THE AUTOMATED DISCOVERY OF KINETIC RATE MODELS – METHODOLOGICAL FRAMEWORKS: **S.I.**

Miguel Ángel de Carvalho Servia
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
m.de-carvalho-servia21@imperial.ac.uk

Ilya Orson Sandoval
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
o.sandoval-cardenas20@imperial.ac.uk

King Kuok (Mimi) Hii
Department of Chemistry
Imperial College London
White City, London, W12 0BZ, UK
mimi.hii@imperial.ac.uk

Klaus Hellgardt
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
k.hellgardt@imperial.ac.uk

Dongda Zhang *
Department of Chemical Engineering
The University of Manchester
Manchester, M13 9PL, UK
dongda.zhang@manchester.ac.uk

Ehecatl Antonio del Rio Chanona *
Department of Chemical Engineering
Imperial College London
South Kensington, London, SW7 2AZ, UK
a.del-rio-chanona@imperial.ac.uk

April 3, 2024

A Model Selection Analysis

Model selection is a pivotal aspect of the proposed methodologies, determining the most suitable model from an array of generated and optimized models. Consequently, an in-depth analysis on the behavior of four different information criteria is conducted: Akaike information criterion (AIC), sample corrected Akaike information criterion (AIC_c), Hannan-Quinn criterion (HQC) and Bayesian information criterion (BIC). Each criteria value for a given model m is calculated using the equations presented below:

$$\text{AIC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + 2d_m \tag{1a}$$

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

$$\text{HQC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + 2cd_m \log(\log(n)) \tag{1c}$$

$$\text{BIC}_m = 2\mathcal{L}(\theta_m \mid \mathcal{D})_m + d_m \log(n), \tag{1d}$$

where $\mathcal{L}$ represents the negative log-likelihood (NLL), θ_m are the parameters of model m, $\mathcal{D}$ represents the data set, n represents the number of data points within set $\mathcal{D}$, and d_m represents the number of parameters contained in θ_m. For HQC, c stands for any constant equal to or greater than 1 to ensure model selection consistency (i.e., the best model within a model set is selected with probability going to one if the number of samples tend to infinity).

The analysis employs the hypothetical isomerization reaction detailed in Section 3.3, mimicking the data generation protocol as outlined therein.

This study examines seven competing kinetic models, with r_5 representing the data-generating kinetic model and the desired selection target for the information criteria.

$$r_1 = k_1 C_A \tag{2a}$$
$$r_2 = k_1 C_A - k_2 C_B \tag{2b}$$
$$r_3 = \frac{k_1 C_A - k_2 C_B}{k_3 C_A} \tag{2c}$$
$$r_4 = \frac{k_1 C_A - k_2 C_B}{k_3 C_A + k_4 C_B} \tag{2d}$$
$$r_5 = \frac{k_1 C_A - k_2 C_B}{k_3 C_A + k_4 C_B + k_5} \tag{2e}$$
$$r_6 = \frac{k_1 C_A^2 - k_2 C_B - k_3 C_A}{k_4 C_A + k_5 C_B + k_6} \tag{2f}$$
$$r_7 = \frac{k_1 C_A^2 - k_2 C_B^2 - k_3 C_A - k_4 C_B}{k_5 C_A + k_6 C_B + k_7} \tag{2g}$$

The study presented herein aims at analyzing the behavior of the presented information criteria with respect to the noise and size of a data set.

A.1 Noise Dependency

The initial focus of this study is the exploration of the noise level employed in the generation of the kinetic data set, how it influences the behavior of the information criteria, and its eventual effect on model selection. For this, the same five experimental points specified in Section 3.3 are used to create 13 distinct kinetic data sets with varying degrees of Gaussian noise, dictated by the user-defined variance σ^2. The variance range explored is $\sigma^2 \in [0.04, 0.25]$, at equally spaced intervals.

