



Short Communication

Grammar-guided symbolic regression as tool for model development of enzymatic reactions

Thomas Waluga^{a,*}, Jorge Adrian Peña Ayala^a, Miriam Edel-Teichmann^b, Mirko Skiborowski^a^a Institute of Process Systems Engineering, Hamburg University of Technology, Hamburg, Germany^b Institute of Technical Microbiology, Hamburg University of Technology, Hamburg, Germany

ARTICLE INFO

Keywords:

Grammar-Guided Symbolic Regression

Kinetic modelling

Enzyme reactions

Isomerase

ABSTRACT

The transition of the chemical industry toward renewable biocatalysis requires robust kinetic models to scale and optimize enzymatic processes. However, deriving these models typically relies on manual, trial-and-error fitting of predefined equations, while data-driven machine learning approaches often produce accurate but physically meaningless “black-box” models. Here, we present an adaptive kinetic model evaluation framework that automates the discovery of enzymatic rate laws directly from batch progress curves. By employing Grammar-Guided Symbolic Regression (GGSR) the algorithm autonomously constructs rate equations using only physically permissible kinetic building blocks. The framework’s reliability was first validated against 13 synthetic kinetic mechanisms, confirming structural and parametric identifiability. The pipeline was then applied to a real experimental case study of glucose-fructose isomerization using *Streptomyces murinus* isomerase. The algorithm successfully recovered a reversible kinetic equation featuring an apparent uncompetitive-type interaction term providing specific activities and Michaelis constants that align well with literature. This work demonstrates that GGSR can bridge the gap between data-driven automation and mechanistically interpretable biochemical engineering.

1. Introduction

Enzymes offer enormous potential for the transition of the chemical industry from finite fossil-based resources toward renewable feedstocks. Their high selectivity makes them attractive catalysts for a wide range of chemical reactions (Waluga et al., 2026). To transfer such processes to industrial scale, process modelling represents a powerful and indispensable tool (Boodhoo et al., 2022). The core of such models is the kinetic description of the enzymatic reaction in the form of mathematical equations. This equation is used to describe concentration profiles under varying reaction conditions. The mathematical structure itself can also provide insight into the underlying reaction mechanism and thereby contribute to a molecular-level understanding of the process (Cornish-Bowden, 2012). The most fundamental equation in enzyme kinetics is the Michaelis–Menten kinetics model as shown in Eq. (1),

$$v = \frac{v_{\max} c_S}{K_M + c_S} \quad (1)$$

with $v_{\max}$ as maximal reaction rate, c_S as substrate concentration and K_M

as Michaelis–Menten constant. Nevertheless, this equation only represents the simplest case of an enzymatic reaction and can be extended by additional mechanistic contributions such as external, substrate or product inhibition as well as multiple substrates and products as well as reversible reaction steps or Hill-type cooperativity (Cornish-Bowden, 2012; Waluga and Nagshi, 2026).

Symbolic regression has emerged as a powerful approach for deriving mathematical relationships directly from experimental data (Reissmann et al., 2025; Lima et al., 2025). These methods follow the principle of parsimony seeking the most simple and accurate models, by penalizing their complexity. Parametric methods, such as the ALAMO approach (Cozad et al., 2014) derive a simple model by the optimal weighting of an a-priori known combination of basis functions, solving an optimal subset selection problem with a model fitness metric derived from information theory (Wilson and Sahinidis, 2017). Non-parametric methods for a symbolic regression do not require a-priori selection of basis functions, but simultaneously seek an optimal model structure and parametrization represented by a syntax tree or directed acyclic graph, applying genetic programming (McKay et al., 1997) or deterministic global optimization (Cozad and Sahinidis, 2018). Such methods have e.

* Corresponding author.

E-mail address: thomas.waluga@tuhh.de (T. Waluga).<https://doi.org/10.1016/j.dche.2026.100346>

Received 25 June 2026; Received in revised form 31 August 2026; Accepted 22 September 2026

Available online 23 September 2026

2772-5081/© 2026 The Authors. Published by Elsevier Ltd on behalf of Institution of Chemical Engineers (IChemE). This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Open access funding enabled and organized by Projekt DEAL.

g. successfully identified basic physicochemical laws (Neumann et al., 2020), kinetic laws in material science (Wang et al., 2019) or phenomenological models for flux and rejection predictions of membrane separations (Goebel et al., 2020; Goebel and Skiborowski, 2020).

