

A Bayesian approach to understanding artificial enzymatic networks

Mathieu G. Baltussen Jeroen van de Wiel Cristina Lía Fernández Regueiro
Miglė Jakštaitė Wilhelm T.S. Huck

Summary/Abstract

In order to create artificial enzymatic networks capable of increasingly complex behavior, an improved methodology in understanding and controlling the kinetics of these networks is needed. Here, we introduce a Bayesian analysis method allowing for the accurate inference of enzyme kinetic parameters and determination of most likely reaction mechanisms, by combining data from different experiments and network topologies in a single probabilistic analysis framework. This Bayesian approach explicitly allows us to continuously improve our parameter estimates and behavior predictions by iteratively adding new data to our models, while automatically taking into account uncertainties introduced by the experimental setups or the chemical processes in general. We demonstrate the potential of this approach by characterizing systems of enzymes compartmentalized in beads inside flow reactors. The approach we introduce here could be relevant to the design of increasingly complex artificial enzymatic networks, making the design of such networks more efficient, and robust against the accumulation of experimental errors.

Introduction

Enzymatic reaction networks (ERN's) play key roles in many cellular processes, such as energy metabolism, signaling pathways, and cell division¹⁻³. The fields of synthetic biology and systems chemistry aim to understand and reproduce the behavior of these ERNs in artificial systems⁴⁻⁸. Previous work has shown the development of small network motifs⁹ by autocatalysis and delayed inhibition¹⁰, photochemical control of oscillations by reversible photo-inhibitors¹¹, coupling to DNA-based circuits¹², logic-gate responses¹³, pattern-formation¹⁴, adaptive responses to environmental perturbations¹⁵, and coupling to dynamic environments¹⁶. While these networks can show complex behavior, such as oscillations and adaptation, scaling up their size towards metabolic scales remains a significant challenge. To construct complex, yet functional ERNs, estimating the mechanisms and kinetics of the enzymatic reactions in these systems is essential in order to reliably predict the relevant experimental regimes in which a desired functional output will be observed¹⁷. But while the development of artificial ERNs with more complex behavior continues, methods are missing to not only obtain realistic kinetic parameter estimates, but to simultaneously allow for the evaluation of the relevance and correctness of existing kinetic models.

This lack of accurate and experimentally realistic parameter and mechanism estimation greatly limits the efficient exploration of more complex systems. Furthermore, while the fitting of a model to experimental data is in principle relatively simple, in practice numerous sources of uncertainty are encountered, including experimental errors and unknown inhibitory or allosteric effects. Typically, the kinetic parameters of an enzymatic reaction are estimated from a single dataset, using least-squares regression or similar maximum likelihood estimation methods. Although this approach is well-established, there are multiple downsides¹⁸. First, sources of uncertainty must be explicitly modelled in, which would require an exact knowledge of the influence of these uncertainties on the final experimental results^{19,20}. Secondly, this approach often neglects additional sources of data, either from previous or additional experiments, or from literature. And lastly, estimation of enzyme kinetics is often done using rather limited datasets, which should increase the uncertainty of the obtained parameter values, but in practice potentially leads to overfitting of the proposed model^{21,22}.

Here, we demonstrate an analysis framework based on Bayesian methods for the inference of kinetic parameters and comparison of reaction mechanisms of compartmentalized enzymes in a flow reactor. Bayesian methods are probabilistic in nature, so that any knowledge of kinetic parameters or reaction mechanisms obtained from experimental data is expressed in terms of probability-distributions, instead of specific values. Furthermore, they allow for the explicit incorporation of any information previously obtained on the system in question through the prior, either from literature or previous experiments, resulting in a coherent framework for combining data from different sources²³. Additionally, they are ideally suited for estimations that contain uncertainty and a lack of data²⁴. Bayesian methods have been more widely implemented in recent years, mainly due to an increase in available computational power and an increase in general availability of powerful, yet accessible, algorithms. They are used in a wide range of fields, from applications in pure physics²⁵, medicine²⁶, and sociology²⁷, to large-scale metabolomics in systems biology²⁸ and optimized peak detection in chromatographic methods²⁹. Previous research on the applications of Bayesian methods in enzymatic networks have mostly been attempted from a systems biology perspective, focusing on whole-cell metabolomics^{28,30,31}, or focusing on simulated datasets and evaluating the feasibility of an alternative enzyme rate equation³². The approach introduced here instead focusses on experimental relevance and the combination of multiple datasets, specifically for the construction of encapsulated enzymes in flow, but is readily adoptable in most experimental enzymatic reactions setups, without requiring extensive computational expertise to employ.

Methods

Bead production

Two methods were used to immobilize enzymes through polymerisation of the enzyme into polyacrylamide hydrogel beads. The first method consists of enzyme-first functionalization with a 6-acrylamino-hexanoic acid succinate linker (AAH-Suc) by coupling to amino groups of lysine using NHS chemistry, followed by the production of hydrogel beads using droplet-based microfluidics by UV-polymerising water-in-oil droplets.

containing the functionalized enzyme, acrylamide, N,N'-methylenebisacrylamide and a photoinitiator yielding monodisperse polyacrylamide-enzyme beads (PEBs). In the second method, droplet-based microfluidics technique are used to produce empty hydrogel beads consisting of acrylamide, N,N'-methylenebisacrylamide, acrylic acid and 2,2'-Azobis(2-methylpropionamidine) dihydrochloride. After UV polymerisation the carboxyl groups of acrylic acid were activated by EDC/NHS chemistry and later enzyme was coupled via amine groups of lysine residues forming PEBs. Details of the procedure used per type of PEB can be found in the Supporting Information.

Flow experiments

Flow experiments were conducted similar to description in previous work³³, but replacing the inflow of the desired enzymes with the desired volume of PEBs, which remained compartmentalized in the reactor during the experiment. The openings of the reactors were sealed with Whatman Nuclepore polycarbonate membranes (5 μm poresize) to prevent outflow of PEBs. Cetoni Low-Pressure High-Precision Syringe Pumps neMESYS 290N were used to control the dispense of the different solutions, prepared in Gastight Hamilton syringes (2500 – 10000 μL), into the CSTR. The precise flow profile of the desired flowrates was programmed using the Cetoni neMESYS software.

To detect and determine outflow concentrations from the CSTR both online and offline detection was employed. Online absorbance detection was achieved with an Avantes AvaSpec2048 Fiber Optic spectrometer and Avantes AvaLight 355 nm LED combined with a custom designed flow cuvette provided to us by LabM8. Alternatively, offline measurement could be achieved by means of connecting the outflow to a BioRad Model 2110 fraction collector. These fractions could subsequently be probed for NADH absorbance using a Tecan Spark M10 platereader, or probed for ATP, ADP, NAD+ and NADH using a Shidmadzu Nexera X3 HPLC.