For each of these unique data sets, the data is utilized to re-calibrate the parameters for each of the seven candidate models, and subsequently the information criteria values are computed. Figure 1 presents a plot that illustrates the difference of information criteria value between the best kinetic model m_1 (chosen from a subset that excludes the data-generating model) and the data-generating model m_2. Within this graph, the horizontal line $y = 0$ serves as the threshold at which an information criterion starts to select the incorrect model (i.e., above this line, the criterion selects the right model, below it, the criterion selects the wrong model).

It is worth noting that for each of the 13 sets of kinetic data, m_1 is consistently the 4-parameter model, r_4. Upon examining the graph, the AIC emerges as the most noise-resilient criterion, as its profile line intersects the horizontal threshold after all the other criteria (i.e., for a certain noise level, all other criteria select the wrong model, except AIC). Conversely, BIC exhibits the least noise resilience, as it initiates wrong model selection prior to the other criteria ($\sigma_i^2 \approx 0.06$). In general, a robustness hierarchy is preserved across these experiments, where AIC $>$ AICc $>$ HQC $>$ BIC (from most to least robust).

The proposed hierarchy is a reasonable conclusion which can be deduced from mathematics. Given that the correct model is known to be r_5, that m_1 was invariably r_4, and that the number of samples remains constant ($n = 150$), the difference of the penalty imposed by each information criterion to both models, regardless of the data set, remains constant. More formally, where d_m is the number of parameters of a model:

$$d_m \eta(d_m = 4, n = 150)_{\text{IC}} - d_m \eta(d_m = 5, n = 150)_{\text{IC}} = k_{IC}, \tag{3}$$

where $\eta(d_m, n)$ is the penalty coefficient, with $\eta(d_m, n)_{\text{AIC}} = 2$, $\eta(d_m, n)_{\text{AIC}_\text{c}} = \frac{2n}{n - d_m - 1}$, $\eta(d_m, n)_{\text{BIC}} = \log n$, and $\eta(d_m, n)_{\text{HQC}} = 2c \log \log n$. Eq. (3) holds for each of the 13 kinetic data sets, where ***IC*** stands for any of the four examined information criteria, and k_{IC} represents an arbitrary constant.

Considering the definitions of AIC, AICc, BIC and HQC: $k_{\text{AIC}} = -2$, $k_{\text{AICc}} = -2.14$, $k_{\text{BIC}} = -5.01$, $k_{\text{HQC}} = -3.22$. This demonstrates that AIC is the most tolerant criterion towards models of higher complexity (for AIC to favour r_5 over r_4, $2(l_{n=150,m=4} - l_{n=150,m=5}) > 2$; for BIC, $2(l_{n=150,m=4} - l_{n=150,m=5}) > 5.01$). Consequently, the aforementioned hierarchy is not only comprehensible but is also mathematically grounded.

However, the fact that all information criteria begin to choose the incorrect model at a certain noise level offers interesting insights. As previously explained, the penalty term's difference between the models r_4 and r_5 stays fixed across all data sets, hence the only element potentially influencing a selection shift is the NLL term. As the additive

