19}$$

在式（19）中， C_A 和 C_B 分别表示反应物*A* 和产物*B*的浓度。该动力学速率模型的动力学参数以 K_i 表示，其中 $i \in [A, B, ..., E]$。为生成用于测试两个框架所需的数据集，进行了五次计算实验，每次实验采用不同的初始条件。计算实验采用以下初始条件（摩尔单位，$\mathrm{mol\,L^{-1}}$，M）：($C_A(t=0)$ $C_B(t=0)$) $\in${(2， 0)，(10， 2)，(2， 2)，(10， 2)，(10， 1)}。每次计算实验中，反应物和产物的浓度均记录 30 次，记录时间在 $t_0 = 0$ h 至 $t_f = 10$ h 之间等间隔分布。对于这一案例研究，动力学参数定义为： $K_A = 7\ \mathrm{M\,h^{-2}}$， $K_B = 3\ \mathrm{M\,h^{-2}}$， $K_C = 4\ \mathrm{h^{-1}}$， $K_D = 2\ \mathrm{h^{-1}}$ 和$K_E = 6\mathrm{M\,h^{-1}}$。

为模拟真实的化学系统，在计算机模拟测量值中加入高斯噪声。对于*A* 和*B*，所添加噪声的均值均为零，标准差为 0.2。加入噪声可以近似真实系统的响应。

A.2 一氧化二氮的分解

用于分析 ADoK-S 和 ADoK-W 性能的第二个案例研究是对一氧化二氮进行催化分解，其中一氧化二氮（N_2O）转化为氮气（N_2）和氧气（O_2）。反应如下所示：

$$2N_2O \rightleftharpoons 2N_2 + O_2. \tag{20}$$

描述 N_2O、 N_2 和 O_2 浓度随时间演化的动力学速率模型如下所示。该表达式直接引自 Levenspiel [38]。

$$r = -2\frac{dC_{N_2O}}{dt} = 2\frac{dC_{N_2}}{dt} = \frac{dC_{O_2}}{dt} = \frac{K_A C_{N_2O}^2}{1 + K_B C_{N_2O}} \tag{21}$$

在方程（21）中， C_{N_2O}、 C_{N_2} 和 C_{O_2} 分别表示反应物一氧化二氮以及产物氮气和氧气的浓度。该动力学速率模型的动力学参数表示为 K_i ，其中$i \in [A, B]$。为生成用于测试两个框架所需的数据集，进行了五次计算实验，每次实验采用不同的初始条件。实验采用以下初始条件（摩尔单位）：($C_{N_2O}(t=0)$ $C_{N_2}(t=0)$ $C_{O_2}(t=0)$) $\in$ {(5，0，0)，(10，0，0)，(5，2，0)，(5，0，3)，(0，2，3)}。每次实验中，反应物和产物的浓度均记录 30 次，记录时间在 $t_0 = 0$ 小时至 $t_f = 10$ 小时之间等间隔分布。对于这一案例研究，动力学参数定义为： $K_A = 2\ \mathrm{M^{-1}\ h^{-1}}$ 和 $K_B = 5\ \mathrm{M^{-1}}$。

计算机模拟数据集加入高斯噪声，以模拟实际化学系统的行为，从而生成所需的动态浓度轨迹。对于 N_2O, N_2 and O_2，所引入的噪声以零均值和 0.2 的标准差为特征，用于模拟真实系统响应中固有的变化。

A.3 甲苯加氢脱烷基

用于分析 ADoK-S 和 ADoK-W 性能的第三个案例研究是甲苯催化加氢脱烷基生成苯的反应，其中甲苯（$C_6H_5CH_3$）和氢气（H_2）转化为苯（C_6H_6）和甲烷（CH_4）。反应如下所示：

$$C_6H_5CH_3 + H_2 \rightleftharpoons C_6H_6 + CH_4. \tag{22}$$

描述 $C_6H_5CH_3$、 H_2、 C_6H_6 和 CH_4 浓度随时间演化的动力学速率模型如下。此表达式直接引自 Fogler [39]。

$$r = -\frac{dC_T}{dt} = -\frac{dC_H}{dt} = \frac{dC_B}{dt} = \frac{dC_M}{dt} = \frac{K_A C_T C_H}{1 + K_B C_B + K_C C_T} \tag{23}$$

在方程 23 中， C_T、 C_H、 C_B和 C_M分别表示反应物甲苯和氢气以及产物苯和甲烷的浓度。该动力学速率模型的动力学参数用 K_i 表示，其中$i \in [A, B, C]$。

计算实验采用以下初始条件（摩尔单位）：$(C_T(t=0)\ C_H(t=0)\ C_B(t=0)\ C_M(t=0)) \in \{(1, 8, 2, 3), (5, 8, 0, 0.5), (5, 3, 0, 0.5), (1, 3, 0, 3), (1, 8, 2, 0.5)\}$。每次实验均记录反应物和产物的浓度，共记录 30 次，时间间隔均匀，时间范围为 $t_0 = 0\,\mathrm{h}$ 至 $t_f = 10\,\mathrm{h}$。动力学参数定义如下： $K_A = 2\,\mathrm{M}^{-1}\,\mathrm{h}^{-1}$、 $K_B = 9\mathrm{M}^{-1}$ 和 $K_C = 5\mathrm{M}^{-1}$。

向计算机模拟测量值添加高斯噪声，以模拟真实的化学系统。对于 T, H, B, M，所添加噪声的均值为零，标准差为 0.2。添加噪声后，即可近似真实系统的响应。

B 模型选择分析

模型选择是所提出方法的关键环节，旨在从一系列生成并优化的模型中确定最合适的模型。因此，我们深入分析了四种不同信息准则的行为：赤池信息准则（AIC）、样本修正赤池信息准则（AIC$_c$）、汉南-奎因准则（HQC）和贝叶斯信息准则（BIC）。给定模型 m 的各项准则值通过以下公式计算：

$$\mathrm{AIC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + 2d_m \tag{24a}$$

$$\mathrm{AIC}_c, m = \mathrm{AIC}_m + \frac{2(d_m+1)(d_m+2)}{n - d_m - 2} \tag{24b}$$

$$\mathrm{HQC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + 2cd_m \log(\log(n)) \tag{24c}$$

$$\mathrm{BIC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + d_m \log(n), \tag{24d}$$

其中， $\mathcal{L}$ 表示负对数似然（NLL）， θ_m 表示模型 m的参数， $\mathcal{D}$ 表示数据集， n表示集合 $\mathcal{D}$中的数据点数量， d_m 表示 θ_m中包含的参数数量。对于汉南-奎因准则（HQC）， c 表示任何大于或等于 1 的常数，以确保模型选择一致性（即，当样本数趋于无穷大时，从模型集中选出最佳模型的概率趋于 1）。

