

Physics-Informed Automated Discovery of Kinetic Models

Miguel Ángel de Carvalho Servia^a, Ilya Orson Sandoval^a, King Kuok (Mimi) Hii^b, Klaus Hellgardt^a, Dongda Zhang^c, Ehecatl Antonio del Rio Chanona^{a*}

^a Imperial College London, Department of Chemical Engineering, London, United Kingdom

^b Imperial College London, Department of Chemistry, London, United Kingdom

^c The University of Manchester, Department of Chemical Engineering, Manchester, United Kingdom

* Corresponding Author: a.del-rio-chanona@imperial.ac.uk.

ABSTRACT

The industrialization of catalytic processes requires reliable kinetic models for design, optimization, and control. While white box models are preferred for their interpretability, they demand considerable time and expertise for their construction. This research enhances the ADoK-S framework by embedding prior expert knowledge using mathematical constraints and integrating uncertainty quantification. The improved methodology consists of: (I) a genetic programming algorithm with constraints to produce physically coherent candidate models, (II) a sequential optimization algorithm for parameter estimation, (III) model selection based on the Akaike information criterion (AIC), and (IV) uncertainty quantification of the chosen model's predictions. The refined approach not only requires less data for discovering kinetic models but also ensures physically sound proposals. With the inclusion of uncertainty quantification, the method bolsters prediction reliability, and aids in safer system developments – crucial for decision-making and risk management. These improvements enhance data efficiency, model reliability, and position automated knowledge discovery as a real alternative to traditional kinetic modeling techniques.

Keywords: chemical reaction engineering, kinetic model generation, automated knowledge discovery, expert knowledge, uncertainty quantification.

INTRODUCTION

Catalytic processes stand at the forefront of industrial advancement, and their significance is amplified in the face of climate change and the pressing need to reduce waste and enhance efficiency. In the journey from design to industrialization of catalytic processes, the role of reliable kinetic models is paramount, particularly in the development, optimization, and control of chemical reactors. These models act as the centerpiece of process understanding, enabling the refinement, and scaling of catalytic operations to meet environmental and economic goals.

Yet, the task of developing reliable kinetic models is fraught with challenges. The traditional approach has seen experts making closed-form, interpretable models, known as white box models, from the ground up, guided by physical principles and their domain-specific expertise. Despite their obvious appeal, these models are notoriously cumbersome and time-consuming to develop. As an alternative, the rise of black box models – such as

neural networks and Gaussian processes – and hybrid models, which aim to integrate the interpretability of white box models with the flexibility of black box models, has been observed. Nevertheless, these models come with their own caveats, most notably the high volume of data needed for training, which is typically not available in kinetic studies. Moreover, their lack of interpretability raises safety and reliability concerns for deployment in safety-critical processes.

In response to these challenges, a new paradigm has been embraced within the community: automated knowledge discovery or symbolic regression. Symbolic regression, an emerging tool in the field of machine learning and data analysis, offers a novel approach to model discovery. Unlike traditional regression methods that fit predetermined models to data, symbolic regression autonomously searches for mathematical expressions that best describe observed datasets. This method utilizes evolutionary algorithms or other heuristic techniques to explore an extensive space of mathematical constructs, ranging from simple algebraic equations to complex

interactions embodying non-linear behaviors. Its ability to uncover underlying patterns and relationships without pre-specified model structures makes it particularly valuable in scientific research, where it can reveal previously unrecognized phenomena or propose novel theoretical models. As a result, symbolic regression is gaining traction across diverse scientific domains including kinetic modeling, providing a powerful tool for exploratory analysis and advancing our understanding of complex systems.

Techniques such as SINDy (Sparse Identification of Nonlinear Dynamics) and ALAMO (Automated Learning of Algebraic Models for Optimization) have emerged to solve symbolic regression, yet not without limitations. They often depend on significant, accurate prior assumptions about the model structure, are vulnerable to noisy data, and lack a justified model selection routine. These gaps in methodology led the way for the development of ADoK-S and ADoK-W (Automated Discovery of Kinetics using a Strong/Weak formulation of symbolic regression) frameworks, as introduced in the work of de Carvalho Servia et al. [1]. Despite the benefits of this approach, these frameworks were recognized to require enhancements, specifically an uncertainty quantification mechanism for model's predictions and the integration of expert knowledge via mathematical constraints to enrich and ease the model discovery process.