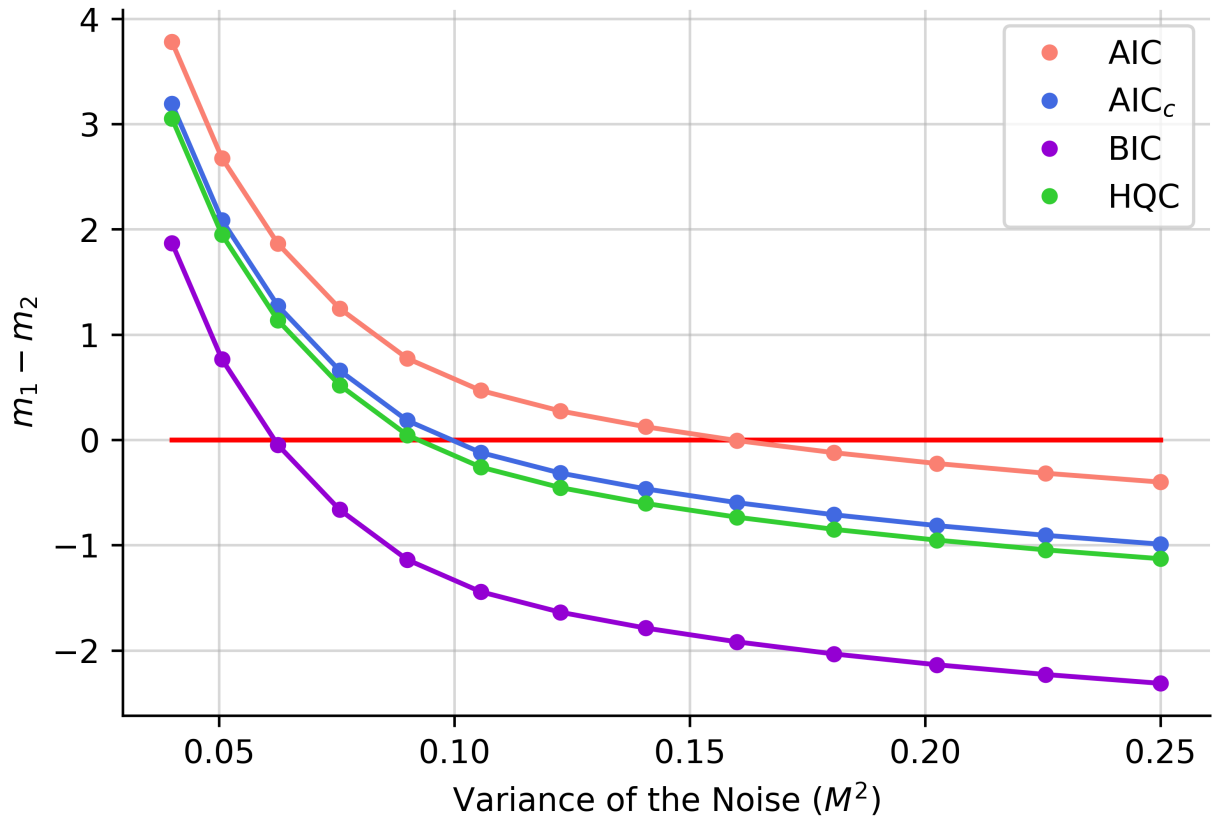

Figure 1: Plot of the difference of information criteria value between m_1 (the best model chosen from a subset that does not include the data-generating kinetic mode) and m_2 (the data-generating kinetic model) with respect to the variance of the Gaussian noise, σ^2, added to the kinetic data simulated, which was used to estimate the parameters of all rival models.

noise increases and the number of samples remains unchanged, the NLL values must also increase. Furthermore, deducing from Figure 1, not only do NLL values increase with noise, but they also increase at different rates for each model.

It becomes evident that the NLL term for the 5-parameter model rises at a higher rate in relation to the noise variance than the NLL term for the 4-parameter model (i.e., $\frac{\mathrm{d}l_{n=150,m=5}}{\mathrm{d}\sigma^2} > \frac{\mathrm{d}l_{n=120,m=4}}{\mathrm{d}\sigma^2}$) causing the criteria to select the wrong model more confidently as the noise is increased. It is important to underscore that this insight cannot be generalized to all 5-parameter and 4-parameter models, as it remains specific to this particular case.

A.2 Quantity of Data Dependency

The next aspect scrutinized is the influence of the number of samples on the values calculated by the information criteria. To investigate this, 18 data sets with varying numbers of data points are generated (i.e., the same five experiments detailed in Section 3.3 are simulated, but with different number of samples per experiment). Gaussian noise, with a variance of $\sigma^2 = 0.2$, is introduced into the kinetic simulations. The outcomes of these computational experiments are presented in Figure 2.

A noteworthy feature from the graph deserving of discussion is how some criterion profiles intersect one another at different points of the plot, a phenomenon not identified in the previously presented graph. In the low-data regime, the HQC criterion is closest to identifying the correct model, followed closely by AIC, BIC, and AIC$_c$, in that order. The previously proposed hierarchy does not hold here, as the penalty terms now significantly differ, given their dependency on the number of samples.