本分析采用 A.1 节中详细介绍的假想异构化反应，并模拟其中概述的数据生成方案。

本研究考察了七种相互竞争的动力学模型，其中 r_5 代表数据生成动力学模型，也是信息准则所要选出的目标模型。

$$r_1 = K_1 C_A \tag{25a}$$
$$r_2 = K_1 C_A - K_2 C_B \tag{25b}$$
$$r_3 = \frac{K_1 C_A - K_2 C_B}{K_3 C_A} \tag{25c}$$
$$r_4 = \frac{K_1 C_A - K_2 C_B}{K_3 C_A + K_4 C_B} \tag{25d}$$
$$r_5 = \frac{K_1 C_A - K_2 C_B}{K_3 C_A + K_4 C_B + K_5} \tag{25e}$$
$$r_6 = \frac{K_1 C_A^2 - K_2 C_B - K_3 C_A}{K_4 C_A + K_5 C_B + K_6} \tag{25f}$$
$$r_7 = \frac{K_1 C_A^2 - K_2 C_B^2 - K_3 C_A - K_4 C_B}{K_5 C_A + K_6 C_B + K_7} \tag{25g}$$

本研究提出了 rein 旨在分析所提出的信息准则在噪声影响下的表现。
以及数据集的大小 .

B.1 噪声依赖性

本研究首先关注生成动力学数据集时所采用的噪声水平，探究其如何影响信息准则的表现，以及最终对模型选择产生的影响。为此，采用 A.1 节中指定的相同五个实验点，生成 13 个具有不同高斯噪声水平的动力学数据集，噪声水平由用户定义的方差 σ^2 决定。所考察的方差范围为 $\sigma^2 \in [0.04, 0.25]$，且间隔相等。

对于这些不同的数据集，利用数据重新校准七个候选模型各自的参数，随后计算信息准则值。图6绘制了最佳动力学模型 m_1 （从不包含数据生成模型的子集中选出）与数据生成模型 m_2 的信息准则值之差。在该图中，水平线 $y = 0$表示信息准则开始选择错误模型的阈值（即，在线上方，准则选择正确模型；在线下方，准则选择错误模型）。

值得注意的是，对于 13 组动力学数据中的每一组， m_1 始终都是四参数模型， r_4。观察图表可知，赤池信息准则是最能抵御噪声的准则，因为其曲线在其他所有准则之后才与水平阈值相交（也就是说，在某个噪声水平下，除赤池信息准则外，其他所有准则都会选错模型）。相反，贝叶斯信息准则的抗噪能力最弱，因为它比其他准则更早开始选错模型（$\sigma_i^2 \approx$0.06）。总体而言，在这些实验中，稳健性层级关系始终保持不变：赤池信息准则 $>$ 修正赤池信息准则 $>$ 汉南-奎因信息准则 $>$ 贝叶斯信息准则（从最稳健到最不稳健）。
所提出的层级关系是一个合理的结论，可以从数学上推导得出。鉴于已知正确模型是 r_5，且 m_1 始终为 r_4，并且样本数保持不变（$n = 150$），无论数据集如何，各信息准则对两个模型施加的惩罚之差都保持不变。更形式化地说，其中 d_m 是模型的参数数量：

$$d_m \eta(d_m = 4, n = 150)_{\text{IC}} - d_m \eta(d_m = 5, n = 150)_{\text{IC}} = k_{IC}, \tag{26}$$

其中， $\eta(d_m,\ n)$为惩罚系数， $\eta(d_m,\ n)_{\text{AIC}} = 2$、 $\eta(d_m,\ n)_{\text{AIC}_\text{c}} = \frac{2n}{n - d_m - 1}$、 $\eta(d_m,\ n)_{\text{BIC}} = \log\, n$，以及$\eta(d_m,\ n)_{\text{HQC}} = 2c\log\log\, n$。式（26）适用于 13 个动力学数据集中的每一个，其中*IC* 代表所考察的四种信息准则中的任意一种，而 k_{IC} 代表任意常数。
考虑 AIC、AICc、BIC 和 HQC 的定义： $k_{\text{AIC}} = -2$、 $k_{\text{AICc}} = -2.14$、 $k_{\text{BIC}} = -5.01$、 $k_{\text{HQC}} = -3.22$。这表明，AIC 是对复杂度更高的模型最宽容的准则（若要使 AIC 倾向于r_5 而非 r_4，则 $2(l_{n=150,\ m=4} - l_{n=150,\ m=5}) > 2$；对于 BIC，则 $2(l_{n=150,\ m=4} - l_{n=150,\ m=5}) > 5.01$）。因此，前述层级关系不仅容易理解，而且有数学依据。