However, since non-parametric symbolic regression methods allow for a rather unrestricted combination of mathematical operators, they may generate functions that accurately describe experimental data but lack physically or mechanistically meaningful interpretation (Reissmann et al., 2025; Crochepierre et al., 2022; Lima et al., 2025). To avoid this problem, Grammar-Guided Symbolic Regression (GGSR) provides a highly promising alternative by restricting the symbolic model search to expressions generated by predefined grammar. This grammar specifies the admissible mathematical structures and can encode prior knowledge such as physical consistency, mechanistic assumptions, dimensional constraints, or known functional relationships. This enables the resulting model to be constructed exclusively from physically meaningful building blocks into the resulting kinetic model (Reissmann et al., 2025; Narayanan et al., 2022). Moreover, the modular nature of GGSR (Hoai et al., 2026) aligns particularly well with the structure of enzyme kinetic equations, where individual mechanistic contributions can be incorporated as building blocks of the overall kinetic model.

While GGSR has been proposed as a general framework for incorporating domain knowledge into symbolic model discovery, to the best of our knowledge, this study represents the first investigation of GGSR applied to enzyme kinetics. As an initial proof-of-concept study, the scope is limited to a simple uni-uni reaction system involving one substrate molecule and one product molecule. The objective is to evaluate whether grammar-guided approaches can recover mechanistically meaningful kinetic expressions while maintaining sufficient predictive accuracy for experimental data. GGSR relocates domain expertise from manual equation selection toward grammar construction and model adjudication, rather than eliminating it.

2. Material and methods

2.1. Grammar-Guided Symbolic Regression

A custom workflow for adaptive kinetic model evaluation was implemented in MATLAB. The overall framework consisted of four consecutive blocks (cf. Fig. 1): (i) data loading or synthetic data generation and standardization, (ii) pre-processing and extraction of kinetic information from experimental data by the use of spline interpolation (Markus and Plesser, 1981; Waluga et al., 2025) as constrained smoothing splines fitted under monotonicity and concavity constraints that regularize the concentration data prior to differentiation showed the lowest average parameter-estimation error under Gaussian noise of up to 10% relative amplitude, and the widest range of viable initial

parameter guesses, among the methods compared in a dedicated prior study (Waluga et al., 2025), (iii) GGSR for kinetic model discovery, and (iv) model validation and visualization. Block (ii) also produces a preliminary Michaelis-Menten fit used only to scale parameter bounds and initial guesses in Block (iii), not to restrict candidate model structure. All code as well as experimental data are available in our repository, listed in the Data Availability statement.

The symbolic regression approach was based on a grammar-guided search strategy in which predefined mechanistic building blocks were iteratively combined to construct kinetic equations. The grammar consisted of physically interpretable numerator and denominator terms representing common enzymatic reaction motifs such as substrate binding, reversibility, and inhibition phenomena. Exemplary building blocks are illustrated in Fig. 2. By construction, denominator building blocks are dimensionless concentration ratios with matching-unit constants, so admissible combinations remain dimensionally homogeneous.

In contrast to unrestricted symbolic regression, the search space was constrained by a predefined reaction grammar to avoid physically implausible model structures (Crochepierre et al., 2022). The regression procedure was implemented as a custom multi-phase adaptive beam search algorithm based on the principles of beam search (Lemons et al., 2022). Initially, experimental data were flattened into concentration vectors and corresponding observed reaction rates. The search started from a null model and iteratively expanded candidate equations by adding admissible numerator or denominator terms according to the predefined grammar. The search was restricted to a one-parameter orthogonal grammar during the mechanistic discovery phase. Hill-type cooperativity terms were intentionally excluded at this stage to prevent phenomenological overfitting during early model construction.