Further details on the instrumentation and experimental protocols can be found in the Supporting Information.

Modelling of enzyme kinetics in flow

We generally assume that the enzymatic reactions behave according to Michaelis-Menten-like mechanisms, although other mechanisms might also be considered, and we add flow-dependent terms to model the dynamics of the flow reactor. Inclusion of these flow-terms yields the following system of Ordinary Differential Equations (ODEs) for a single-substrate single-product reaction:

$$\mathbf{f}(\mathbf{C}, \phi, \theta) : \begin{cases} \frac{d[P]}{dt} &= \frac{V_{max}[S]}{K_M + [S]} - k_f[P] \\ \frac{d[S]}{dt} &= -\frac{V_{max}[S]}{K_M + [S]} + k_f([S]_{in} - [S]) \end{cases}$$

Where $V_{max} = k_{cat}[E]$ and K_M are the kinetic parameters ϕ that generally need to be estimated, and where we introduce a set of control parameters θ , in the form of $[S]_{in}$, the effective substrate concentration flown into the reactor, and $k_f = \frac{\text{flowrate}(\frac{\mu\text{L}}{\text{min}})}{\text{reactor volume}(\mu\text{L})}$ the flow-constant. Measurements of the product-concentration are performed when the system has reached steady-state conditions ($\frac{d\mathbf{C}}{dt} = 0$), resulting in a set of steady-state concentrations $[P]_{ss}$ and $[S]_{ss}$. These values can then be used in a fitting procedure to obtain kinetic parameter estimates.

Creation of Bayesian models

In a Bayesian approach, the probability distributions for parameters of interest are obtained by application of Bayes' theorem

$$P(\phi|y) \propto P(\phi) P(y|\phi)$$

which relates the posterior probability $P(\phi|y)$ of a specific parameter value ϕ given the data y observed during an experiment, to the likelihood $P(y|\phi)$ of observing that specific data given the parameter value,

and any previously available information of the parameter, the prior $P(\phi)$. As the likelihood is a function of both the observed data and the values of the kinetic parameters, and not a probability distribution, we write $P(y|\phi) = \mathcal{L}(y, \phi)$.

In the case of a single-substrate, steady-state enzymatic network, the observed data y is given simply by the set of observed steady-state concentrations $[P]_{ss}$ at specific experimental conditions $[S]_{in}$ and k_f , while the parameter ϕ can be any of the kinetic parameters that is unknown, such as k_{cat} or K_M . Furthermore, because the data we collect is inherently noisy, we assume that the concentrations we observe are part of a normal distribution $[P]_{obs} \sim \mathcal{N}([P]_{ss}, \sigma)$ with a mean equal to the true steady-state concentration and an unknown standard-deviation σ . This assumption allows us to write down the form of the likelihood:

$$\mathcal{L}(y, \phi) = \mathcal{N}([P]_{ss} - [P]_{obs}, \sigma)$$

where $[P]_{ss} = g(\phi, \theta)$ is a function of the kinetic parameters and the experimental conditions. The likelihood incorporates most sources of uncertainty, such as the intrinsic fluctuations of product concentration at steady-state and noise from the used measurement technique, inside the uncertainty-term σ . This parameter is then inferred simultaneously with the kinetic parameters k_{cat} and K_M , allowing us to directly estimate the uncertainty in our observations as well. Any other sources of uncertainty, such as inconclusive data, or a wrong assumed reaction-mechanism, are implicitly encoded into the posterior probability distributions of the kinetic parameters.

We use the Python package PyMC3³⁴ and custom-written likelihood functions for observed steady-state concentrations, from which we can sample with the No-U-Turn Sampler (NUTS³⁵) by inclusion of likelihood gradients. The steady-states can be obtained either by symbolically solving $\mathbf{f}(\mathbf{C}, \phi, \theta) = 0$, or by numerical root-finding of the vector-function \mathbf{f} . The gradients for the numerical steady-states are obtained from the Implicit Function Theorem, which relates the sensitivity of the steady-state concentrations $\frac{\partial g}{\partial \phi}$ to the kinetic parameters without needing to explicitly write down an expression for the steady-state concentrations, via $\frac{\partial g}{\partial \phi}(\phi, \theta) = -\left[\frac{\partial \mathbf{f}}{\partial \mathbf{C}}(\mathbf{C}, \phi, \theta)\right]^{-1} \left[\frac{\partial \mathbf{f}}{\partial \phi}(\mathbf{C}, \phi, \theta)\right]$, while gradients for symbolic steady-states are automatically obtained via automatic differentiation in PyMC3. Using the NUTS-algorithm, we can obtain correlated probability distributions for the value of every kinetic parameter of interest. Implementations of the computational models used in this paper and the scripts to generate the figures can be found in the Supporting Information and on the associated Github page.

Results and Discussion

Obtaining improved accuracy from correlated parameter estimates

We first show the relevancy of our Bayesian approach by estimating the kinetic parameters of Trypsin PEBs cleaving a substrate (Cbz-Arg-7-amino-4-methylcoumarin, R-AMC) while in the presence of an inhibitor (Suc-Ala-Ala-Ala-7-amino-4-methylcoumarin, AAA-AMC), shown in figure 1. Two experiments were performed, one where the inhibitor was absent and one where the inhibitor was present (figure 1A). Both experiments on their own did not yield enough information to obtain conclusive estimates of all kinetic parameters involved (k_{cat} , K_M , K_I), as shown in figure 1B. Clearly, from the experiment without inhibitor relatively precise estimates can be obtained on k_{cat} and K_M , but no information is obtained on the value of the inhibition constant K_I . Thus, our posterior estimate of the inhibition constant is equivalent to our prior estimate (a uniform distribution between 1 and $10 \times 10^3 \mu M$). In contrast, from the experiment with inhibitor present, a posterior estimate for the inhibition constant can be obtained, albeit not a precise one. Additionally, from this experiment alone, the posterior estimates for the other kinetic parameters are also uncertain.

However, while the posterior estimates of the individual parameters remain uncertain, we do obtain additional information by analyzing the posterior correlations, shown in figure 1C. While the experiment without inhibitor does not show any correlation between the value of the estimated k_{cat} and K_I values, the experiment with inhibitor present shows a nonlinear correlation between low estimated values of k_{cat} and high values of K_I , and vice versa.