然而，所有信息准则都会在某一噪声水平下开始选择错误模型，这一事实提供了有趣的见解。如前所述，模型 r_4 和 r_5 之间的惩罚项差异在

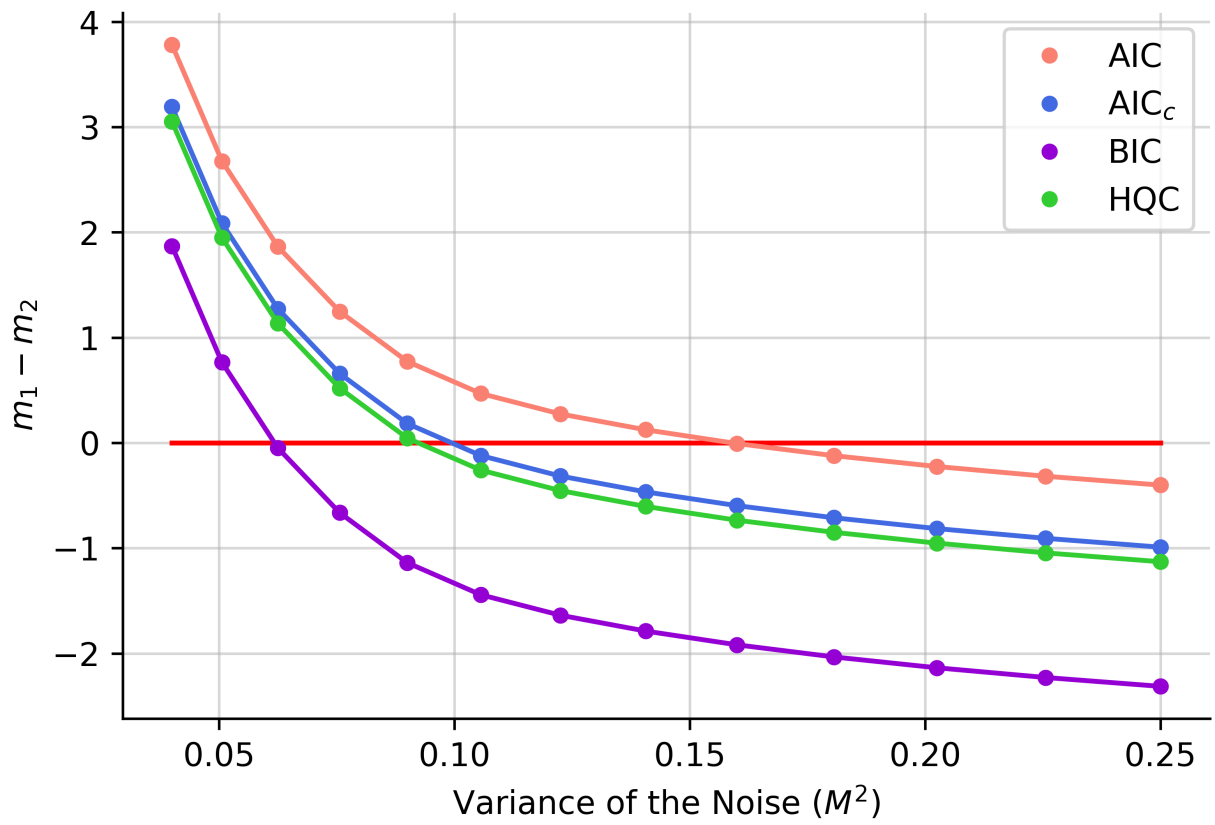


图6：绘制信息准则值之差随添加到模拟动力学数据中的高斯噪声方差 σ^2 的变化关系；该数据用于估计所有竞争模型的参数。此差值为 m_1（从不包含数据生成动力学模型的子集中选出的最佳模型）与 m_2（数据生成动力学模型）之间的差。

所有数据集中都保持固定，因此，可能影响选择转变的唯一因素就是负对数似然项。随着加性噪声增加，而样本数保持不变，负对数似然值也必然增加。此外，从图6可以推断，负对数似然值不仅会随着噪声增加而增大，而且各模型的增速也不同。

显然，五参数模型的负对数似然项随噪声方差增加而上升的速率，高于四参数模型的负对数似然项（即 $\frac{d}{d\sigma^2} l_{n=150,m=5} > \frac{d}{d\sigma^2} l_{n=120,m=4}$），因此随着噪声增大，准则会以更高的置信度选出错误模型。需要强调的是，这一发现不能推广到所有五参数模型和四参数模型，因为它仅适用于这一特定情形。

B.2 数据量依赖性

接下来考察的是样本数对信息准则计算值的影响。为此，生成了 18 个包含不同数量数据点的数据集（即模拟 A.1 节中详述的相同五个实验，但每个实验的样本数不同）。在动力学模拟中加入了方差为 $\sigma^2 = 0.2$的高斯噪声。这些计算实验的结果见图 7。

图中一个值得讨论的显著特征是，一些准则曲线在图中的不同位置彼此相交；此前展示的图中并未出现这一现象。在低数据量区间，HQC准则最接近识别出正确模型，其次依次是 AIC、BIC 和 AIC$_c$。此前提出的层级关系在这里并不成立，因为惩罚项现在有显著差异，这是由它们对样本数的依赖性所致。

然而，仅有六个样本时，图 7 **a)** 中便可见第一个交点：AIC$_c$ 开始比 BIC 更有把握地选择数据生成动力学模型，尽管两者仍然都选择了正确模型。约有十个数据点时，图 7**b)** 中出现第二个交点，此时 AIC$_c$ 选择数据生成动力学模型的把握度超过了 HQC，尽管 HQC 仍然选择了正确模型。此后，原有的层级关系重新出现并保持成立（AIC $>$ AIC$_c$ $>$ HQC $>$ BIC）。
这一发现再次得到了数学的支持。随着数据点数量增加，HQC 和贝叶斯信息准则的惩罚项也随之增大，而 AIC$_c$ 的惩罚项则减小，AIC 的惩罚项保持不变。AIC$_c$ 的惩罚项渐近趋近于 AIC 的惩罚项；随着数据点数量增加，贝叶斯信息准则的惩罚项增长速度高于汉南-奎因信息准则。

显而易见，随着收集到的过程信息增多，所有信息准则都能够识别出正确的模型（针对该模型集），这凸显了充足数据对于稳健地进行模型结构选择的重要性。必须注意图 7 中明显的“噪声”，它源于 m_1的动态定义。虽然在噪声效应研究中，m_1 始终是 r_4，但本研究中并非如此，因此 m_1的身份会发生变化。

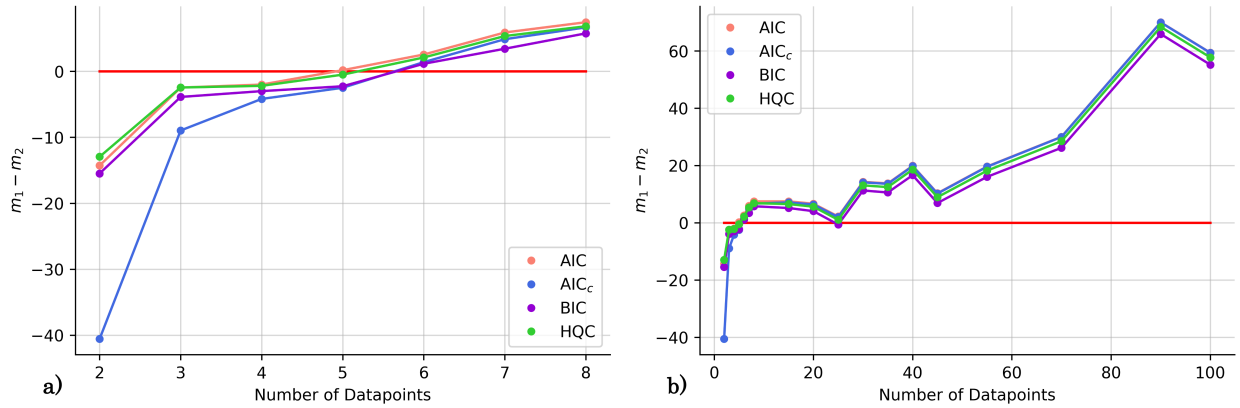


图 7：绘制 m_1 （从不包含数据生成动力学模型的子集中选出的最佳模型）与 m_2 （数据生成动力学模型）之间的差值随可用数据点数量的变化情况；这些数据点用于估计所有竞争模型的参数。**a)**：展示低数据量区间图表的更详细版本。**b)**：从整体上呈现高数据量区间内各项准则的趋势。