This paper introduces these novel enhancements to the ADoK-S framework, incorporating a sampling-based uncertainty quantification algorithm to evaluate the reliability of a chosen model's predictions – a critical component for the safe design, optimization, and control of processes. Additionally, by assimilating expert knowledge through mathematical constraints, the data requirements for the model discovery phase are reduced. The subsequent sections of this paper are organized as follows: Section 2 revisits the ADoK-S framework, and presents the new methods for uncertainty quantification and constraint handling; Section 3 discusses the case study employed to evaluate the enhanced algorithm, now termed physics-informed ADoK-S (PI-ADoK-S); Section 4 presents the results where PI-ADoK-S improves upon the performance of the unconstrained algorithm; and Section 5 concludes with the study's main contributions.

METHODOLOGY

In prior works, ADoK-S was presented as a solution to the challenges of conventional modeling paradigms and positioned itself as an alternative to other automated knowledge discovery methodologies such as SINDy and ALAMO [1]. The ADoK-S framework is structured around three main components: (I) a genetic programming (GP) algorithm, which generates candidate models; (II) a sequential optimization algorithm designed for the

estimation of parameters in promising models; and (III) a transparent and reasoned model selection routine that utilizes the Akaike Information Criterion (AIC) for determining the best-suited model (for a detailed discussion on the choice of AIC as the criterion of choice, readers are referred to de Carvalho Servia and del Rio Chanona [2]).

ADoK-S implements symbolic regression in its traditional form, often referred to as the strong formulation. This method necessitates the development of rate models to stem from rate measurements, which, being experimentally inaccessible, must be estimated. Following the outlined three-step process, ADoK-S first uncovers the optimal concentration profiles (i.e., a time dependent function), describing the dynamic evolution of species' concentrations in the system. These profiles are then differentiated to approximate the rate measurements of the reaction system. Subsequently, utilizing the same three steps, ADoK-S aims to discover kinetic rate models that best fit these estimated rates. The selected rate model is integrated and compared against the original concentration data. Should the model output prove unsatisfactory – owing to either contradictions of prior knowledge or poor model fitting – ADoK-S offers a closed-loop approach. The modeler can run an optimal experiment, as determined by a Model-Based Design of Experiments (MBDoe) scheme, which is then concatenated with the initial dataset. The Hunter-Reiner criterion is used to find the optimal experiment that maximizes the error between the prediction of two competing models (in this study, the competing models are the best two models found using our methodology) [3]. With the new experimental data, ADoK-S can be iterated, and the subsequent model output examined. This iterative process can continue as required by the modeler or until the experimental budget is exhausted. For a detailed explanation of the ADoK-S algorithm, readers are referred to de Carvalho Servia et al. [1].

Despite favorable results, ADoK-S had its own limitations. Particularly, its lack of a direct and flexible mechanism for injecting prior knowledge through mathematical constraints during the model generation phase, and an absence of a quantifiable measure of the uncertainty in the predictions of the selected rate model. The following subsections delve into these newly incorporated features that have led to the development of PI-ADoK, an improved version of the ADoK-S framework. Figure 1 provides a detailed and visual representation of the workflow of PI-ADoK.

Inclusion of Mathematical Constraints

The effectiveness of integrating mathematical constraints within a symbolic regression framework has been a subject of debate in the literature. On one hand, studies such as Kronberger et al. [4] indicate that constraints