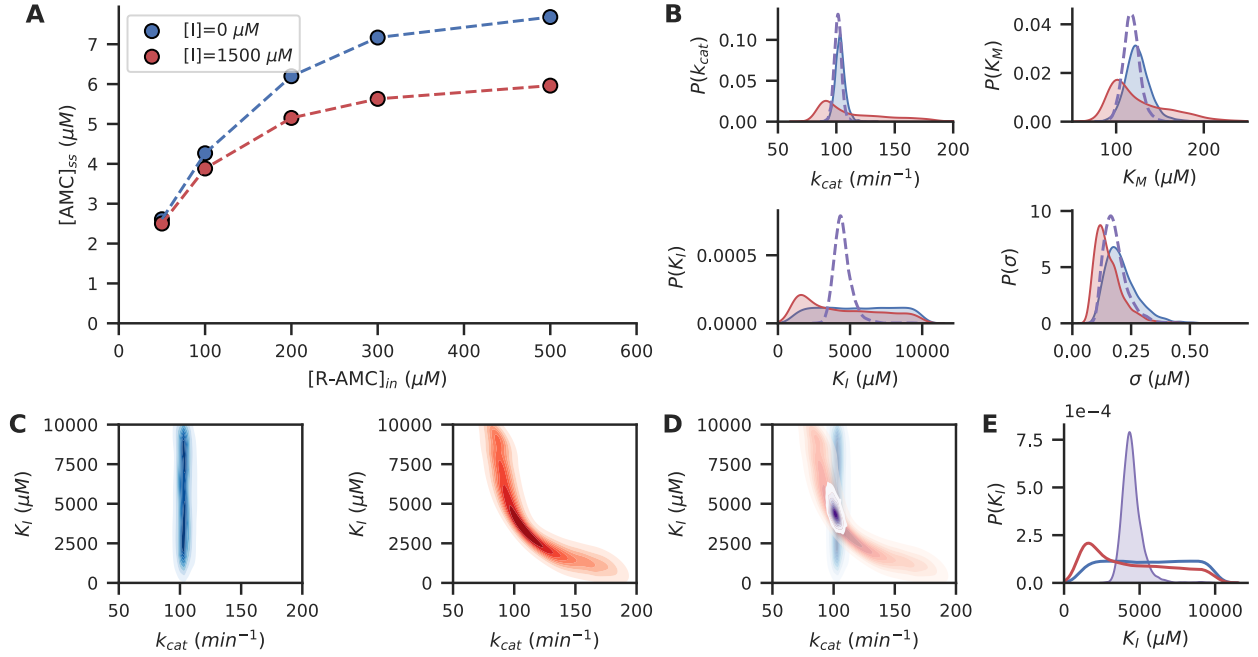


Figure 1: Improving parameter estimates by exploiting correlations **A**, Steady-state concentrations of R-AMC cleavage by Trypsin, with and without inhibitor (AAA-AMC). **B**, Posterior parameter estimates obtained from the data without inhibitor present (blue) and with inhibitor present (red). Combining both datasets in one model yields more precise posterior estimates (dashed, purple) **C**, Posterior correlation plots of k_{cat} and K_I from the data without inhibitor present (blue, left), showing no correlation, and with inhibitor present (red, right), showing high nonlinear correlation. **D**, Combining data from both experiments yields a new posterior distribution (purple) that exactly corresponds to the intersection from the two experiments separately. **E**, Comparison of posterior K_I estimates from the individual datasets (blue, red) to the estimate obtained from the combined dataset (purple).

Combining data from both experiments in a single likelihood function allows us to combine the certainty of the parameter estimates present in the first experiment with the highly correlated parameter estimates of the second experiment, to obtain a posterior distribution that is essentially an intersection of those obtained from the individual experiments (figure 1D). As expected, this allows us to obtain a much more precise estimate of the inhibitor constant, as shown in figure 1E. Moreover, this procedure yields improved estimates for every parameter in the system, not just the inhibition constant, which can be observed in figure 1B.

Consequently, the Bayesian approach greatly simplifies the iterative addition of experimental data to update parameter estimates. As shown here, subsequent measurements of enzyme activity in the presence of an inhibitor will not only allow an estimation of the inhibition constant, but also retro-actively improves the estimates for the Michaelis constant K_M and the turnover number k_{cat} .