B.3 总结

本研究以一个简单的异构化案例研究为例，剖析了多种信息准则的行为，并全面分析了多种因素（包括数据集中的固有噪声和数据集的大小）如何影响这些准则的表现。每项结论均结合相应准则的数学形式进行了阐释并得到了依据。所有发现似乎都表明，这些信息准则的排名如下，排名越靠前，准则越优：

- 赤池信息准则，
- 样本修正赤池信息准则，
- 汉南-奎因准则，
- 贝叶斯信息准则。

然而，必须承认，这些发现具有特定情境性，未必普遍适用于不同的案例研究。此外，值得指出（并牢记）的是，模型选择准则的选择可能会延伸为一场哲学讨论，因为不同的信息准则在推导时依据的哲学假设各不相同，且回答的问题也略有不同。因此，尽管本文给出的排序可以作为有用的参考指南——事实上，所提出的方法论框架也采用了这一排序——但不应将其误解为衡量所讨论的信息准则在所有学科中的表现的绝对标准。

C 模型辨别

基于模型的实验设计（MBDoE）对于模型判别至关重要，尤其是在实验预算尚未用尽时。本文提出的方法论框架主要采用 Hunter-Reiner 准则，因为参数不确定性的评估超出了本研究的范围，因此排除了所有需要估计任意实验点处模型响应不确定性的准则。此外，在这一阶段，赤池权重设计准则的可解释性并非特别关注的重点。Hunter-Reiner 准则的主要作用是确定使两个模型响应差异最大化的最优实验。

在理想情况下，模型发现的首要目标——也是本研究的核心目的——是确定最优实验，使所提出模型与数据生成模型的响应之间产生最大的偏差。然而，实际的底层模型未知，因此必须对其进行近似。本研究考察了两种旨在近似数据生成动力学模型行为的建模方法：高斯过程状态空间模型（GPSSM）和 ADoK-S 生成的次优模型，以下称为 ADoK_2。

GPSSM 描述一种非线性动力系统，其中使用高斯过程预测系统的状态空间动力学转移（例如反应系统）[44, 45]。该模型包含基于贝叶斯原理的系统动力学非参数表示，并辅以用于控制这一非参数表示行为的超参数 [44]。选择 GPSSM 作为一种建模方法，是因为它具有非参数特性，能够有效地从小数据集中学习（这类数据集在动力学研究中很常见），因此优于循环神经网络等参数化方法 [45]。尽管 GPSSM 通常因其概率属性（即能够考虑预测不确定性）而备受青睐，但鉴于 Hunter-Reiner 准则并未考虑不确定性，这一特性在此处并无特别意义。

本研究旨在针对指定实验空间内对数据生成模型的近似效果，对这两个模型在最坏情况下的性能进行比较评估。本文采用以下流程研究三个案例（一种假想异构化反应、一氧化二氮分解和甲苯加氢脱烷基化）：

1. 按照各自章节中详述的方法，生成与给定案例研究相对应的数据；2. 对数据进行归一化，以构建 GPSSM 的训练集。输入训练数据集包含 $t \in [t_0, \cdots, t_{n-1}]$ 时所有可观测物种的浓度，输出训练数据集则包含X在 $t \in [t_1, \cdots, t_n]$ 时的物种浓度；反应系统中每种可观测物种均对应一个 GP；3. 使用整理好的训练数据集和 GPJax（一个使用 Jax 实现 GP 的 Python 软件包）[46] 训练各个 GP；4. 使用来自案例研究的数据执行一次 ADoK-S 迭代，并记录生成的次优模型，记为 ADoK_2；5. 运用 Hunter-Reiner 准则，确定使数据生成模型与训练模型（GPSSM 和 ADoK_2）之间偏差最大的实验。由此可判断，在最坏情况下哪个模型能更准确地表征真实系统。

表 2 总结了最坏情况实验中，训练模型的响应与数据生成模型之间的平方误差和（SSE）。根据这些结果，除了最佳模型之外，选择 ADoK 提出的次优动力学模型来实施基于模型的实验设计似乎是合适的。这种方法使建模者在对 ADoK-S 或 ADoK-W 给出的选项都不满意时，能够进一步生成实验。

表 2：对于每个案例研究，训练后的 GPSSM 和 ADoK_2 在最坏情况下相对于数据生成动力学模型的性能。

案例研究	GPSSM 的 SSE（M）	ADoK_2 的 SSE（M）
假想异构化反应	0.27	0.11
一氧化二氮分解	90.69	0.42
甲苯加氢脱烷基化	49.20	0.14

D ADoK-S 性能详述

D.1 假设的异构化反应

从 A.1节所述的五个初始实验开始，ADoK-S 为每个物种和实验生成、优化并选取了所示的浓度分布。下文中，$C_{i,\ j}$ 是描述实验 j期间物种 i 浓度动态演化的模型。

$$\hat{C}_{A,1}(t) = \exp\left(1.470 - \exp\left(0.642t\right)\right) + 0.634 \tag{27a}$$

$$\hat{C}_{B,1}(t) = \frac{-1.325}{\exp\left(t\right)} + 1.376 \tag{27b}$$

$$\hat{C}_{A,2}(t) = 0.075t^2 - 1.375t + 9.981 \tag{27c}$$

$$\hat{C}_{B,2}(t) = t\exp\left(-0.084t + 0.399\right) \tag{27d}$$

$$\hat{C}_{A,3}(t) = \exp\left(-0.615t\right) + 1.213 \tag{27e}$$

$$\hat{C}_{B,3}(t) = \exp\left(1.040 - \exp\left(-1.189 - t\right)\right) \tag{27f}$$

$$\hat{C}_{A,4}(t) = \frac{9.918}{0.179t + 1.078} + 0.665 \tag{27g}$$

$$\hat{C}_{B,4}(t) = -0.055t^2 + 1.128t + 2.087 \tag{27h}$$

$$\hat{C}_{A,5}(t) = \frac{\exp\left(\exp\left(\exp\left(-0.060t\right)\right)\right)}{1.499} \tag{27i}$$

$$\hat{C}_{B,5}(t) = 0.608t + 2.069. \tag{27j}$$

通过对浓度曲线进行数值微分，计算得到速率的近似值。随后，这些估算值用于生成速率模型。根据 AIC 排名，下面列出了最佳的两个模型:

$$= \hat{r}_{r_1(t)} \quad \frac{C_A(C_A - 0.578)(C_A - 0.443)}{(C_A + 0.540)(0.684C_A(C_A - 0.578) + C_B(C_B - 0.949))} \tag{28a}$$

$$\hat{r}_2(t) = \frac{(C_A - 0.579)(C_A - 0.519)(C_A - 0.161)}{(C_A + 0.293)(C_B(C_B - 0.988) + 0.684(C_A - 0.519)(C_A - 0.161))}. \tag{28b}$$