For each expansion step, model parameters were estimated by nonlinear least-squares regression using the MATLAB function *lsqcurvefit*, via an eight-point multi-start strategy to mitigate convergence to local minima. Parameter uncertainties (95% confidence intervals and relative standard errors reported throughout) were derived from the Jacobian of the nonlinear least-squares fit, which provides an approximate parameter covariance matrix. Candidate models were evaluated

$$v = \frac{v_{max}c_s}{(K_M(1 + \frac{c_{ic}}{K_{ic}}) + c_s(1 + \frac{c_{iu}}{K_{iu}}))(1 + \frac{c_{in}}{K_{in}})}$$

Building Block for: competitive inhibition
uncompetitive inhibition
non-competitive inhibition

Fig. 2. Concept for the building blocks. Selected examples for the inhibition of a simple irreversible uni-uni reaction.

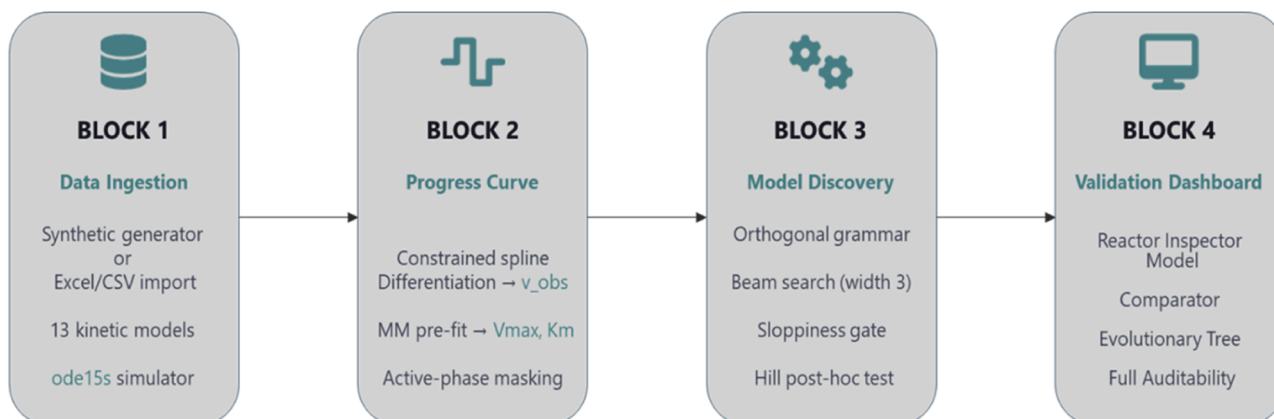


Fig. 1. Pipeline of the adaptive kinetic model evaluation. Each block is a standalone MATLAB function. Struct-based data contracts ensure modularity.

using the corrected Akaike Information Criterion (AICc) (Hurvich and Tsai, 1989), which served as the primary model selection metric. Only models that improved the AICc by at least two units relative to their parent model were retained for further exploration. This threshold is consistent with the interpretation that ΔAICc values below two provide little evidence in favor of one model over another (Burnham and Anderson, 2004). To reduce the risk of convergence toward local minima, multiple candidate branches were maintained simultaneously using a beam width of three. Both settings are user-adjustable rather than fixed constants. To avoid structurally unstable solutions, an additional identifiability filter was incorporated into the workflow. Candidate models were excluded from further consideration if parameter perturbation sensitivity analysis revealed poorly identifiable or weakly constrained parameter combinations, here referred to as “sloppy” behavior. This step was intended to remove models with poorly identifiable parameter combinations despite apparently good numerical fits. After convergence of the mechanistic search phase, a second phase was performed to evaluate potential cooperative substrate effects. In this step, linear substrate-related terms were systematically transformed into Hill-type expressions containing an additional cooperativity exponent. These transformed models were only accepted when two criteria were fulfilled simultaneously: (i) a significant improvement in AICc relative to the corresponding mechanistic ancestor model and (ii) a fitted Hill coefficient of at least 1.6. Candidates are filtered in this fixed order: physical admissibility, the identifiability gate, AICc-based ranking, and finally cross-validation. Candidate Hill models were subsequently ranked using leave-one-out cross-validation (LOOCV) (Arlot and Celisse, 2010), whereby entire progress curves rather than individual time points were withheld to avoid leakage from the per-reactor spline fit, to improve robustness against overfitting.