However, with a mere six samples, the first intersection becomes visible in Figure 2 **a)**, where AIC_c starts selecting the data-generating kinetic model with greater confidence than the BIC, although both are still selecting the correct model. With approximately ten data points, a second intersection emerges in Figure 2 **b)**, with AIC_c now choosing the data-generating kinetic model with more certainty than the HQC, albeit the HQC is still selecting the right model. After this point, the original hierarchy reappears and is respected (AIC $>$ AIC_c $>$ HQC $>$ BIC).

This finding is once again grounded by mathematics. As the number of data points increases, the penalty terms for HQC and BIC also increase, while the penalty term for AIC_c decreases and that for AIC remains constant. The penalty term for AIC_c asymptotically approaches that of the AIC, and BIC's penalty term increases at a higher rate than that of HQC as the number of data points increases.

It is evident that with the collection of more process information, all information criteria are capable of identifying the correct model (for this model set), underscoring the importance of sufficient data for robust model structure selection. It is vital to acknowledge the apparent "noise" in Figure 2, which stems from the dynamic definition of m_1. While in the noise effect study m_1 is always r_4, this phenomenon does not hold in this study, leading to the changing identity of m_1.

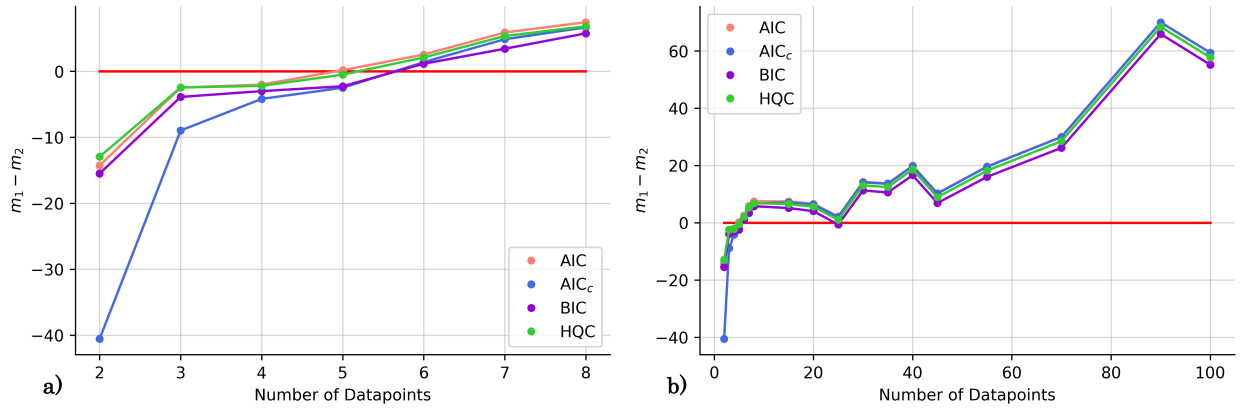

Figure 2: Plot of the difference between m_1 (the best model chosen from a subset that does not include the data-generating kinetic mode) and m_2 (the data-generating kinetic model) with respect to the number of data points available, which were used to estimate the parameters of all rival models. **a)**: shows a more detailed version of the graph in the low-data regime. **b)**: provides a more holistic perspective regarding the trends of the criteria in the high-data regime.

A.3 Summary

In this investigation, utilizing a simple isomerization case study, the behavior of several information criteria was dissected. Comprehensive analyses of how various factors, including inherent noise in the data set and the size of the data set, influence the performance of these criteria were provided. Every conclusion was explained and rooted in the corresponding mathematical form of these criteria. All findings seem to suggest a specific ranking for the information criteria, with the best criterion listed first:

- Akaike Information Criterion,
- Sample Corrected Akaike Information Criterion,
- Hannan-Quinn Criterion,
- Bayesian Information Criterion.