利用这两个速率模型，通过基于模型的实验设计提出了一项新实验。拟议实验为 $(C_A(t = 0)\ C_B(t = 0)) = (4。926，0。000)$ M。对于第六次实验，ADoK-S 生成、优化并选取了所示的浓度曲线。

$$\hat{C}_{A,6}(t) = \exp\left(1.258 - \frac{t}{2.151}\right) + 1.458 \tag{29a}$$

$$\hat{C}_{B,6}(t) = \frac{3.436}{\exp\left(\exp\left(-0.707t + 0.824\right)\right)}. \tag{29b}$$

通过近似第六次实验的速率测量值，并将估计值与先前的数据集合并，ADoK-S 揭示了数据生成速率模型的结构和相似的参数:

$$\hat{r}^* = \frac{8.666C_A - 3.642C_B}{4.976C_A + 2.525C_B + 7.003}. \tag{30}$$

D.2 一氧化二氮的分解

从 A.2 节所述的五个初始实验出发，ADoK-S 为每个实验中的每种物种生成、优化并选出了ˆ呈现的浓度曲线。下方的 $C_{i,\ j}$ 是描述实验 j期间物种 i 浓度动态演化的模型

$$\hat{C}_{N_2O,1}(t) = \exp\left(1.602 - \frac{t}{2.128}\right) + 0.309 \tag{31a}$$

$$\hat{C}_{N_2,1}(t) = 4.866 - \frac{4.703}{\exp(0.383t)} \tag{31b}$$

$$\hat{C}_{O_2,1}(t) = \frac{t}{0.347t + 0.896} + 0.152 \tag{31c}$$

$$\hat{C}_{N_2O,2}(t) = \exp\left(2.287 - \frac{t}{2.486}\right) + 0.113 \tag{31d}$$

$$\hat{C}_{N_2,2}(t) = 9.863 - \frac{2.262}{\exp(0.390t)} \tag{31e}$$

$$\hat{C}_{O_2,2}(t) = t\exp(\exp(-0.094t)) - t \tag{31f}$$

$$\hat{C}_{N_2O,3}(t) = \exp(1.520 - 0.418t) + 0.199 \tag{31g}$$

$$\hat{C}_{N_2,3}(t) = 6.970 - 5.032\exp(-0.385t) \tag{31h}$$

$$\hat{C}_{O_2,3}(t) = \frac{t}{0.318t + 0.850} \tag{31i}$$

$$\hat{C}_{N_2O,4}(t) = \exp(1.533 - 0.380t) + 0.187 \tag{31j}$$

$$\hat{C}_{N_2,4}(t) = 1.560t\exp(-0.122t) \tag{31k}$$

$$\hat{C}_{O_2,4}(t) = 5/343 - \exp(1.741 - \exp(0.270t)) \tag{31l}$$

$$\hat{C}_{N_2O,5}(t) = 0.077 \tag{31m}$$

$$\hat{C}_{N_2,5}(t) = 2.009 \tag{31n}$$

$$\hat{C}_{O_2,5}(t) = 2.978. \tag{31o}$$

通过对浓度曲线进行数值微分，计算出速率的近似值。随后，这些估计值用于生成速率模型。根据 AIC 排名选出的两个最佳模型如下所示：

$$\hat{r}_1(t) = \frac{C_{N_2O}(0.416C_{N_2O} - 0.034)}{C_{N_2O} + 0.200} \tag{32a}$$

$$\hat{r}_2(t) = 0.061 - 0.404C_{N_2O}. \tag{32b}$$

基于这两个速率模型，使用基于模型的实验设计提出了一项新实验。所提出的实验为 ($C_{N_2O}(t=0)$ $C_{N_2}(t=0)$ $C_{O_2}(t=0)$) =(0.000，1.641，1.095) M。对于第六次实验，ADoK-S 生成、优化并选取了所展示的浓度曲线。

$$\hat{C}_{N_2O,6}(t) = 0.158 \tag{33a}$$

$$\hat{C}_{N_2,6}(t) = 1.631 \tag{33b}$$

$$\hat{C}_{O_2,6}(t) = 1.085. \tag{33c}$$

通过近似第六次实验的速率测量值，并将估计值与先前的数据集合并，ADoK-S 揭示了与数据生成速率模型相同的结构和相近的参数：

$$\hat{r}^* = \frac{1.937C_{N_2O}^2}{1 + 4.803C_{N_2O}}. \tag{34}$$

D.3 甲苯加氢脱烷基化

从 A.3节所述的五个初始实验开始，ADoK-S 为每个物种和实验生成、优化并选出了所呈现的浓度曲线ˆ。下方的 $C_{i,\ j}$ 是描述实验 j期间物种 i 浓度动态演化的模型

$$\hat{C}_{T,1}(t) = \exp(0.164 - t) + 0.102 \tag{35a}$$
$$\hat{C}_{H,1}(t) = 7.006 + \exp(-0.534t) \tag{35b}$$
$$\hat{C}_{B,1}(t) = 3.021 + \frac{-0.525}{0.430 + t} \tag{35c}$$
$$\hat{C}_{M,1}(t) = 3.940 - \exp(-0.525(t + 0.115)) \tag{35d}$$
$$\hat{C}_{T,2}(t) = \frac{6.652}{1.172 + \frac{t}{1.592}} - 0.407 \tag{35e}$$
$$\hat{C}_{H,2}(t) = \exp(\exp(-0.229(t - 2.379))) + 2.424 \tag{35f}$$
$$\hat{C}_{B,2}(t) = \frac{t}{0.437 + 0.175t} \tag{35g}$$
$$\hat{C}_{M,2}(t) = \exp\left(1.775 - \frac{1.599}{\exp(-0.414) + t}\right) \tag{35h}$$
$$\hat{C}_{T,3}(t) = 1.295 + \exp(\exp(-0.197(t - 1.305))) \tag{35i}$$
$$\hat{C}_{H,3}(t) = \exp(\exp(-0.070t^2)) - 0.323 \tag{35j}$$
$$\hat{C}_{B,3}(t) = \exp\left(\frac{\exp(1.004)}{-0.793 - t} + 1.192\right) \tag{35k}$$
$$\hat{C}_{M,3}(t) = \exp\left(\frac{-1.330}{\exp(0.472t)} + 1.115\right) \tag{35l}$$
$$\hat{C}_{T,4}(t) = \exp(-0.151\exp(t)) + 0.058 \tag{35m}$$
$$\hat{C}_{H,4}(t) = \frac{-2.282}{-2.029 - t^2} + 1.948 \tag{35n}$$
$$\hat{C}_{B,4}(t) = \exp\left(\frac{-2.523}{\exp(t)}\right) \tag{35o}$$
$$\hat{C}_{M,4}(t) = 0.230t - 0.016t^2 + 3.318 \tag{35p}$$
$$\hat{C}_{T,5}(t) = \frac{t + 0.493}{\exp(t)} + 0.063 \tag{35q}$$
$$\hat{C}_{H,5}(t) = 6.906 + \exp(-0.376t) \tag{35r}$$
$$\hat{C}_{B,5}(t) = 3.010 - \exp(-0.442t) \tag{35s}$$
$$\hat{C}_{M,5}(t) = \frac{\frac{-0.797}{0.557+t} + 1.574}{1.009}. \tag{35t}$$