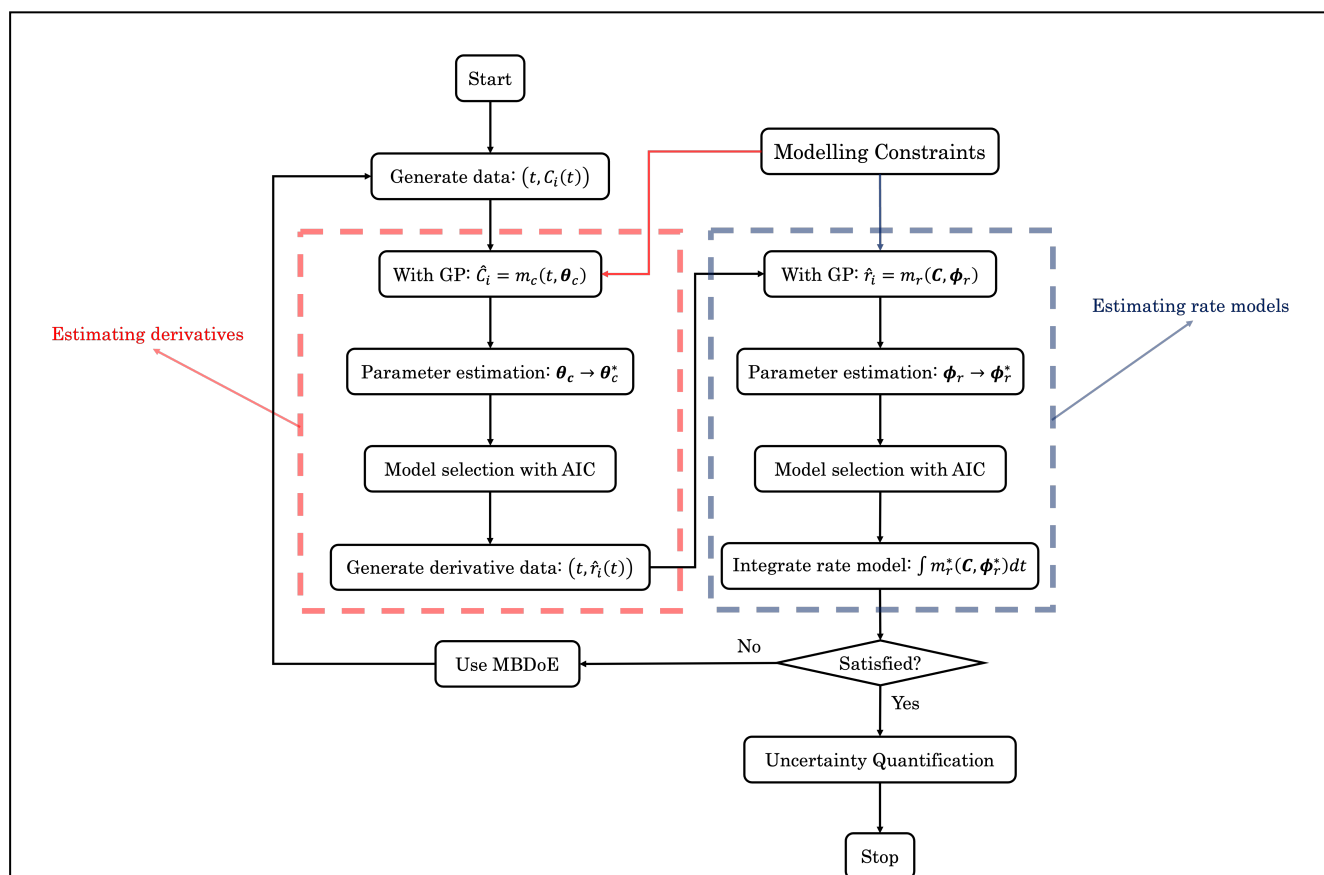


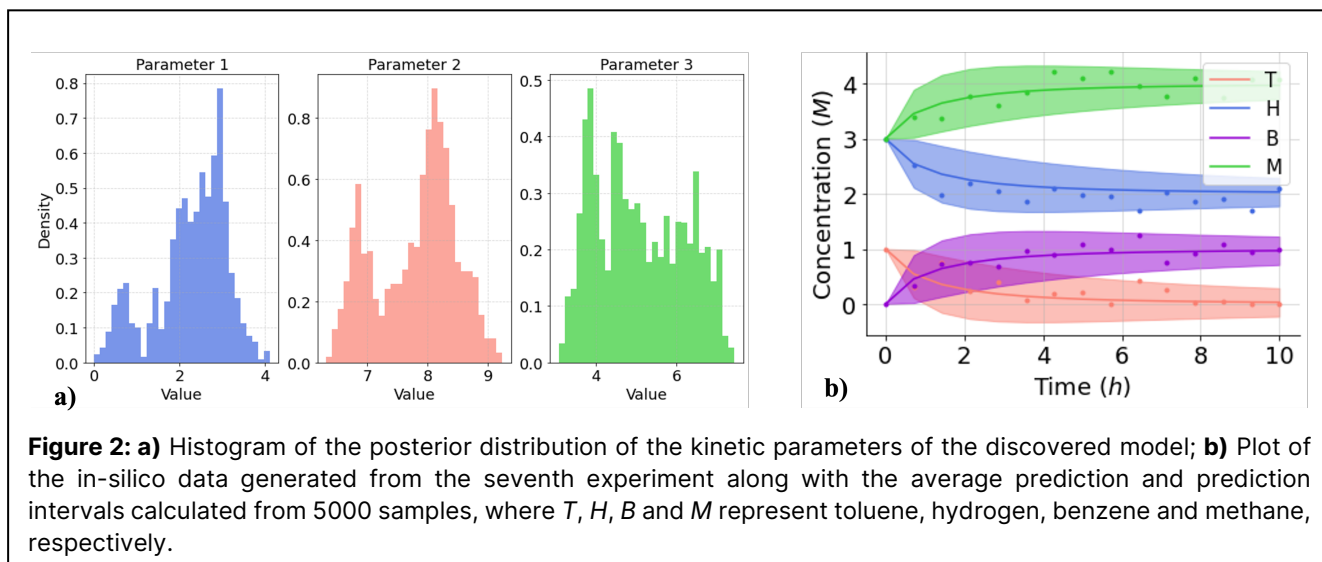
Figure 1: Illustration of the PI-ADoK framework, emphasizing its two main tasks: derivative estimation (red box) and rate model generation (blue box). The derivative estimation phase begins with genetic programming to generate candidate concentration models, followed by parameter estimation and model selection based on AIC. These selected models are then numerically differentiated to approximate reaction rates. In the rate-modeling phase, the estimated rates are used to construct kinetic expressions, which are further refined through parameter estimation and model selection. If the current model is insufficient, Model-Based Design of Experiments (MBDoE) suggests new experiments to improve data quality, iterating until a reliable model is obtained. Finally, uncertainty quantification is performed to assess predictive reliability of the best model. Constraints are applied at each step to guide the genetic programming algorithm toward physically meaningful models.

may lead to higher prediction errors on both training and testing datasets. The authors suggest that this could be attributed to slower convergence or a more rapid loss of diversity. However, they also note that the benefits of including constraints may become more relevant in scenarios with higher noise levels. Following up on this, the same authors published Haider et al. [5], investigating case studies with higher noise. Their findings concluded that constraints could aid in identifying better models, characterized by lower prediction errors, though the differences were not statistically significant compared to cases without constraints. On the other hand, other studies like Błażdek and Krawiec [6] have demonstrated that in smaller datasets, the inclusion of mathematical constraints can lead to statistically significant improvements compared to a GP algorithm without constraints.