It is crucial, however, to acknowledge that these findings are context-specific and may not universally apply across different case studies. Furthermore, it is worth noting (and remembering) that the choice of a model selection criterion can transcend into a philosophical discussion as different information criteria bring diverse philosophical assumptions to their derivation, and all address slightly different questions. Thus, while the presented ranking can serve as a helpful guide – and it is in fact used in the methodological frameworks proposed – it should not be misconstrued as an absolute measure of the performance of the discussed information criteria across all disciplines.

B Model Discrimination

Model-based design of experiments (MBDoE) is critical in model discrimination, especially when the experimental budget has not yet been spent. For the methodological frameworks proposed herein, the Hunter-Reiner criterion has been adopted primarily because the evaluation of parameter uncertainty falls beyond the scope of this study, thus excluding any criteria necessitating the estimation of model response uncertainty at any given experimental point. Furthermore, the interpretability of Akaike weights design criterion is not of particular interest at this stage. The primary function of the Hunter-Reiner criterion is to identify the optimal experiment that maximizes the difference between the responses of two models.

In an ideal world, the primary goal of model discovery – the central objective of this research – is to pinpoint the optimal experiment that yields the largest discrepancy between a proposed model's and the data-generating model's response. Nevertheless, the actual underlying model is unknown, thus necessitating an approximation. This study examines two modeling approaches geared towards approximating the behavior of the data-generating kinetic model: Gaussian process state space model (GPSSM) and the second-best model generated from ADoK-S, hereinafter referred to as ADoK$_2$.

A GPSSM describes a nonlinear dynamical system, in which Gaussian processes are used to predict the state space dynamical transitions of a system (e.g., a reactive system) [1,2]. This model comprises of a non-parametric representation of the system's dynamics rooted in Bayesian principles, complemented by hyperparameters that control the behavior of this non-parametric representation [1]. GPSSMs have been chosen as one of the modeling approaches owing to their non-parametric nature, rendering them proficient in learning from small data sets (commonly found in kinetic studies), thereby outperforming parametric alternatives like recurrent neural networks [2]. Even though GPSSMs are generally favored due to their probabilistic attributes (i.e., their ability to account for prediction uncertainty), this feature is not of particular relevance in this context, given that the Hunter-Reiner criterion does not factor in uncertainty.

This self-contained study aims to comparatively evaluate the worst-case performance of the two models in approximating the data-generating model within a specified experimental space. The following procedure is employed for each of the three case studies used in this work (the hypothetical isomerization reaction, the decomposition of nitrous oxide, and the hydrodealkylation of toluene):

1. Generate data corresponding to a given case study following the methodology detailed in the respective section;
2. Normalize data to construct the training set for the GPSSM. The input training data set comprises of the concentrations of all observable species at $t \in [t_0, \cdots, t_{n-1}]$, while the output training data set consists of the concentrations of species *X* at $t \in [t_1, \cdots, t_n]$; one GP is dedicated to each species observable in the reaction system;
3. Train each GP using the compiled training data set and GPJax (a Python package which implements GPs using Jax) [3];
4. Execute one iteration of ADoK-S using the data sourced from the case studies and capture the second-best model generated, denoted as ADoK$_2$;
5. Implement the Hunter-Reiner criterion to identify the experiment that maximizes the discrepancy between the data-generating model and the trained models (GPSSM and ADoK$_2$). This elucidates which model more accurately represents the real system in the worst-case scenario.

Table 1 summarizes the sum of squared errors (SSE) between the response of the trained models and the data-generating model for the worst-case experiment. For the sake of completeness, we also compared a few naive parametric models with the trained ones for the hydrodelalkylation of toluene; Table 2 summarizes these results. Based on these outcomes, it appears appropriate to choose the second-best kinetic model suggested by ADoK for implementing MBDoE alongside the best model. This approach enables further experiment generation when the modeler is not satisfied with the choices output by either ADoK-S or ADoK-W.

Table 1: The worst-case scenario performance of the trained GPSSM and ADoK$_2$ with respect to the data-generating kinetic model for each case study.