通过对浓度曲线进行数值微分，计算出速率的近似值。随后，这些估算值被用于生成速率模型。根据 AIC 排名选出的两个最佳模型如下：

$$\hat{r}_1(t) = -0.052C_TC_H \tag{36a}$$
$$\hat{r}_2(t) = 0.020C_T - 0.056C_TC_H. \tag{36b}$$

利用这两个速率模型，通过基于模型的实验设计提出了一项新实验。拟议实验为 ($C_T(t=0)$ $C_H(t=0)$ $C_B(t=0)$ $C_M(t=0)$) =(1.948, 7.503, 1.232, 2.504) M。对于第六次实验，ADoK-S生成、优化并筛选出了所示的浓度曲线。

$$\hat{C}_{T,6}(t) = \exp\left(\frac{1.216 - t}{\exp(0.645)}\right) \tag{37a}$$

$$\hat{C}_{H,6}(t) = \frac{4.934}{\exp(t) + \exp(0.330)} + 5.591 \tag{37b}$$

$$\hat{C}_{B,6}(t) = \frac{3.004}{\exp(\exp(-0.862t))} \tag{37c}$$

$$\hat{C}_{M,6}(t) = 4.896 + \frac{4.048}{-1.506 - t}. \tag{37d}$$

通过近似第六次实验的速率测量值，并将估计值与之前的数据集合并，ADoK-S仍无法找出正确的模型结构。本轮迭代中排名最靠前的两个模型如下所示:

$$\hat{r}_1(t) = \frac{-1.000 C_T C_H}{1.286 C_T + 12.586} \tag{38a}$$

$$\hat{r}_2(t) = -0.034 C_T(-0.741 C_B^2 + 1.886 C_B + 1.886 C_H - C_T + 1.612) - 0.021. \tag{38b}$$

基于这两个速率模型，使用基于模型的实验设计提出了一项新实验。拟议实验为 (*C*$_T$(*t* =0) *C*$_H$(*t* = 0) *C*$_B$(*t* = 0) *C*$_M$(*t* = 0)) =(2.560，5.654，0.341，2.337) M。对于第七次实验，ADoK-S 生成、优化并选出了所示的浓度曲线。

$$\hat{C}_{T,7}(t) = \exp(\exp(0.033 - t)) - 0.101t \tag{39a}$$

$$\hat{C}_{H,7}(t) = \exp(\exp(-0.298t - 0.172)) + 2.186 \tag{39b}$$

$$\hat{C}_{B,7}(t) = \exp(1.024) - \exp(-0.442t + 0.802) \tag{39c}$$

$$\hat{C}_{M,7}(t) = 4.818 - \exp(-0.434(t - \exp(0.693))). \tag{39d}$$

通过近似第七次实验的速率测量值，并将估计值与先前的数据集合并，ADoK-S 揭示了与数据生成速率模型相似的结构和参数:

$$\hat{r}^* = \frac{1.669 C_T C_H}{1 + 7.347 C_B + 4.439 C_T}. \tag{40}$$

E ADoK-W 性能详述

E.1 假设的异构化反应

从最初五次实验出发，ADoK-W 生成并优化了大量速率模型，并输出了以下两个最佳模型（根据赤池信息准则排名）：

$$\hat{r}_1(t) = \frac{0.025 C_A^2 + 0.967 C_A - 0.597 C_B + 0.438}{C_A} \tag{41a}$$

$$\hat{r}_2(t) = \frac{1.143 - 0.490_B}{C_A}. \tag{41b}$$

利用这两个速率模型，通过基于模型的实验设计提出了一项新实验。拟议实验为 (*C*$_A$(*t* =0) *C*$_B$(*t* = 0)) = (7。319，2。000) M。将第六次实验收集的测量数据合并后，ADoK-W 几乎成功地重新发现了数据生成动力学速率模型；所选模型的分子中多出了一个参数。尽管如此，有观点认为，这个额外的动力学参数远小于其余参数，因此在进一步研究或模型降阶时必然会被删除。

$$\hat{r}^* = \frac{9.998 C_A - 4.496 C_B + 0.386}{6.038 C_A + 2.137 C_B + 7.892} \tag{42}$$

E.2 一氧化二氮的分解

从最初五次实验出发，ADoK-W 生成并优化了大量速率模型，并根据赤池信息准则排名，输出了以下两个最佳模型：

$$\hat{r}_1(t) = \frac{(0.411C_{N_2O} - 0.062)(C_{N_2O} - 0.189) - 0.002}{C_{N_2O} - 0.189} \tag{43a}$$

$$\hat{r}_2(t) = 0.411C_{N_2O} - 0.068. \tag{43b}$$

基于这两个速率模型，使用基于模型的实验设计提出了一项新实验。拟议实验为（$C_{N_2O}(t=0)$ $C_{N_2}(t=0)$ $C_{O_2}(t=0)$）=(0.189,0.913, 0.926) M。将第六次实验收集的测量数据合并后，ADoK-W 仍无法发现数据生成模型。相反，根据 AIC 排名，该方法输出了以下模型，分别作为最佳和次佳模型：

$$\hat{r}_1(t) = \frac{C_{N_2O}(0.927C_{N_2O} - 0.147)}{2.252C_{N_2O} + 0.046} \tag{44a}$$