Given the ambiguity in the literature, the necessity for a flexible method to incorporate prior knowledge

(which can be extensive in some kinetic studies), and the intuition that including constraints should improve the physical soundness of proposed models, PI-ADoK was developed and benchmarked against the original ADok-S framework. The integration of mathematical constraints within the GP algorithm is nuanced by a delicate balance of exploration versus exploitation, and the preservation of diversity. GP typically operates in vast search spaces, and constraints can reduce this space and facilitate the exploitation aspect of the argument. However, overly stringent constraints risk premature convergence to suboptimal solutions. Furthermore, maintaining diversity in the population while respecting constraints is a challenging task, that often also leads to early convergence to suboptimal solutions.

With these considerations in mind, our approach to integrating constraints is intentionally straightforward. In PI-ADoK, constraint violation is treated as a penalty in the



calculation of a model's performance metric. Specifically, the GP algorithm operates similarly to a standard GP algorithm, with a crucial modification: each constraint acts as a test for candidate models. If a model satisfies the constraints, its performance metric is simply its prediction error. In contrast, if a model fails to meet a constraint, its performance metric is penalized based on the degree of constraint violation and a predefined hyperparameter. This inclusion of hyperparameters allows the modeler to fine-tune the algorithm, ensuring the search space is neither overly constrained, which could yield poor models, nor under-constrained, which could yield physically implausible models.