Combining diverse experimental datasets

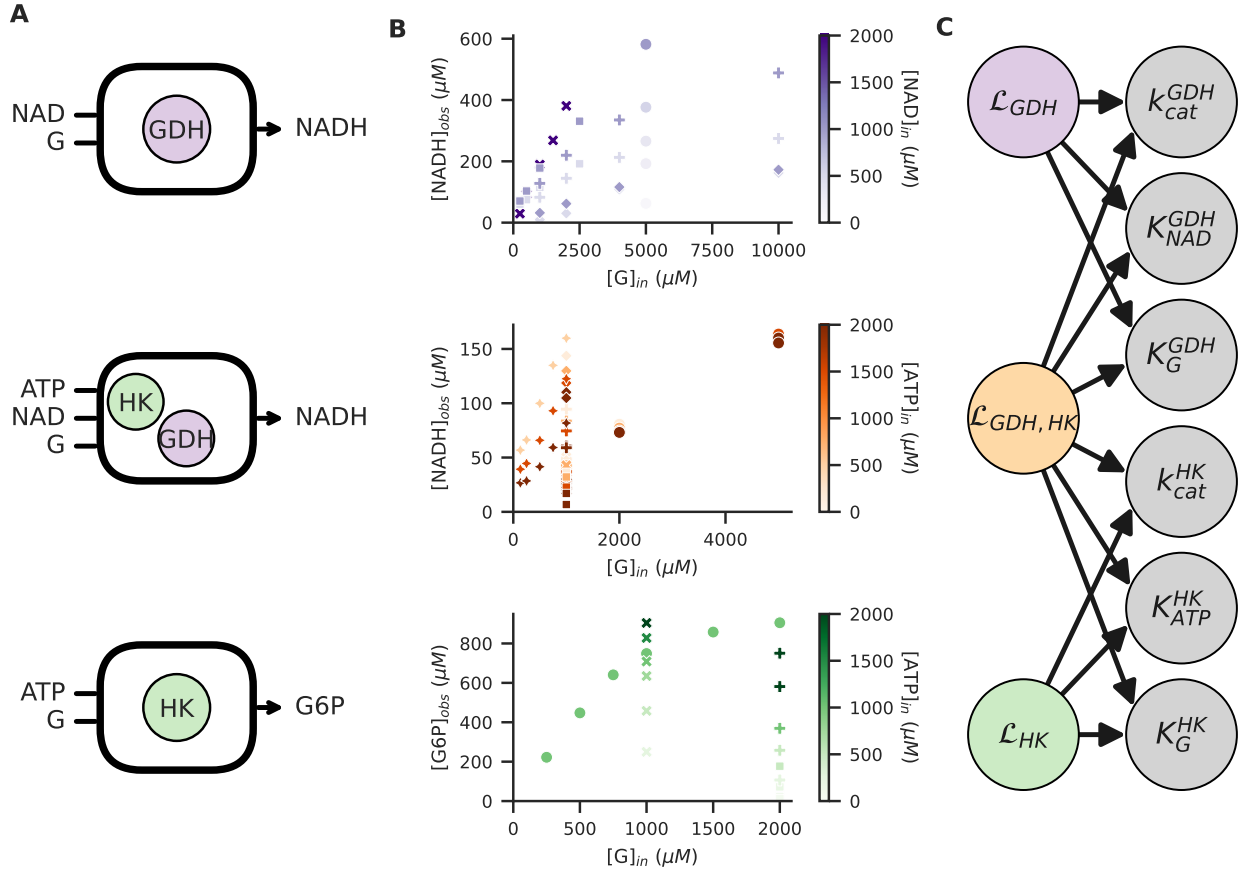


Figure 2: **Combining diverse experimental datasets** **A**, Three different ERN topologies are used in different reactors, and at different experimental conditions (varying input concentrations and volume of PEBs). **B**, Plots showing all collected observations at different input concentrations of glucose (x-axis) and co-factor (color intensity). The observed species is topology-dependent. **C**, Schematic of the causal network relating the observation likelihoods \mathcal{L}_x to the inferable parameters, where likelihoods corresponding to either the GDH or HK topology only relate to a subset of the parameters. The combined GDH,HK likelihood relates to every kinetic parameter in the probabilistic model. **D,E** Posterior parameter estimates obtained from the model combining all three (GDH, HK, GDH+HK) observation likelihoods. For the GDH k_{cat} , two estimates are obtained because PEBs with two different enzyme concentrations were used in different experiments.

More complex ERNs introduce a number of additional challenges in modelling the system's behavior. One of these challenges is combining data from a diverse range of experiments, both with variations in experimental

conditions, and variations in network topologies due to the enzymes that are present. Additionally, for some experiments only partial data can be obtained, for example in the case where only substrates involved in a single reaction can be observed, while substrates from a different reaction remain undetected.

In figure 2, we show how data obtained from these different types of experiments can be captured in a single probabilistic model. In figure 2A, we distinguish between 3 different network topologies, two with only a single type of enzyme PEB present, either glucose-dehydrogenase (GDH) or hexokinase (HK), and one where both enzymes PEBs are present simultaneously. For all three topologies, multiple experiments are performed at different conditions, such as different substrate input-concentrations and PEB volumes used. For the two single-enzyme topologies, detection of a single substrate is enough for full observability of the network (through stoichiometric conservation), while for the combined GDH+HK topology, only NADH is observed. Thus, the substrates involved in the hexokinase-reaction are not directly detected. These observations are shown in figure 2B.

All three topologies have a corresponding likelihood function that relates the observations to the kinetic parameters in question, as shown schematically in figure 2C (see the SI for the programmatic implementation of these likelihoods). While the GDH+HK system does not allow for full observability of the network, its likelihood does allow us to correlate the GDH and HK kinetic parameters, consequently leading to improved estimates of all parameters involved.

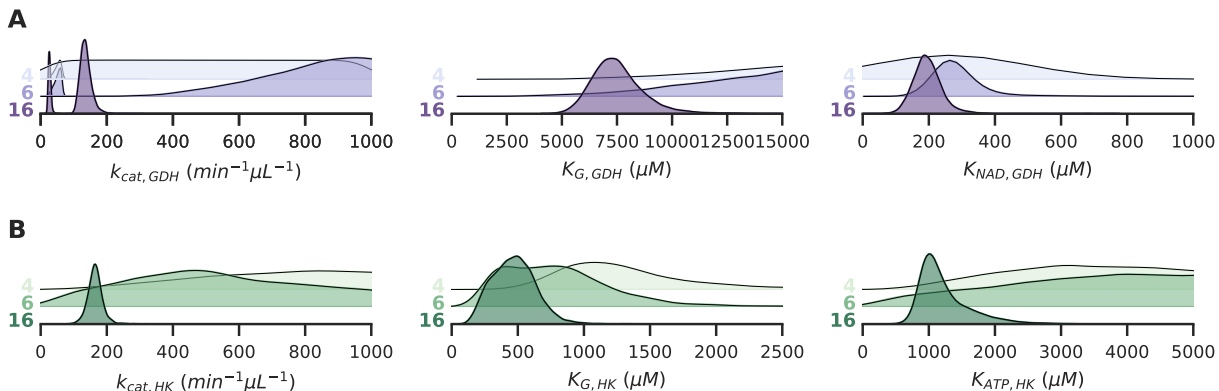


Figure 3: Iterative posterior updating **A,B** Posterior parameter estimates obtained from the model combining all three (GDH, HK, GDH+HK) observation likelihoods. For every parameter, the distributions are shown for 3 different dataset sizes, with respectively 4, 6, and 16 experiments included. Distributions are shifted and scaled to increase visibility. For the GDH k_{cat} , two estimates are obtained because PEBs with two different enzyme concentrations were used in different experiments.

The resulting posterior estimates of combining all available data are shown in figure 3. For the GDH PEB's, two different batches were used with different enzyme concentration, resulting in two distinct effective k_{cat} parameters. Estimation of their respective values are performed under the assumption that the K_M for the two substrates remain the same for both batches. By directly encoding this assumption into the combined likelihood functions, observations on both batches become relevant for estimation of all the parameters involved.

Correlating the parameter estimates of the individual enzymes through a joint likelihood function allows us to potentially improve parameter estimates by observing a system not directly related to those parameters. Thus, as more and more observations are made, any parameter estimates will increase in accuracy simply by the inclusion of more data. This iterative improvement of estimates as more data becomes available is shown by the gradual shrinkage of posterior distributions, implying that the estimates become more precise. This improvement is most pronounced when little data is available (for example from 4 to 6 experiments), but gradually becomes less for larger datasets (the same figure extended to stepwise addition of every experiment up to 16 can be found in the SI), and eventually converges to a final posterior distribution. For these final

posterior distributions, adding new data will not significantly alter the results, but the estimates will increase in robustness and become less susceptible to outliers in the data.

