Data-driven Discovery of Reaction Kinetic Models in Dynamic Plug Flow Reactors using Symbolic Regression

Ben Cohen,a,c Burcu Beykal,a,b George M. Bollasa,c

aDepartment of Chemical & Biomolecular Engineering, University of Connecticut, Storrs, CT, USA

bCenter for Clean Energy Engineering, University of Connecticut, Storrs, CT, USA

cPratt & Whitney Insitute for Advanced Systems Engineering, University of Connecticut, Storrs, CT, USA

george.bollas@uconn.edu

Abstract

The method of characteristics is combined with symbolic regression to identify kinetic models of reactions taking place in a dynamic, ideal plug flow reactor (PFR). A change in coordinates facilitates the transformation of time-series measurements collected at a reactor’s outlet to data along characteristic curves. The method is applied to three synthetic reactors: one with a nonisothermal, irreversible chemical reaction, one with a reaction described by Hougen-Watson kinetics, and one with measurement noise. The results show that the proposed combination of tools can use time-series data collected at the outlet of a dynamic PFR to discover the Arrhenius expression, rate limited kinetic models, and even analytical solutions to the PFR equation. The toolchain is also flexible enough to allow for discovery of kinetic models even in the presence of measurement noise. Overall, the proposed method offers a simple approach to learn concise kinetic models from dynamic, ideal PFRs and more generally models of source terms in systems described by partial differential equations.

**Keywords**: Symbolic regression, surrogate modelling, machine learning, genetic algorithms, chemical kinetics

* 1. Introduction

The surge in available data within the chemical industry has prompted widespread adoption of data-driven modeling tools among process engineers (McBride and Sundmacher, 2019). Many of these methods generate black- or grey-box models and require significantly less oversight than traditional first principles models during development. While these models can offer improved process control (Beykal et al., 2022) and design, a persistent challenge lies in establishing trust due to their poor extrapolative ability beyond the bounds of their training data and limited interpretability.

To improve the extrapolative and interpolative power of data-driven models, Physics-informed Machine Learning (PIML) emerged as an approach to incorporate domain knowledge into the model training process. PIML methods construct models that are consistent with physics, enhancing their ability to extrapolate beyond the training data and making them easier to trust in many applications (Karniadakis et al., 2021). Despite this integration of domain knowledge, interpreting models built using many methods in PIML remains challenging, in part due to the many parameters in data-driven models.

Symbolic regression (SR), as a method in PIML, improves the interpretability of data-driven models while leveraging domain knowledge. It can transform data into concise mathematical representations, resembling and often embodying mechanistic models. In essence, SR provides a data-driven and automated alternative to the traditional expert-driven model ideation process, where a domain expert formulates mechanistic models based on theory, heuristics, and experience. The challenging nature of kinetic model discovery makes SR particularly attractive for this purpose.

The application of SR to discover kinetic models has predominately centered around identifying ordinary differential equation (ODE) models for systems varying in one dimension, typically time (Narayanan et al., 2022). Current SR methods for ODEs often necessitate extensive data sampled across the time domain or involve computationally expensive integration steps (Cornforth and Lipson, 2013). Methods that avoid forward integration approximate derivatives directly from data or rely on surrogate models for determining derivatives analytically.

Despite the progress in SR for discovery of kinetic models in systems described by ODEs, its application to reactors better characterized by partial differential equation (PDE) models – like dynamic plug flow reactors (PFRs) – remains limited. SR methods for PDE discovery encounter similar challenges to their ODE counterparts: dependence on extensive or challenging-to-obtain data (Rudy et al., 2017) or expensive integration (Cohen et al., 2023 under review). Consequently, the endeavor of learning kinetic models using SR from input-output relationships in systems best described using PDEs remains prohibitively expensive.

This work introduces an innovative approach to the discovery of kinetic models specifically tailored for ideal PFRs with constant thermophysical properties. In contrast to prior SR methods for PDE discovery, the proposed approach acknowledges that the PFR equation is already known and directs its focus towards identifying the unknown reaction term. To achieve this, the method of characteristics (MoC) is used to enact a change in coordinates such that the reactor dynamics occur in one dimension. The new coordinates are then leveraged to transform the time-series data collected at the PFR outlet into a dataset sampled across the direction that captures the reactor’s dynamics. Subsequently, established methods in ODE model discovery can be applied to reveal the unknown kinetic model without forward integration.