$$\hat{r}_2(t) = 0.414C_{N_2O} - 0.074. \tag{44b}$$

基于这两个速率模型，使用基于模型的实验设计提出了一项新实验。拟议实验为（$C_{N_2O}(t=0)\, C_{N_2}(t=0)\, C_{O_2}(t=0)$）= (0.189,0.913, 0.926) M。将第七次实验收集的测量数据合并后，ADoK-S 发现了数据生成速率模型的结构和相近的参数：

$$\hat{r}^* = \frac{1.584C_{N_2O}^2}{1 + 3.798C_{N_2O}}. \tag{45}$$

E.3 甲苯加氢脱烷基化

如本文正文所述，ADoK-W 仅用最初的五次实验，便重新发现了此案例研究的真实动力学速率模型。所发现的模型在结构上与数据生成模型完全相同，其动力学数值也与之非常相近：

$$\hat{r}^* = \frac{1.124C_TC_H}{1 + 4.932C_B + 2.928C_T}. \tag{46}$$

用于生成本文所示所有结果和图表的代码可通过 https://github.com/ MACServia/auto_discov_kin_rate_models 获取。

参考文献

[1] C. Song、T. Koren、P. Wang 和 A. L. Barabási。人类流动规模特性的建模。*Nat. Phys.*，6(10):818–823，2010 年 9 月。doi:10.1038/nphys1760。[2] D. Brockmann、L. Hufnagel 和 T. Geisel。人类出行的标度定律。*Nature*，439(7075):462–465，2006 年 1 月。doi:10.1038/nature04292。[3] L. C. Franssen、T. Lorenzi、A. E. F. Burgess 和 M. A. J. Chaplain。癌症转移扩散建模的数学框架。*Bull. Math. Biol.*，81(6):1965–2010，2019 年 3 月。doi:10.1007/s11538-019-00597-x。[4] D. H. Margarit 和 L. Romanelli。用于理解转移路径的吸收马尔可夫链数学模型。*Biomath*，5(1)，2016 年 8 月。doi:10.11145/j.biomath.2016.07.281。[5] N. S. Schbib、M. A. García、C. E. Gígola 和 A. F. Errazu。乙烯生产中前端乙炔加氢的动力学。*Ind. Eng. Chem. Res.*，35(5):1496–1505，1996 年 1 月。doi:10.1021/ie950600k。[6] G.C. Battiston、L. Dalloro 和 G.R. Tauszik。乙炔选择性加氢催化剂的性能与老化：微型中试装置研究。*Appl. Catal.*，2(1-2):1–17，1982 年 1 月。doi:10.1016/0166-9834(82)80170-x。[7] R. E. Baker、J. M. Peña、J. Jayamohan 和 A. Jérusalem。机理模型与机器学习：一场值得生物学界参与的较量？*Biol. Lett.*，14(5):20170660，2018 年 5 月。doi:10.1098/rsbl.2017.0660。

[8] K. V. Gernaey。发酵过程开发与运行中的过程系统工程视角。*Comput. Aided Chem.Eng.*，第 123–130 页，2015 年。doi:10.1016/b978-0-444-63578-5.50016-5。[9] I. Jimenez del Val、J. M. Nagy 和 C. Kontoravdi。用于描述成熟中的高尔基体内单克隆抗体 N-连接糖基化和核苷酸糖供体转运的动态数学模型。*Biotechnol.Prog.*，27(6):1730–1743，2011 年。doi:https://doi.org/10.1002/btpr.688。[10] P. M. Jedrzejewski、I. Jimenez Jimenez del Val、A. Constantinou、A. Dell、S. M. Haslam、K. M. Polizzi 和 C. Kontoravdi。迈向糖型调控：连接细胞外代谢物与抗体糖基化的模型框架。*Int. J. Mol. Sci.*，15(3):4492–4522，2014 年。ISSN 1422-0067。doi:10.3390/ijms15034492。[11] R. T. Giessmann、N. Krausch、F. Kaspar、M. N. Cruz Bournazou、A. Wagner、P. Neubauer 和 M. Gimpel。嘧啶核苷磷酸化酶催化的磷酸解裂解动态建模。*Processes*，7(6)，2019 年。ISSN 2227-9717。doi:10.3390/pr7060380。[12] H. R. Sant Anna、A. G. Barreto、F. W. Tavares 和 M. B. de Souza。用于甲烷-氮气分离的变压吸附单元机器学习模型与优化。*Comput ChemEng*，104:377–391，2017 年 9 月。doi:10.1016/j.compchemeng.2017.05.006。[13] D. Zhang、E. A. del Rio-Chanona、P. Petsagkourakis 和 J. Wagner。用于生物过程在线仿真与优化的基于物理和数据驱动的混合建模。*Biotechnol. Bioeng.*，116(11):2919–2930，2019 年 7 月。doi:10.1002/bit.27120。[14] E. A. del Rio-Chanona、X. Cong、E. Bradford、D. Zhang 和 K. Jing。动态生物过程仿真的先进物理模型与数据驱动模型综述：藻菌联合体废水处理案例研究。*Biotechnol.Bioeng.*，116(2):342–353，2018 年 12 月。doi:10.1002/bit.26881。[15]S. Y. Park、C. H. Park、D. H. Choi、J. K. Hong 和 D. Y. Lee。用于先进生物制造的哺乳动物细胞培养生物过程数字孪生。*Curr. Opin. Chem. Eng.*，33:100702，2021 年 9 月。doi:10.1016/j.coche.2021.100702。[16] Y. Sun、W. Nathan-Roberts、T. D. Pham、E. Otte 和 U. Aickelin。小数据条件下用于生物制造过程建模的多保真高斯过程。*arXiv:2211.14493*，2022 年。doi:10.48550/ARXIV.2211.14493。[17]P. Petsagkourakis、I. O. Sandoval、E. Bradford、D. Zhang 和 E.A. del Rio-Chanona。用于批式生物过程优化的强化学习。*Comput Chem Eng*，133:106649，2020 年 2 月。doi:10.1016/j.compchemeng.2019.106649。[18] E. A. del Rio-Chanona、J. L. Wagner、H. Ali、F. Fiorelli、D. Zhang 和 K. Hellgardt。用于微藻生物燃料生产和光生物反应器设计的基于深度学习的代理建模与优化。*AIChE J*，65 (3):915–923，2018 年 12 月。doi:10.1002/aic.16473。[19] G. Wu、M. Á. de Carvalho Servia 和 M. Mowbray。多级供应链库存管理的分布式强化学习。*Digital Chemical Engineering*，6:100073，2023 年 3 月。doi:10.1016/j.dche.2022.100073。[20] P. Natarajan、R. Moghadam 和 S. Jagannathan。基于在线深度神经网络的叶黄素生物过程反馈控制。*J. Process Control*，98:41–51，2021 年 2 月。doi:10.1016/j.jprocont.2020.11.011。[21] M. Mowbray、H. Kay、S. Kay、P. Castro Caetano、A. Hicks、C. Mendoza、A. Lane、P. Martin 和 D. Zhang。基于概率机器学习的批式过程产品质量预测软测量仪。*ChemometrIntell Lab Syst*，228:104616，2022 年 9 月。doi:10.1016/j.chemolab.2022.104616。[22] S. Kay、H. Kay、M. Mowbray、A. Lane、C. Mendoza、P. Martin 和 D. Zhang。融合自编码器与异方差噪声神经网络的批式过程软测量仪设计。*Ind. Eng. Chem. Res.*，61(36): 13559–13569，2022 年 9 月。doi:10.1021/acs.iecr.2c01789。[23] P. Kadlec、B. Gabrys 和 S. Strandt。过程工业中的数据驱动软测量仪。*Comput Chem Eng*，33(4): 795–814，2009 年 4 月。doi:10.1016/j.compchemeng.2008.12.012。[24] F. Vega-Ramon、X. Zhu、T. R. Savage、P. Petsagkourakis、K. Jing 和 D. Zhang。不确定性条件下酵母虾青素生产的动力学与混合建模。*Biotechnol. Bioeng.*，118(12):4854–4866，2021 年 10 月。doi:10.1002/bit.27950。[25] M. R. Mowbray、C. Wu、A. W. Rogers、E. A. del Rio-Chanona 和 D. Zhang。用于生物过程动力学辨识的基于强化学习的混合建模框架。*Biotechnol.Bioeng.*，120(1):154–168，2022 年 10 月。doi:10.1002/bit.28262。[26] D. Zhang、T. R. Savage 和 B. A. Cho。结合模型结构辨识与混合建模，实现光生产过程的预测仿真与优化。*Biotechnol. Bioeng.*，117(11):3356–3367，2020 年 7 月。doi:10.1002/bit.27512。