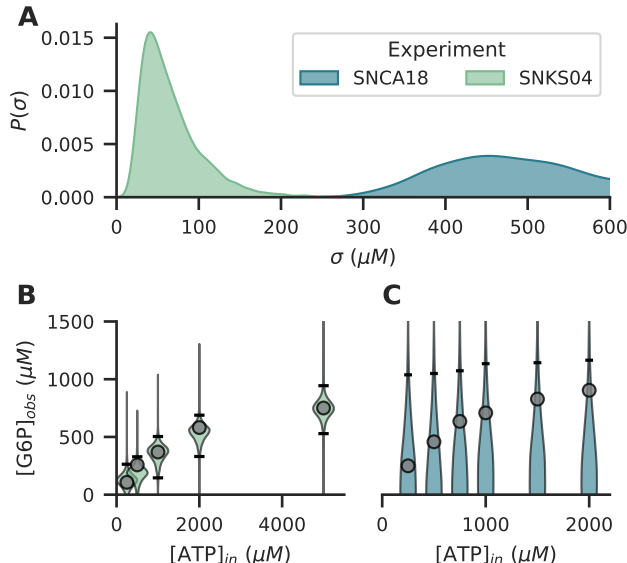


Figure 4: **Evaluating uncertainty estimates** **A**, Posterior experimental uncertainty estimates for two specific HK-experiments, obtained from the posterior distributions calculated from the full dataset of all experiments. One experiment (green, SNKS04) has a low estimated uncertainty, while the other experiment (blue, SNCA18) has a much higher estimated uncertainty. **B**, Associated observed datapoints of the low-uncertainty experiment, the posterior predictive distribution of expected observations, and 95% CI quantiles (black) **C**, Associated observation datapoints of the high-uncertainty experiment, the posterior predictive distributions of the expected observations, and 95% CI quantiles (black).

Additionally, by estimating the uncertainty in every experiment individually, it becomes more practical for a large number of experiments to determine which ones have corresponding results, and which ones are potential outliers or contain experimental errors. This can be observed especially in the uncertainty estimates for two specific HK experiments, as shown in figure 4. One experiment, with experiment code SNKS04, has a relatively low uncertainty estimates (figure 4A) and correspondingly, the posterior predictions obtained from the model are similar to the actual observations (figure 4B). However, another experiment stands out with a much higher uncertainty estimate ($\sigma \sim 400 - 600 \mu M$), which indicates some unknown error in the observations made during that experiment. Consequently, the posterior predictions show a very large spread and do not correlate well with the actual observations (figure 4C). Importantly, these uncertainty estimates are obtained by conditioning of the individual observations on the complete dataset of all experiments. Large uncertainty estimates therefore imply results that do not correspond with most other performed experiments.

By estimating the uncertainty parameters alongside all of the kinetic parameters, individual experiments are allowed to be ‘wrong’, and consequently influence the final parameter estimates less than other experiments. While not a solution for badly performed experiments, it does protect against drawing incorrect conclusions from incorrect data. Consequently, the uncertainty estimates indirectly act as an automatic weighting factor for individual experiments, where experiments with higher estimated uncertainty are less relevant towards the kinetic parameter estimations. It can also function as a key indicator for experiments influenced by an unknown source of error or systematic bias, especially in cases where large datasets collected over longer time periods are involved.

Comparing reaction mechanism hypotheses

The microscopic mechanisms underlying enzymatic reactions often allow for the creation of more complex kinetic models than simple Michaelis-Menten kinetics. However, a more complex model, with more kinetic