Case Study	SSE for GPSSM (M^2)	SSE for ADoK$_2$ (M^2)
Hypothetical isomerization reaction	0.27	0.11
Decomposition of nitrous oxide	90.69	0.42
Hydrodealkylation of toluene	49.20	0.14

Table 2: The worst-case scenario performance of the trained GPSSM, $ADoK_2$ and a few naive parametric models with respect to the data-generating kinetic model for the hydrodealkylation of toluene.

Model	SSE (M^2)
GPSSM	49.20
$ADoK_2$	0.14
kC_T	104.712
kC_H	408.179
kC_T^2	156.259
kC_H^2	498.939
kC_TC_H	19.503
$k_1C_T + k_2C_H$	145.104
$k_1C_T^2 + k_2C_H^2 + k_3C_TC_H$	50.046

C Benchmark Study of Derivative Estimation Methods

In our pursuit of refining and validating the adaptability and flexibility of our GP-based rate estimation framework, we recognized the importance of a comprehensive evaluation against current state-of-the-art derivative estimation methods. To this end, we performed a benchmarking study to compare the performance of our GP-based approach with leading methodologies mentioned in relevant literature [4]. Details into each of the methodologies can be found in Van Breugel et al. [4].

This comparative analysis was designed to assess the effectiveness of each method in accurately estimating reaction rates within the context of chemical kinetics. Our findings indicate that our GP-based method, even in the absence of mathematical constraints, exhibits competitive performance. It achieved a squared error of 26.528 $(\text{M h}^{-1})^2$, which is marginally higher than the 25.951 $(\text{M h}^{-1})^2$ squared error associated with the Iterative Total Variation Regularization approach. The detailed results are presented in Table 3 These metrics were derived from an evaluation involving the calculation of squared errors between estimated and actual reaction rates across various chemical species in all seven experiments conducted as part of the hydrodealkylation of toluene case study presented in the main manuscript.

Table 3: Results from benchmarking state-of-the-art derivative-estimation methods against our GP-based approach.

Rate Estimation Method	SSE ($(\text{M h}^{-1})^2$)
Finite Difference: First Order	41.648
Finite Difference: Second Order	88.637
Finite Difference: Iterated First Order	37.697
Smooth Finite Difference: Median smoothing	130.883
Smooth Finite Difference: Mean smoothing	79.449
Smooth Finite Difference: Gaussian smoothing	30.174
Smooth Finite Difference: Friedrichs smoothing	32.586
Smooth Finite Difference: Butterworth smoothing	176.221
Smooth Finite Difference: Spline smoothing	41.711
Iterative Total Variation Regularization (regularized velocity)	25.951
Linear Models: Spectral derivative	142.249
Linear Models: Sliding polynomial fit	145.857
Linear Models: Savitzky-Golay filter	676.194
Kalman smoothing: constant velocity (forward-backward)	53.044
Kalman smoothing: constant acceleration (forward-backward)	70.167
Kalman smoothing: constant jerk (forward-backward)	70.167
GP-Based Approach	26.528

D ADoK-S Performance in Multi-Reaction Systems

To demonstrate the versatility of the ADoK-S methodology in the discovery of kinetic models in a multi-reaction system setting, we have applied ADoK-S on a specific part of a metabolic pathway, particularly the phospholipid cycle, more details can be found in Iba [5]. The portion of the network that is of importance in this case study, involves two key reactions facilitated by enzymes glycerol kinase and glycerol-1-phosphatase. ATP, which is referred to as "A", serves as the external input for the network. The reactions within this subnetwork result in the production and utilization of

sn-glycerol-3phosphate ("B") and glycerol ("C"). The dynamics of this metabolic subnetwork are described using the Michaelis-Menten law and can be approximated to:

$$\frac{dC_A}{dt} = \frac{-k_A C_A C_C}{1 + K_B C_A C_B} \tag{4a}$$

$$\frac{dC_B}{dt} = \frac{k_A C_A C_C}{1 + K_B C_A C_B} - k_C C_B \tag{4b}$$

$$\frac{dC_C}{dt} = \frac{-k_A C_A C_C}{1 + K_B C_A C_B} + k_C C_B \tag{4c}$$

The kinetic parameters were arbitrarily defined as: $k_A = 9$ M^{-1} h^{-1}, $K_B = 5$ M^{-2} and $K_C = 2$ h^{-1}. From Eq. (4), we generated five datasets by integrating the ODE system under five different initial conditions, collecting 50 datapoints per dataset and adding Gaussian noise (i.e.; zero mean and a standard deviation of 0.01 for A, B and C) to mimic the response from a real chemical system. Due to the complexity of the system, we increased the sampling frequency that we have used in the main manuscript. The initial conditions were: $(C_A(t=0), C_B(t=0), C_C(t=0)) \in (2.0, 1.0, 2.0), (2.0, 1.0, 0.2), (0.2, 1.0, 0.2), (2.0, 0.0, 2.0), (0.2, 0.0, 0.2)$ M. The initial conditions were randomly selected from a 2^k factorial design of experiments. Our findings show that ADoK-S an almost identical version of the data-generating kinetic model after five iterations (i.e., four extra experiments proposed by MBDoE). The extra MBDoE experiments were: $(C_A(t=0), C_B(t=0), C_C(t=0)) \in (1.101, 0.0, 0.298), (2.0, 0.177, 0.667), (2.0, 1.0, 0.2), (1.136, 0.0, 2.2)$ M. The generated model by ADoK-S is shown below:

$$\frac{dC_A}{dt} = \frac{-9.126 C_A C_C}{1 + 5.036 C_A C_B + 0.018 C_A} \tag{5a}$$

$$\frac{dC_B}{dt} = \frac{9.144 C_A C_C}{1 + 5.082 C_A C_B} - 2.015 C_B \tag{5b}$$

$$\frac{dC_C}{dt} = \frac{-9.150 C_A C_C}{1 + 5.084 C_A C_B} + 2.016 C_B \tag{5c}$$

We see that only the first rate equation is different from the data-generating model, and that is by a single parameter that is considerably less significant than the rest. Below, we show in Fig. 3 a summary of the results of this case study, similar to the ones we presented in the manuscript.

References

[1] Frigola R, Lindsten F, Schön TB, Rasmussen CE. Identification of Gaussian Process State-Space Models with Particle Stochastic Approximation EM. IFAC Proceedings Volumes. 2014;47(3):4097-102. 19th IFAC World Congress.

[2] Eleftheriadis S, Nicholson T, Deisenroth M, Hensman J. Identification of Gaussian Process State Space Models. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. p. 1-11. Available from: `https://proceedings.neurips.cc/paper_files/paper/2017/file/1006ff12c465532f8c574aeaa4461b16-Paper.pdf`.

[3] Pinder T, Dodd D. GPJax: A Gaussian Process Framework in JAX. J Open Source Softw. 2022;7(75):4455. Available from: `https://doi.org/10.21105/joss.04455`.

[4] Van Breugel FV, Kutz JN, Brunton BW. Numerical Differentiation of Noisy Data: A Unifying Multi-Objective Optimization Framework. IEEE Access. 2020;8:196865–196877. Available from: `http://dx.doi.org/10.1109/access.2020.3034077`.

[5] Iba H. Inference of differential equation models by genetic programming. Inf Sci. 2008;178(23):4453-68. Including Special Section: Genetic and Evolutionary Computing.