[27] C. Haider、F.O. de Franca、B. Burlacu 和 G. Kronberger。用于符号回归的形状约束多目标遗传编程。应用软计算，132：109855，2023年1月。doi:10.1016/j.asoc.2022.109855。[28] Z. T. Wilson 和 N. V. Sahinidis。ALAMO 机器学习方法。计算机与化学工程，106：785–795，2017年11月。doi:10.1016/j.compchemeng.2017.02.010。[29] S. L. Brunton、J. L. Proctor 和 J. N. Kutz。通过非线性动力系统的稀疏辨识，从数据中发现控制方程。美国国家科学院院刊，113(15)：3932–3937，2016年3月。doi:10.1073/pnas.1517384113。[30] C. J. Taylor、M. Booth、J. A. Manson、M. J. Willis、G. Clemens、B. A. Taylor、T. W. Chamberlain 和 R. A. Bourne。快速自动确定反应模型和动力学参数。化学工程学报，413：127017，2021年6月。doi:10.1016/j.cej.2020.127017。[31] P. Neumann、L. Cao、D. Russo、V. S. Vassiliadis 和 A. A. Lapkin。用于从实验数据中识别物理化学定律的一种新符号回归方法。化学工程学报，387：123412，2020年5月。doi:10.1016/j.cej.2019.123412。[32] T. Forster、D. Vázquez、M. N. Cruz-Bournazou、A. Butté 和 G. Guillén-Gosálbez。通过基于 MINLP 的 S 系统形式符号回归对生物过程建模。计算机与化学工程，170：108108，2023年2月。doi:10.1016/j.compchemeng.2022.108108。[33] J. R. Koza。遗传编程作为通过自然选择编写计算机程序的手段。统计计算，4(2)，1994年6月。doi:10.1007/bf00175355。[34] Marco Virgolin 和 Solon P. Pissis。符号回归是 NP-hard，2022。[35] Miles Cranmer。利用 PySR 和 SymbolicRegression.jl 实现科学领域的可解释机器学习，2023。[36] William G. Hunter 和 Albey M. Reiner。区分两个竞争模型的实验设计。计量学，7(3)：307–323，1965年8月。doi:10.1080/00401706.1965.10490265。[37] G. B. Marin、G. S Yablonsky 和 D. Constales。化学反应动力学：解码复杂性。John Wiley & Sons，2019。[38] O. Levenspiel。化学反应工程。John Wiley & Sons，美国田纳西州纳什维尔，第三版，1998年8月。[39] H. S. Fogler。化学反应工程基础。Prentice Hall，美国宾夕法尼亚州费城，第五版，2016年1月。[40] Robert Mee。两水平因子实验综合指南。Springer Science & Business Media，2009。[41] L. Schrecker、J. Dickhaut、C. Holtze、P. Staehle、M. Vranceanu、K. Hellgardt 和 K. K. Hii。通过瞬态流动实验发现 Knorr 吡唑合成中出乎意料的复杂反应路径。反应化学与工程，8(1)：41–46，2023。doi:10.1039/d2re00271j。[42] Conor Waldron、Arun Pankajakshan、Marco Quaglio、Enhong Cao、Federico Galvanin 和 Asterios Gavriilidis。基于模型设计瞬态流动实验，以识别动力学参数。反应化学与工程，5：112–123，2020。doi:10.1039/C9RE00342H。[43] Dimitris Bertsimas 和 Wes Gurnee。通过混合整数优化学习稀疏非线性动力学。非线性动力学，111(7)：6585–6604，2023年1月。doi:10.1007/s11071-022-08178-9。[44] Roger Frigola、Fredrik Lindsten、Thomas B. Schön 和 Carl E. Rasmussen。利用粒子随机近似 EM 辨识高斯过程状态空间模型。*IFAC* 论文集，47(3)：4097–4102，2014。ISSN 1474-6670。doi:https://doi.org/10.3182/20140824-6-ZA-1003.01843。第19届 IFAC 世界大会。[45] Stefanos Eleftheriadis、Tom Nicholson、Marc Deisenroth 和 James Hensman。高斯过程状态空间模型的辨识。载于 I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vish-wanathan 和 R. Garnett 编，神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017。网址：https://proceedings.neurips.cc/paper_files/paper/2017/file/1006ff12c465532f8c574aeaa4461b16-Paper.pdf。[46] Thomas Pinder 和 Daniel Dodd。GPJax：基于 JAX 的高斯过程框架。开放源码软件期刊，7(75)：4455，2022。doi:10.21105/joss.04455。网址：https://doi.org/10.21105/joss.04455。