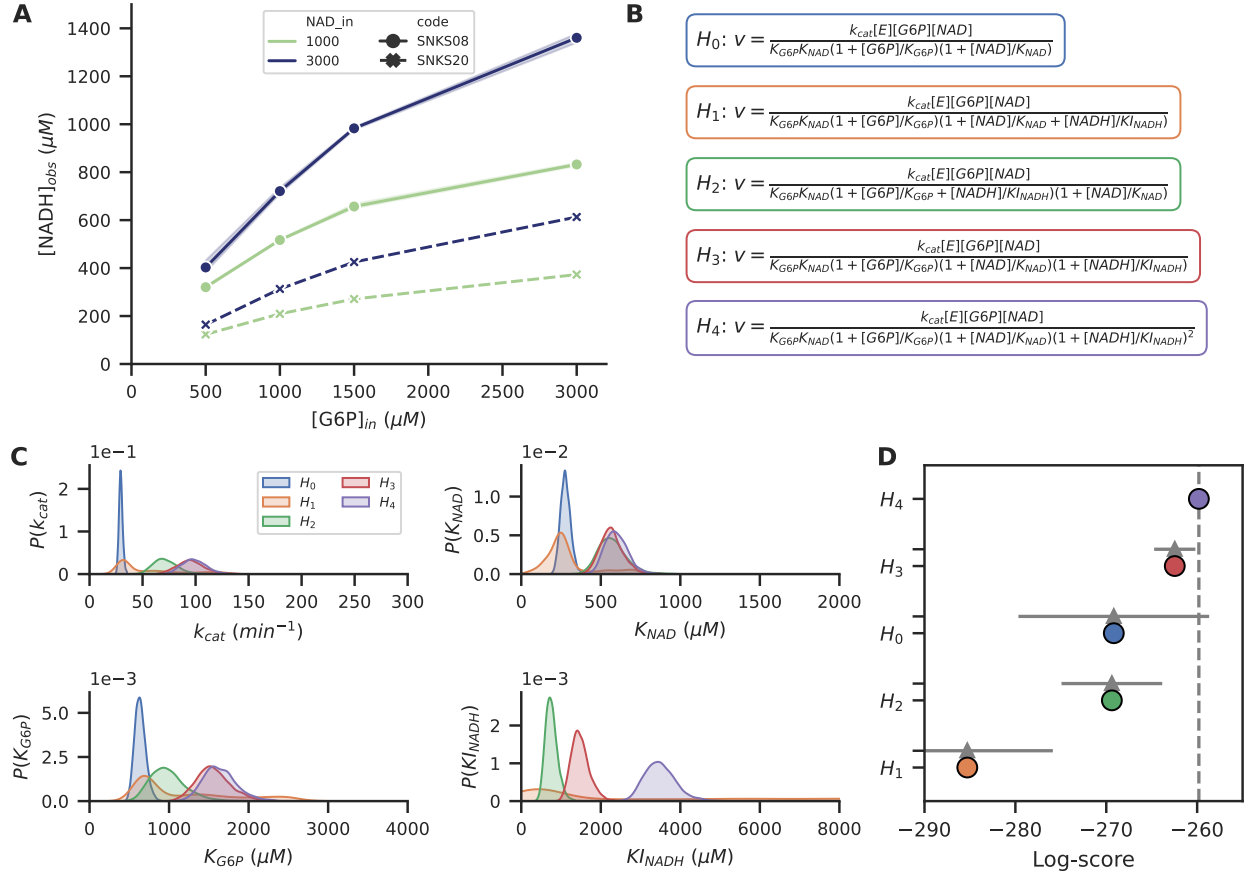


Figure 5: **Comparing reaction mechanism hypotheses** **A**, Steady-state concentrations obtained during two experiments, from a G6PDH system at different glucose-6-phosphate input concentrations and NAD input concentrations. **B**, Five different hypotheses for Michaelis-Menten mechanisms without (H_0), and with NADH product inhibition (H_1 - H_4). Only the reaction-rate is shown, but full sets of ODE's with additional flow-terms are used in the probabilistic model. **C**, Posterior parameter estimates for all five hypotheses. H_0 does not include an inhibition constant $K_{I_{NADH}}$, but all other hypotheses do. **D**, Comparison of the PSIS-LOO information criterion for all five hypotheses (colored), and standard errors of the difference in information criterion with respect to the top-ranked model (grey).

parameters, does not necessarily imply a more useful model. Instead, in the presence of uncertain data, it can lead to overfitting and unrealistically high certainty in the estimates.

In figure 5 we show how the posterior estimates obtained using our Bayesian approach, can be used to compare different hypotheses for the reaction mechanism and associated kinetics of glucose-6-phosphate dehydrogenase (G6PDH) PEBs. From a set of experiments performed at varying experimental conditions (figure 5A), we propose a number of different hypotheses describing the suspected mechanism of product inhibition by NADH on the reaction rate (figure 5B). We also include a 0-hypothesis describing a mechanism where the formation of NADH has no inhibiting effect, although the inclusion of a 0-hypothesis is not necessary for using this methodology.

We consider 4 modes of NADH inhibition: competitive inhibition of the NAD-binding site (H_1), competitive inhibition of the G6P-binding site (H_2), non-competitive inhibition of the enzyme activity (H_3), and cooperative non-competitive inhibition of the enzyme activity (H_4). All five hypotheses result in posterior distributions that give well-defined parameter estimates (figure 5C), from which it is difficult to conclude the most likely hypothesis. Instead, because the experimental noise per experiment is estimated alongside the kinetic parameters, the zero-hypothesis yields unrealistically precise estimates, due to the algorithm indicating that under the assumption that this hypothesis is true the experimental uncertainties are much larger. The four hypotheses that do model the influence product-inhibition have similar precisions in their parameter estimates.

To compare all hypotheses, and determine the most (and least) likely ones, we performed a Pareto Smoothed Importance Sampling Leave-one-out (PSIS-LOO) cross-validation directly from the posterior probability distributions³⁶. This test efficiently determines the model that approximates the observed data best and possesses the highest predictive accuracy, while taking into consideration the complexity of the models (e.g. the number of kinetic parameters involved) to prevent overfitting. From this hypothesis comparison we can conclude that given the experiments performed up to this point, a cooperative non-competitive inhibition (H_4) is the most likely product-inhibition mechanism occurring in the G6PDH PEBs, although the non-cooperative variant (H_5) cannot be ruled out. The zero-hypothesis cannot be ruled out completely, only requiring high experimental uncertainties to be explained by the data, although it remains unlikely. Both hypothesis for competitive inhibition can be ruled at with some confidence. Similarly, while the zero-hypothesis is likely not correct, a small chance exists that indeed one experiment does contain large experimental errors and is therefore unreliable.

Hypotheses comparisons can be a useful exploratory or diagnostic tool to evaluate which proposed reaction mechanisms are relevant and worth investigating further. However, numerous sources of error and uncertainty still exist in PSIS-LOO and similar cross-validation algorithms, making them susceptible to decreased robustness when comparing many models, or when little data is available³⁶. This should be taken into consideration when evaluating these result and considering new experiments to perform.

Conclusion & Outlook

We have demonstrated how a Bayesian approach towards analyzing enzymatic reaction networks allows for more accurate inference of the kinetics in these networks, while simultaneously taking into account any experimental or model-related uncertainties. Using this approach, we have shown how experimental data can be combined in one coherent framework, in order for us to correlate the findings in these experiments and improve the estimation of parameters, as well as outlier detection. This approach essentially allows us to continuously improve these estimates further by iteratively adding more experimental data to our models. Moreover, this means that any new experiment might have the potential to unlock more information from older experiments in the process, enabling much more efficient data gathering. Lastly, we have shown how this approach can be used to compare the likelihood of different reaction mechanism hypotheses. Comparing reaction mechanisms from a probabilistic perspective is a potentially powerful tool that can be used to make informed decisions about the next best experiments to perform when a lot of different mechanisms are under consideration. Importantly, it can equally well be used in re-analyzing old datasets in light of newly discovered or proposed mechanisms, or when new data becomes available. However, care should be taken in interpreting

the results from these comparisons as final conclusions. It is not a suitable method to make statements about the absolute truth of a hypothesis, as the test only check the predictive power of each hypothesis relative to all other hypotheses under consideration. Therefore, if no correct reaction mechanism is included in the hypotheses, then it will also not be considered in the test.

Bayesian methods open up multiple new areas of possibilities for the design of more complex enzymatic reaction networks, and for systems chemistry in general. In addition to the findings presented here, a lot of potential exists in the usage of knowledge from literature for more realistic prior distributions, which could improve the obtained estimates further, and could allow for direct comparison between new results and previous studies, as well as help in deciding initial experimental designs. Furthermore, more advanced hierarchical models and the inclusion of latent variables could potentially aid in discovering previously unknown interactions or hidden factors affecting the behavior of ERNs³⁷, both from a chemical point-of-view (allosteric effects, influence of pH) and an experimental point-of-view (systematic measurement errors, equipment deterioration). Finally, calculation of the full posterior probability distributions opens the door for determining optimal experimental designs^{38–40}. These designs could be aimed at a variety of different goals, such as experimental conditions for the maximum information gain for a certain kinetic parameter, but also the maximum production of a specific substrate or set of substrates, taking automatically into account any uncertainties that still exist about the behavior of these systems.

We do note that the methods introduced here are still computationally relatively expensive, and some of the sampling techniques are not yet suitable for every type of data. We have also limited our present study to enzymatic networks at steady-state. While inference on time-dynamic data is in principle possible, we found that current implementations are computationally demanding, while they do not yet significantly improve parameter estimates. This, however, is an area of active research that we hope can be improved in the future to allow for the inclusion of time-varying data and more complex models. Additionally, while our approach can indicate the presence of bad data and experimental errors, it does not guarantee the absence of sources of error. Care should still be taken to avoid a false sense of security when precise parameter estimates are obtained.