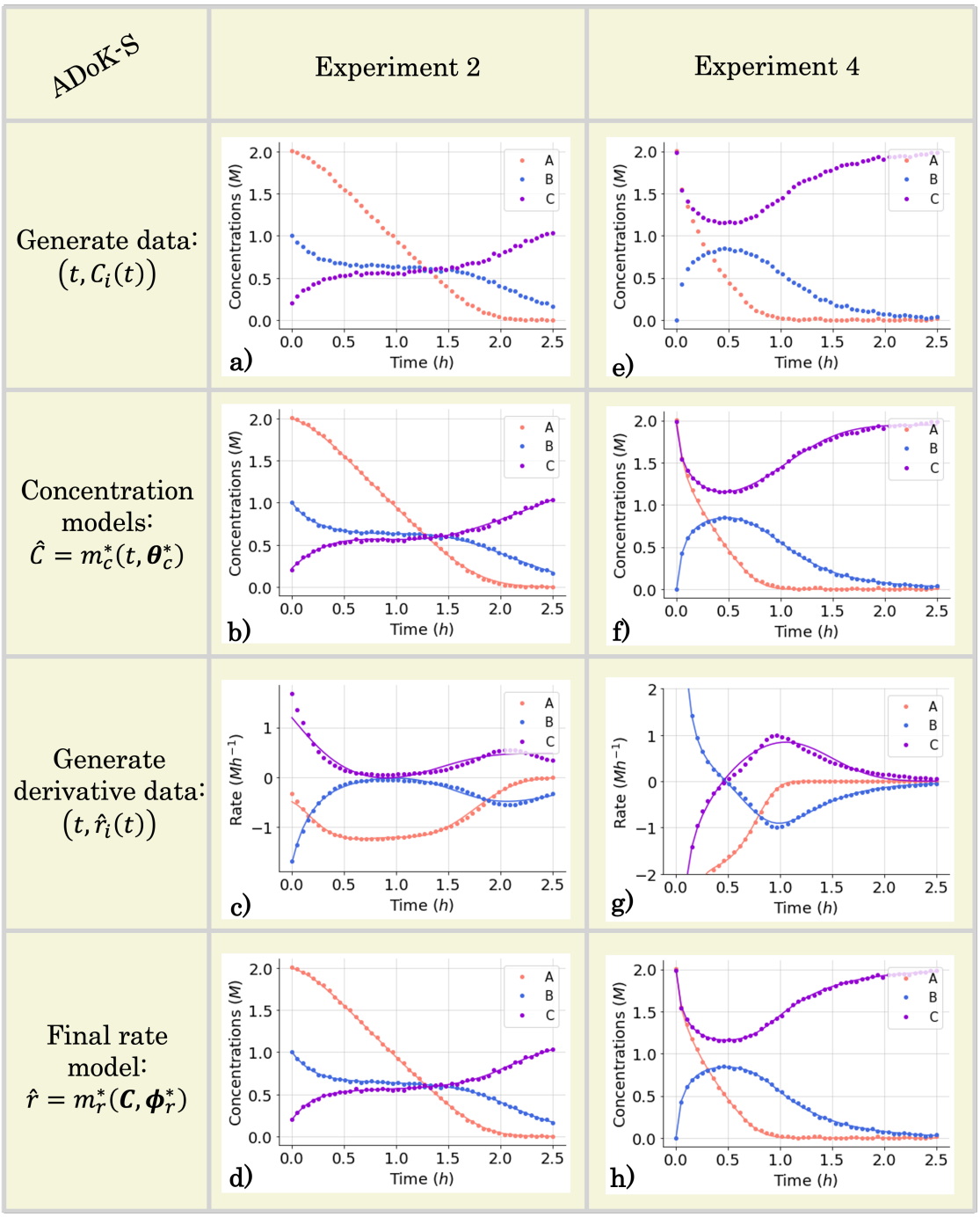

Figure 3: The conditions for the second and fourth computational experiment $(C_A(t=0), C_B(t=0), C_C(t=0)) \in (2.0, 1.0, 2.0), (2.0, 0.0, 2.0)$ M, respectively, where A, B and C denote ATP, sn-glycerol-3phosphate and glycerol, respectively. **a)** and **e)**: the measured concentration data for the second and fourth experiments which are used in the execution of ADoK-S. **b)** and **f)**: the concentration profiles selected by AIC that model the dynamic trajectories of the observable species' concentrations in the second and fourth experiments as a function of time. These models are used to approximate the rate measurements. **c)** and **g)**: numerical derivatives of the selected concentration profiles and the true rate measurements (which realistically are inaccessible). **d)** and **h)**: response of the selected rate model after the fifth iteration of the ADoK-S with the initial set of experiments and four additional MBDoE-proposed experiments for the second and fourth experiments.