Automated discovery of enzymatic reaction kinetics using symbolic regression guided model-based design of experiments

Harry Kay[1], Alexander Rogers[1], Dongda Zhang[1]*

[1]*Department of Chemical Engineering, University of Manchester, Oxford Road, M1 3AL, Manchester, UK,*

**Corresponding author: dongda.zhang@manchester.ac.uk*

Highlights

- Developed a framework for the accelerated automatic knowledge discovery of novel systems.
- Proposed new experiments based on a balance of exploitation and exploration.
- Symbolic regression proved capable of generating accurate and generalisable models.
- Framework tested for a variety of enzymatic reaction kinetics of increasing complexity.

1. Introduction

Developing accurate and predictive kinetic models to describe and optimise chemical and biochemical reactions is a key priority within the field of reaction engineering. It is typical that, for any novel system, a series of experiments must be conducted to obtain data for gaining knowledge and for parameter estimation for any developed models. After a thorough analysis using literature and making assumptions throughout the modelling process, a set of rate expressions can be constructed to represent the system. This approach often impose inductive bias and, if the assumptions made are incorrect or are only valid within specific operating regions, the developed model will lack generalisability and may not be suitable for optimisation purposes. Ideally, a situation exists where we can automatically construct models based on the data provided through experimental analysis and use these expressions to optimally guide the next experimental evaluation. In doing so, we gain the maximum information possible with the smallest experimental burden and impose no modelling inductive bias onto the generated expressions.

Previous works have aimed to complete these tasks independently such as employing a Bayesian optimisation framework for model-based design of experiment (MBDOE) or exploiting the paradigm of machine learning to accelerate modelling efforts. However, the simultaneous and autonomous construction of expressions that guide optimal experimental evaluations to maximise information gain (such that we can develop the most robust and representative model structures) is scarcely found within the literature. In this work we introduce a symbolic regression-based design of experiment in order to achieve this goal, allowing for accelerated and interpretable understanding, optimisation and control of novel systems.

2. Methods

The case studies chosen are different types of enzymatic reactions given their ever-increasing interest in developing sustainable biotechnologies, where we aim to discover the ground-truth rate expressions that describe the underlying reaction mechanisms using a symbolic regression-based optimal design of experiment framework as shown in Figure 1.

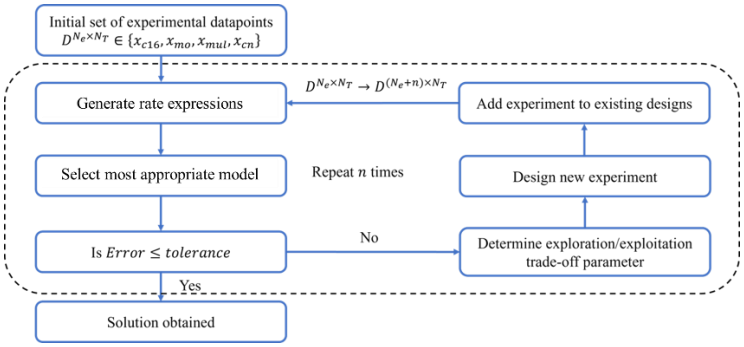


Figure 1 - Flowchart representation of the symbolic regression-based design of experiment framework. N_e and N_T represent the number of initial experiments and the number of timesteps respectively.

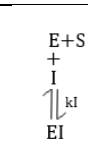
In Figure 1, the system of rate expression generator is constructed through symbolic regression. This is an interpretable machine learning technique that employs genetic algorithms to obtain the "fittest" of the populations of symbolic expressions. This is achieved by searching through a list of operators, functions, and numerical constants to find the expressions that most accurately correlate the predictor matrix to the response matrix. To assist the genetic algorithm, the operator domain can be manually restricted to contain a feasible set relevant to the case study as shown within enzymatic reactions in the literature (i.e., $+, -, \times, \div$). In addition, we can enforce a tree structure for the reaction networks to follow to improve the results of the algorithm and reduce the search space, hence increasing the solution quality.

Subsequent experimental datapoints are chosen through a balance of exploitation and exploration hence, maximising information gain and generating models valid within a large region of the operating space. Here, exploration refers to the discrepancy between the previous experimental outcome and the predicted outcome over the different proposed mechanisms, whereas exploitation is the discrepancy between the experimental outcome and a target outcome. The tradeoff parameter, α is determined such that a value of 0 represents pure exploitation and a value of 1 represents pure exploration.

3. Results and discussion

The framework is tested on a series of enzymatic reactions of increasing complexity as shown in Table 1. The first equation developed from traditional Michaelis-Menten kinetics, the second for allosteric enzymes, and the third for inhibited enzyme kinetics.

Table 1 - Table describing the models for the enzymatic reactions used to test the symbolic regression-based MBDOE framework (Shuler & Kargi. Fikret, 2017).

Model	Reaction	Description
$v = \frac{V_m[S]}{K_m + [S]}$	$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightarrow} E + P$	Traditional Michaelis-Menton kinetics
$v = \frac{V_m[S]^n}{K''_m + [S]^n}$		Allosteric enzymes – n represents the cooperativity coefficient.
$v = \frac{V_m}{\left(1 + \frac{[I]}{K_1}\right) + \left(1 + \frac{K_m}{[S]}\right)}$		Enzymes with inhibition effects

The framework proved capable of recovering knowledge for all complexities of models, boasting low errors and high generalisability within the search space. It was also able to discriminate expressions not abiding by correct structures. In addition, the use of the exploration/exploitation trade-off parameter can systematically determine if the next experiment should prioritise knowledge exploration or enhancing the accuracy of the proposed model. As such, no human intervention is involved in the framework.

4. Conclusions

In conclusion, symbolic regression was shown capable for automatic discovery of expressions for varying complexities of systems. Through the implementation of constrained tree structures, and a suitable domain of operators, accurate rate expressions were able to be generated for enzymatic reaction networks. Although exact expressions may not be recovered for extremely complex reaction systems, the current framework can still provide physical insight into the reaction network, and model the systems to high accuracies. In addition, due to the balance of exploration and exploitation when evaluating proposing optimal experimental datapoints, the models are valid within large regions of the operating domain (are generalisable). Therefore, the developed framework shows potential for modelling of novel enzymatic reaction kinetics and on a wider scale.

References

[1] M.L. Shuler, F.Kargi, & M. DeLisa (2017), Bioprocess Engineering: Basic Concepts 3rd Edition. 75.

Keywords

Symbolic regression, design of experiment, enzymatic reaction network, automatic knowledge discovery