In conclusion, we have shown that the Bayesian approach we demonstrate here is highly relevant for the construction of complex enzymatic networks, allowing researchers to increase the predictability and reproducibility of artificial enzymatic networks, and allowing the field of enzymatic reaction networks to mature beyond toy models and proof-of-concepts.

Associated Content

Supporting Information

All supporting information, including data files, python scripts, and Jupyter notebooks are available at https://github.com/huckgroup/Bayesian-enzymatic-networks_manuscript_2022

- Supplementary materials and methods (PDF)
- Supplementary data (ZIP)
- Supplementary static versions of Jupyter notebooks (ZIP)

SI references

Droplet-production⁴¹

Bead production device⁴²

Linker synthesis⁴³

Author Information

Corresponding Author

- Wilhelm T.S. Huck

- Address: *Institute for Molecules and Materials, Radboud University Nijmegen, 6525 AJ, Nijmegen, the Netherlands;*
- Email: w.huck@science.ru.nl

Author Contributions

Notes

The author declares no competing financial interest.

Acknowledgments

We thank Max Derks (LabM8) for his help and work in designing the flow cuvette used for online absorbance detection. And we thank Arjan H. de Kleine for his help and work in the design and manufacturing of the CSTRs. Furthermore, we thank Bob van Sluijs for fruitful discussions and advice. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (ERC Adv. Grant Life-Inspired, grant agreement no. 833466) and the Dutch Ministry of Education, Culture and Science (Gravity program 024.001.035).

References

- (1) Barabási, A.-L.; Oltvai, Z. N. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics* **2004**, *5* (2), 101–113. <https://doi.org/10.1038/nrg1272>.
- (2) Kholodenko, B. N. Cell-signalling dynamics in time and space. *Nature Reviews Molecular Cell Biology* **2006**, *7* (3), 165–176. <https://doi.org/10.1038/nrm1838>.
- (3) Boccaletti, V. A. M.; Stefano And Latora. Complex networks: Structure and dynamics. *Physics Reports* **2006**, *424* (4-5), 175–308. <https://doi.org/10.1016/j.physrep.2005.10.009>.
- (4) Purnick, P. E. M.; Weiss, R. The second wave of synthetic biology: From modules to systems. *Nature Reviews Molecular Cell Biology* **2009**, *10* (6), 410–422. <https://doi.org/10.1038/nrm2698>.
- (5) Novák, B.; Tyson, J. J. Design principles of biochemical oscillators. *Nature Reviews Molecular Cell Biology* **2008**, *9* (12), 981–991. <https://doi.org/10.1038/nrm2530>.
- (6) Ludlow, R. F.; Otto, S. Systems chemistry. *Chem. Soc. Rev.* **2008**, *37* (1), 101–108. <https://doi.org/10.1039/B611921M>.
- (7) Roedel, H. W. H. van; Rosier, B. J. H. M.; Meijer, L. H. H.; Hilbers, P. A. J.; Markvoort, A. J.; Huck, W. T. S.; Greef, T. F. A. de. Programmable chemical reaction networks: emulating regulatory functions in living cells using a bottom-up approach. *Chemical Society Reviews* **2015**, *44* (21), 7465–7483. <https://doi.org/10.1039/C5CS00361J>.
- (8) Ashkenasy, G.; Hermans, T. M.; Otto, S.; Taylor, A. F. Systems chemistry. *Chemical Society Reviews* **2017**, *46* (9), 2543–2554. <https://doi.org/10.1039/C7CS00117G>.
- (9) Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; Alon, U. Network motifs: Simple building blocks of complex networks. *Science* **2002**, *298* (5594), 824–827. <https://doi.org/10.1126/science.298.5594.824>.
- (10) Semenov, S. N.; Wong, A. S. Y.; Van Der Made, R. M.; Postma, S. G. J.; Groen, J.; Van Roedel, H. W. H.; De Greef, T. F. A.; Huck, W. T. S. Rational design of functional and tunable oscillating enzymatic networks. *Nature Chemistry* **2015**, *7* (2), 160–165. <https://doi.org/10.1038/nchem.2142>.
- (11) Pogodaev, A. A.; Fernández Regueiro, C. L.; Jakšaitė, M.; Hollander, M. J.; Huck, W. T. S. Modular Design of Small Enzymatic Reaction Networks Based on Reversible and Cleavable Inhibitors. *Angewandte Chemie - International Edition* **2019**, *58* (41), 14539–14543. <https://doi.org/10.1002/anie.201907995>.
- (12) Meijer, L. H. H. H.; Joesaar, A.; Steur, E.; Engelen, W.; Santen, R. A. van; Merckx, M.; Greef, T. F. A. de. Hierarchical control of enzymatic actuators using DNA-based switchable memories. *Nature Communications* **2017**, *8* (1), 1117. <https://doi.org/10.1038/s41467-017-01127-w>.