Uncertainty Quantification

In uncertainty quantification, various methods range from simpler approximations like Laplace approximations and sigma points to more complex sampling algorithms like Metropolis-Hastings (MH) and Hamiltonian Monte Carlo (HMC). The choice depends on balancing computational efficiency with accuracy. For our purposes, where accuracy is prioritized, we have chosen the MH algorithm for its suitability in propagating uncertainty in kinetic models. MH handles complex, non-linear distributions well, does not require gradient information, and offers flexibility in proposal distributions, making it adaptable for complex models and high-dimensional spaces. While computationally intensive and requiring careful tuning, MH's adaptability and convergence make it a robust choice for uncertainty quantification in complex systems.

The MH algorithm is a Markov Chain Monte Carlo method designed to sample from a complex target distribution, such as a posterior distribution in Bayesian inference. Starting from an initial state, the algorithm iteratively generates a candidate sample from a proposal distribution. It then computes an acceptance probability

based on the ratio of the target densities (adjusted by the proposal densities) to ensure detailed balance. If a random draw from a uniform distribution is less than this acceptance probability, the candidate becomes the new state; otherwise, the current state is retained. Over many iterations, this process produces a chain of samples that, after discarding a burn-in period, approximates the target distribution accurately. For more details on MH, see Chib and Greenberg [7].

CASE STUDY

The case study used for the performance analysis of PI-ADoK is the catalytic toluene hydrodealkylation to benzene, where toluene ($C_6H_5CH_3$) and hydrogen gas (H_2) is transformed to benzene (C_6H_6) and methane (CH_4). The reaction is shown below:



The kinetic rate model that describes the evolution of the concentrations of $C_6H_5CH_3$, H_2 , C_6H_6 and CH_4 through time is shown below. This expression has been directly borrowed from Fogler [8].

$$r = -\frac{dC_T}{dt} = -\frac{dC_H}{dt} = \frac{dC_B}{dt} = \frac{dC_M}{dt} = \frac{K_A C_T C_H}{1 + K_B C_B + K_C C_T}. \quad (2)$$

In Equation (2), C_T , C_H , C_B and C_M represent the concentration of reactants toluene and hydrogen, and of products benzene and methane, respectively. The kinetic parameters of the kinetic rate model are represented by K_i where $i \in [A, B, C]$.

The computational experiments, chosen randomly from a 2^k factorial design, are run with the following initial conditions (in molar units): $(C_T(t=0), C_H(t=0), C_B(t=0), C_M(t=0)) \in \{(1, 8, 2, 3), (5, 8, 0, 0.5), (5, 3, 0, 0.5), (1, 3, 0, 3), (1, 8, 2, 0.5)\}$. For each experiment, the concentration of the reactant and products are recorded 15 times, at evenly spaced intervals between time $t_0=0$ h and $t_f=10$ h. The kinetic parameters were defined as: $K_A=2$ $M^{-1} h^{-1}$,

$K_B=9 \text{ M}^{-1}$ and $K_C=5 \text{ M}^{-1}$.

Gaussian noise is added to the in-silico measurements to simulate a realistic chemical system. The added noise had zero mean and a standard deviation of 0.2 for C_T , C_H , C_B and C_M . This noise addition allows the approximation of the response of a real system.

RESULTS AND DISCUSSION

Our primary objective was to evaluate if the implementation of constraints could enhance the performance of our discovery algorithm, PI-ADoK-S. For this purpose, we introduced specific constraints at two different stages: the estimation of derivatives and the estimation of rate models. During the derivative estimation phase, the constraints included are: (I) ensuring that the proposed concentration models align with the accurately measured initial conditions of the system, (II) requiring concentration models to converge to a concentration value to simulate equilibrium, (III) ensuring that concentration models do not predict negative concentrations to maintain physical plausibility, and (IV) specific to our case study, ensuring reactant concentrations always decrease while product concentrations always increase. In the rate model estimation stage, we applied the following constraints: (I) ensuring the rate of consumption (or production) of reactants (or products) is consistently negative (or positive), and (II) that these rates should always be decreasing (or increasing) as per the specifics of our case study.