* 1. Methodology

The ultimate objective of this work is to demonstrate how to automatically discover the unknown reaction term ( from input-output data of an ideal PFR with constant thermophysical properties governed by Eq. (1). The state variables encompass chemical species concentrations and temperature , consolidated as the state vector . The variables , , and respectively represent flow velocity through the reactor, space, and time. The reactor’s domain is described as where is the final time. The subscripts and denote partial derivatives with respect to time and space.

|  |  |
| --- | --- |
|  | (1) |

Automatic discovery of the reaction term using existing SR methods requires either significant data collected across , or the forward integration of the PFR equation many times. In this work, is reparametrized into and where curves along are curves of constant and curves along are characteristic curves. This change of coordinates can be leveraged to transform the time-series data collected at the outlet of the PFR into an -series dataset which captures all the dynamics in the reactor. That -series dataset can then be used to approximate using either finite differences or taking the analytical derivative of some surrogate model describing in . The derivative can then be used as a target value by SR to discover .

* + 1. Change of Coordinates

The change of coordinates from to is done using the Method of Characteristics (MoC). The MoC requires Cauchy data collected along a manifold, . To discover the kinetic model, the Cauchy data provided is the time-series state measurements collected on , where is defined as the outlet of the reactor (. These measurements are denoted as on . Using the MoC and , the PDE provided in Eq. (1) can be rewritten as the system of ODEs in Eq. (2).

|  |  |
| --- | --- |
|  | (2) |

|  |  |
| --- | --- |
| Since the thermophysical properties of the reactor are assumed to be constant, the ODEs which describe and are independent of the ODEs which describes in Eq. (2). This means given state measurements that span some range of determined by solving the ODEs for and for many values of , the derivative of with respect to can be approximated and used to discover using SR. Care, however, must be taken when selecting a dynamic operating condition that will result in data spanning some range in . In this work, the flow velocity in Eq. (3), realized by a synthetic control valve that opens at the start of each experiment, helps produce varying values of across . | (3) |

* + 1. Symbolic Regression

Once the time-series data is transformed into an -series dataset using the MoC, SR via Genetic Programming (GP) can be used to discover the structure and parameter values of . GP can search a symbol space for an optimal expression by applying genetic operations inspired by biological evolution to expressions represented as expression trees. These genetic operators, crossover and mutation, can be iteratively applied to a population of expression trees to evolve an optimal expression. To search a symbol space, GP needs an argument set, or set of variables that it can include in the expressions, and a primitive set, or set of mathematical operators it can use to create mathematical relations between the arguments. Both sets are selected based on domain expert knowledge of the system of interest. The GP also needs a probability of crossover and a probability of mutation that determine how likely it is that any expression tree within a population will be subjected to the genetic operations crossover and mutation respectively.

To identify an optimal expression, the GP searches for an expression which accomplishes two goals: minimize the mean squared error () between the expression when evaluated using the measured data and the target value, ; and be as concise as possible. The Bayesian Information Criterion () considers both these expectations and thus was used as a fitness criterion for the SR as defined in Eq. (4), where represents the complexity of the expression defined as the number of nodes in the expression tree and represents the number of data points used to train the model.

|  |  |
| --- | --- |
|  | (4) |

To improve the performance of the GP, a gradient-based parameter estimation scheme was used during the evaluation of each expression. The objective of the parameter estimation step was to minimize the for each expression. This step helps prevent good model structures with bad parameter estimates from being discarded due to poor fitness.

* 1. Results

To demonstrate the proposed method, the discovery of three kinetic models were explored: Discovery of a nonisothermal kinetic model, discovery of a rate limited kinetic model, and discovery of kinetic models in the presence of noise. Every case study was tested 30 times. Since SR via GP is not deterministic, there are no guarantees that the underlying model will be returned from each trial, but the results demonstrate that the method is robust and can return a wide range of different types of kinetic models.

* + 1. Discovery of nonisothermal kinetic models

The first cast study investigates an ideal plug flow reactor with constant thermophysical properties home to the nonisothermal reaction . Three synthetic experiments were conducted and the time-series data for and were collected at the outlet of the reactor. The details of the experiments are shown in Table 1 for the Nonisothermal case where the subscript denotes feed, and the subscript denotes coolant. The time-series data collected at the outlet from the three experiments were then transformed into an -series dataset.

Table 1: Data-generating experiments for each reactor and the genetic programming hyperparameters argument set (Arg. Set), primitive set (Prim. Set), number of individuals in a population (POP), and number of generations of evolution (GEN).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Case | Experiments | | | Arg. Set | Prim. Set | POP | GEN |
| Nonisothermal |  |  |  |  |  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Hougen-Watson |  |  |  |  |  |  |  |
|  |  |  |
|  |  |  |
| Analytical |  | | |  |  |  |  |
|  | | |
| Isothermal |  | | |  |  |  |  |
|  | | |

From the -series dataset, was calculated using finite difference approximations. These approximations served as the target for the GP with the argument set, primitive set, number of individuals in the population, and number of generations shown in Table 1. The unknown parameters were represented by . In this work, the parameters were tuned using a gradient-based parameter estimation scheme as described in Section 2.2. The probabilities of crossover and mutation were and respectively.