- (13) Ikeda, M.; Tanida, T.; Yoshii, T.; Kurotani, K.; Onogi, S.; Urayama, K.; Hamachi, I. Installing logic-gate responses to a variety of biological substances in supramolecular hydrogel–enzyme hybrids. *Nature Chemistry* **2014**, *6* (6), 511–518. <https://doi.org/10.1038/nchem.1937>.
- (14) Zhang, Y.; Tsitkov, S.; Hess, H. Complex dynamics in a two-enzyme reaction network with substrate competition. *Nature Catalysis* **2018**, *1* (4), 276–281. <https://doi.org/10.1038/s41929-018-0053-1>.
- (15) Helwig, B.; Sluijs, B. van; Pogodaev, A. A.; Postma, S. G. J.; Huck, W. T. S. Bottom-Up Construction of an Adaptive Enzymatic Reaction Network. *Angewandte Chemie International Edition* **2018**, *57* (43), 14065–14069. <https://doi.org/10.1002/anie.201806944>.
- (16) Maguire, O. R.; Wong, A. S. Y.; Baltussen, M. G.; Duppen, P.; Pogodaev, A. A.; Huck, W. T. S. Dynamic Environments as a Tool to Preserve Desired Output in a Chemical Reaction Network. *Chemistry – A European Journal* **2020**, *26* (7), 1676–1682. <https://doi.org/10.1002/chem.201904725>.
- (17) Wong, A. S. Y. Y.; Huck, W. T. S. S. Grip on complexity in chemical reaction networks. *Beilstein Journal of Organic Chemistry* **2017**, *13*, 1486–1497. <https://doi.org/10.3762/bjoc.13.147>.
- (18) Efron, B.; Hastie, T. *Computer Age Statistical Inference*; Cambridge University Pr., 2016; p 496.
- (19) Gábor, A.; Villaverde, A. F.; Banga, J. R. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Systems Biology* **2017**, *11* (1), 54. <https://doi.org/10.1186/s12918-017-0428-y>.
- (20) Lillacci, G.; Khammash, M. Parameter Estimation and Model Selection in Computational Biology. *PLoS Comput Biol* **2010**, *6* (3), 1000696. <https://doi.org/10.1371/journal.pcbi.1000696>.
- (21) Brown, K. S.; Sethna, J. P. Statistical Mechanical Approaches to Models with Many Poorly Known Parameters. *Physical Review E* **2003**, *68* (2), 021904. <https://doi.org/10.1103/PhysRevE.68.021904>.
- (22) Transtrum, M. K.; Machta, B. B.; Sethna, J. P. Why Are Nonlinear Fits to Data so Challenging? *Physical Review Letters* **2010**, *104* (6), 060201. <https://doi.org/10.1103/PhysRevLett.104.060201>.
- (23) Schoot, R. van de; Depaoli, S.; King, R.; Kramer, B.; Märten, K.; Tadesse, M. G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; Yau, C. Bayesian statistics and modelling. *Nature Reviews Methods Primers* **2021**, *1* (1), 1. <https://doi.org/10.1038/s43586-020-00001-2>.
- (24) McNeish, D. On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal* **2016**, *23* (5), 750–773. <https://doi.org/10.1080/10705511.2016.1186549>.
- (25) Toussaint, U. von. Bayesian inference in physics. *Reviews of Modern Physics* **2011**, *83* (3), 943–999. <https://doi.org/10.1103/RevModPhys.83.943>.
- (26) Ashby, D. Bayesian statistics in medicine: A 25 year review, 2006, *25*, 3589–3631. <https://doi.org/10.1002/sim.2672>.
- (27) Lynch, S. M.; Bartlett, B. Bayesian Statistics in Sociology: Past, Present, and Future. *Annual Review of Sociology* **2019**, *45* (1), 47–68. <https://doi.org/10.1146/annurev-soc-073018-022457>.
- (28) St. John, P. C.; Strutz, J.; Broadbelt, L. J.; Tyo, K. E. J.; Bomble, Y. J. Bayesian inference of metabolic kinetics from genome-scale multiomics data. *PLoS Computational Biology* **2019**, *15* (11), e1007424. <https://doi.org/10.1371/journal.pcbi.1007424>.
- (29) Vivó-Truyols, G. Bayesian Approach for Peak Detection in Two-Dimensional Chromatography. *Analytical Chemistry* **2012**, *84* (6), 2622–2630. <https://doi.org/10.1021/ac202124t>.
- (30) Liepe, J.; Kirk, P.; Filippi, S.; Toni, T.; Barnes, C. P.; Stumpf, M. P. H. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols* **2014**, *9* (2), 439–456. <https://doi.org/10.1038/nprot.2014.025>.
- (31) Jayawardhana, B.; Kell, D. B.; Rattray, M. Bayesian inference of the sites of perturbations in metabolic pathways via Markov chain Monte Carlo. *Bioinformatics* **2008**, *24* (9), 1191–1197. <https://doi.org/10.1093/bioinformatics/btn103>.

- (32) Choi, B.; Rempala, G. A.; Kim, J. K. Beyond the Michaelis-Menten equation: Accurate and efficient estimation of enzyme kinetic parameters. *Scientific Reports* **2017**, *7* (1), 17018. <https://doi.org/10.1038/s41598-017-17072-z>.
- (33) Teders, M.; Pogodaev, A. A.; Bojanov, G.; Huck, W. T. S. Reversible Photoswitchable Inhibitors Generate Ultrasensitivity in Out-of-Equilibrium Enzymatic Reactions. *Journal of the American Chemical Society* **2021**, *143* (15), 5709–5716. <https://doi.org/10.1021/jacs.0c12956>.
- (34) Salvatier, J.; Wiecki, T. V.; Fonnesbeck, C. Probabilistic Programming in Python Using PyMC3. *PeerJ Computer Science* **2016**, *2*, e55. <https://doi.org/10.7717/peerj-cs.55>.
- (35) Hoffman, M. D.; Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv:1111.4246 [cs, stat]* **2011**.
- (36) Vehtari, A.; Gelman, A.; Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **2017**, *27* (5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>.
- (37) Engelhardt, B.; Kschischo, M.; Fröhlich, H. A Bayesian approach to estimating hidden variables as well as missing and wrong molecular interactions in ordinary differential equation-based mathematical models. *Journal of The Royal Society Interface* **2017**, *14* (131), 20170332. <https://doi.org/10.1098/rsif.2017.0332>.
- (38) Kreutz, C.; Timmer, J. Systems Biology: Experimental Design. *The FEBS Journal* **2009**, *276* (4), 923–942. <https://doi.org/10.1111/j.1742-4658.2008.06843.x>.
- (39) Letham, B.; Letham, P. A.; Rudin, C.; Browne, E. P. Prediction Uncertainty and Optimal Experimental Design for Learning Dynamical Systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2016**, *26* (6), 063110. <https://doi.org/10.1063/1.4953795>.
- (40) Huang, Y.; Gilmour, S. G.; Mylona, K.; Goos, P. Optimal Design of Experiments for Hybrid Nonlinear Models, with Applications to Extended Michaelis-Menten Kinetics. **2020**. <https://doi.org/10.1007/s13253-020-00405-3>.
- (41) Matula, K.; Rivello, F.; Huck, W. T. S. Single-Cell Analysis Using Droplet Microfluidics. *Advanced Biosystems* **2020**, *4* (1), 1900188. <https://doi.org/10.1002/adbi.201900188>.
- (42) Klein, A. M.; Mazutis, L.; Akartuna, I.; Tallapragada, N.; Veres, A.; Li, V.; Peshkin, L.; Weitz, D. A.; Kirschner, M. W. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **2015**, *161* (5), 1187–1201. <https://doi.org/10.1016/j.cell.2015.04.044>.
- (43) Semenov, S. N.; Markvoort, A. J.; Gevers, W. B. L.; Piruska, A.; Greef, T. F. A. de; Huck, W. T. S. Ultrasensitivity by Molecular Titration in Spatially Propagating Enzymatic Reactions. *Biophysical Journal* **2013**, *105* (4), 1057–1066. <https://doi.org/10.1016/j.bpj.2013.07.002>.