To assess the efficacy of these constraints, we conducted a benchmarking study between the unconstrained ADok-S and the constrained PI-ADok-S. Assuming an unlimited experimental budget, we ran as many extra experiments as required to uncover the data-generating rate model presented in Eq. (2) (the extra experiments are determined by solving the Hunter-Reiner criterion – a MBDok technique for optimal model discrimination [1]). Due to the inherent stochasticity of genetic programming and its sensitivity to initial seeding, we conducted three independent runs. From these, we observed that ADok-S required 16 experiments to identify the ground-truth model, whereas PI-ADok-S needed only 10 – representing a 37.5% reduction in experimental effort. However, further investigations – such as ablation studies and exploring the impact of hyperparameters in the constraint penalty – would provide a more in-depth understanding of the methodology itself.

CONCLUSIONS

This paper introduces PI-ADok-S, an advancement of the ADok-S algorithm, featuring the integration of mathematical constraints for prior knowledge injection and an uncertainty quantification method using the

Metropolis-Hastings algorithm. The inclusion of constraints streamlines the genetic programming algorithm, significantly reducing the experimental burden by 37.5% on average, though further testing is beneficial for deeper understanding of these results. For instance, additional testing could examine the impact of hyperparameters, individual constraints, and initial DoE strategies on the performance of PI-ADok-S. The uncertainty quantification aspect enhances the understanding of the reliability of a certain model's predictions, which can be used to guide future investigative efforts. These improvements underscore PI-ADok-S as a significant upgrade over ADok-S, both in terms of efficiency and depth of analysis, marking it as a valuable tool for experts in kinetic modeling and chemical process research.

ACKNOWLEDGEMENTS

This work was supported by the EPSRC Centre of Doctoral Training for Next Generation Synthesis & Reaction Technology (rEaCt) funding grant EP/S023232/1.

REFERENCES

1. Miguel Ángel de Carvalho Servia, Ilya Orson Sandoval, King Kuok (Mimi) Hii, Klaus Hellgardt, Dongda Zhang, Ehecatl Antonio del Rio Chanona. The automated discovery of kinetic rate models – methodological frameworks. *Digit Discov* 3(5):954–968 (2024) <https://doi.org/10.1039/D3DD00212H>
2. Miguel Ángel de Carvalho, Ehecatl Antonio del Rio Chanona. Model Structure Identification. In: *Machine Learning and Hybrid Modelling for Reaction Engineering*, Ed: Dongda Zhang, Ehecatl Antonio del Rio Chanona. Royal Society of Chemistry (2023).
3. William Hunter, Albey Reiner. Designs for Discriminating Between Two Rival Models. *Technometrics*. 1965 Aug;7(3):307-23.
4. Gabriel Kronberger, Fabricio Olivetti de França, Bogdan Burlacu, Christian Haider, Michael Kommenda. Shape-Constrained Symbolic Regression – Improving Extrapolation with Prior Knowledge. *Evol Comput* 30(1): 75-98 (2022) https://doi.org/10.1162/evco_a_00294
5. Christian Haider, Fabricio Olivetti de França, Bogdan Burlacu, Gabriel Kronberger. Shape-constrained multi-objective genetic programming for symbolic regression. *Applied Soft Computing* 132: 109855 (2023) <https://doi.org/10.1016/j.asoc.2022.109855>
6. Iwo Bładek, Krzysztof Krawiec. Solving symbolic regression problems with formal constraints. In: *GECCO '19: Proceedings of the Genetic and Evolutionary Computation Conference*, Ed: Manuel

López-Ibáñez. Association for Computing Machinery (2019)

7. Siddhartha Chib, Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. Am Stat 49(4):327-335 (1995)
<https://doi.org/10.2307/2684568>
8. H. Scott Fogler. Elements of Chemical Reaction Engineering. Pearson (2016).

© 2025 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See <https://creativecommons.org/licenses/by-sa/4.0/>