The proposed method was successful in discovering the underlying kinetic model, including the Arrhenius expression, in of the trails conducted. The small errors, less than error, in the parameter estimates shown in Table 2 for the Nonisothermal case are due to numerical errors in the derivative approximations. The average time to return the kinetic model was seconds.

* + 1. Discovery of rate limited kinetic models

While many reactions are nonisothermal, few are represented as simply as the monotonic reaction in Section 3.1. To apply the proposed toolchain to a more complicated kinetic model, a reactor with Hougen-Watson kinetics was considered. Two synthetic experiments were run as presented in Table 1. The argument set, primitive set, size of the population, and the number of generations are also shown in Table 1. The probabilities of crossover and mutation were and respectively. In of the trials, the proposed method identified the correct kinetic model with parameter error less than , shown in Table 2, in an average of 104 seconds.

* + 1. Discovery of kinetic models in the presence of noise

Although the previous two examples used noiseless data, data collected from physical reactors always exhibit some level of noise. To demonstrate how the proposed method can handle noise, a simple PFR home to an isothermal reaction was considered. One experiment was run with a constant feed concentration of , , as shown in Table 1 for the Analytical and Isothermal cases.

The synthetic time series data, with added Gaussian noise collected at the outlet was transformed into an -series dataset. SR was then used to discover a function using the argument and primitive set shown in Table 1 for the Analytical case. The derivative of was then taken analytically to build the target for the SR of with the argument and primitive sets shown for the Isothermal case in Table 1. The GP hyperparameters for both cases are shown in Table 1 and the probabilities of crossover and mutation were and respectively.

Table 2: Discovered Expressions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Case | Ground Truth | Discovered | Successful Runs | Average Time (s) |
| Nonisothermal |  |  |  |  |
| Hougen-Watson |  |  |  |  |
| Analytical Model |  |  |  |  |
| Isothermal |  |  |  |  |

The resulting expressions and are shown in the last two rows of Table 2. of the trials returned the correct with less than parameter error and of the trials returned the correct isothermal kinetic model with less than parameter error. The lower success in discovery of is due to the simple dynamics in which can be well represented using many expressions creating using the large primitive and argument sets, and the smaller population than the previous two studies. The average time to complete the search for the analytical model and the isothermal kinetics were 3.92 seconds and 4.93 seconds respectively.

Using the two-stage approach, the proposed method can identify simple reaction models from noisy state measurements collected at the outlet of a PFR. It should be noted that the discovered function, , is the analytical solution to the PFR equation with the chemical reaction model discovered. This is especially convenient when such an analytical solution exists, however even when an analytical solution does not exist, SR can discover simple expressions that describe ODEs well as shown in Tsoulos and Lagaris (2006).

* 1. Conclusions

Discovery of kinetic models form input-output data of PFRs remains a challenging task. By using the MoC to change coordinates in ideal PFRs, SR becomes a realistic and inexpensive approach to learn kinetic models automatically from data. This chain of methods reliably identified the Arrhenius expression, Hougen-Watson Kinetics, and discovered the analytical solution to the PFR equation.

The demonstrated method is reliable, consistently returning the ground truth model from data. It is also flexible and capable of returning interesting model structures that are not defined a priori. The many methods in SR make the proposed tool chain flexible and adaptable to different reaction schemes and available data so long as the reactor has near constant thermophysical properties.

References

Beykal, B., Diangelakis, N.A., Pistikopoulos, E.N., 2022. Continuous-Time Surrogate Models for Data-Driven Dynamic Optimization, in: Montastruc, L., Negny, S. (Eds.), 32nd European Symposium on Computer Aided Process Engineering, Computer Aided Chemical Engineering. Elsevier, pp. 205–210.

Cohen, B., Beykal, B., Bollas, G., 2023 under review. Physics-informed genetic programming for discovery of partial differential equations from scarce and noisy data.

Cornforth, T.W., Lipson, H., 2013. Inference of hidden variables in systems of differential equations with genetic programming. Genet Program Evolvable Mach 14, 155–190.

Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L., 2021. Physics-informed machine learning. Nature Reviews Physics 3, 422–440.

McBride, K., Sundmacher, K., 2019. Overview of Surrogate Modeling in Chemical Process Engineering. Chemie Ingenieur Technik 91, 228–239.

Narayanan, H., Cruz Bournazou, M.N., Guillén Gosálbez, G., Butté, A., 2022. Functional-Hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. Chemical Engineering Journal 430, 133032.

Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N., 2017. Data-driven discovery of partial differential equations. Sci Adv 3.

Tsoulos, I.G., Lagaris, I.E., 2006. Solving differential equations with genetic programming. Genet Program Evolvable Mach 7, 33–54.