In the validation phase, the final candidate model was compared against all mechanistic ancestor models within its evolutionary lineage. LOOCV was performed to evaluate the generalization capability of the discovered equations. If a simpler ancestor model exhibited superior predictive performance while only showing a moderate loss in AICc, the simpler model was preferred. This final lineage-wide validation step was introduced to identify and remove unnecessary mechanistic terms that improved the in-sample fit but reduced generalizability.

2.2. Experimental details for a real experimental case study

In addition to synthetic data, real experimental data were generated. To this end, a reversible isomerization of fructose to glucose was carried out using an isomerase. The experimental conditions can be summarized as follows: the reaction took place in a 50 mL glass reactor at 45 °C in 0.05 M MgCl_2 and 0.01 M TRIS HCl buffer (pH 7.5) assumed as an ideal stirred tank reactor. Substrate was added for the desired concentration. 5.55 g L^{-1} Sweetzyme IT extra (from *Streptomyces murinus* by Novozymes) was used. Samples of 200 μL were periodically withdrawn and filtered. A 100 μL aliquot of each filtrate was immediately quenched with 150 μL of 1 M HCl and 1250 μL of buffer to stop the reaction prior to HPLC analysis. HPLC was performed as described in the Supporting Information.

3. Results

3.1. Code validation with synthetic data

To validate the developed code for parameter estimation using GGSr, synthetic datasets were generated in Block 1. In this case, the true kinetic model and parameters are known *a priori* and can therefore be directly compared with the generated model candidate and estimated parameters. The synthetic data was also superimposed with Gaussian noise of 2% relative amplitude (plus a small additive floor to avoid a degenerate zero-noise condition at zero concentration). The synthetic data generation enables the representation and evaluation of 13

different kinetic models. Exemplary results of model validation are shown in Table 1 and 2 for a simple irreversible Michaelis-Menten kinetic according to Eq. (1) and a more complex reversible Michaelis-Menten kinetic with uncompetitive product inhibition, respectively. During model evolution, candidate models were retained only if they improved the AICc of their direct parent by at least two units and passed an identifiability screening. The final model was selected using a combination of AICc-based ranking and LOOCV to balance goodness-of-fit and generalization performance. The selected models are marked in italic in the Tables 1 and 2. Details of this model validation for all 13 models can be found in our Supporting Information.

The results demonstrate that, under the tested noise level and data-richness conditions, GGSr is capable of correctly identifying the correct model structure as well as true kinetic parameters with a low estimation error (cf. Supporting Information). Near-nested candidates with comparable AICc, such as Models 1 and 2 in Table 1, indicate that some structural ambiguity remains even under these favorable conditions. This holds provided that the rule proposed by Stroberg and Schnell (Stroberg and Schnell, 2016) is fulfilled: “The time scale of time-course curvature dictates accuracy of parameter estimation.” This relationship is illustrated in Fig. 3. Experimental data points must necessarily be located within the curved region of the progress curve. If measurements are only recorded at early time points and low initial substrate concentrations, only the ratio of v_{max} and K_M can be identified reliably, which is a known issue in parameter estimation of enzyme kinetics (Hong et al., 2022). In contrast, at high initial substrate concentrations, only v_{max} can be estimated. Furthermore, if measurements are recorded too late during the reaction progress, reliable parameter estimation is no longer possible.

3.2. Real experimental case study with isomerase

For the case study based on real experimental data, progress curves obtained from an isomerase catalyzed reaction were used. Representative experimental concentration profiles are shown in Fig. 4, while the complete dataset is available in the associated repository.

As is commonly known, the isomerization of sugars represents an equilibrium-limited reaction (Nam, 2022). In principle, it would be sufficient to estimate the kinetic parameters using data from only one reaction direction. However, this approach is known to result in increased uncertainties for the kinetic parameters associated with the reverse reaction (Waluga et al., 2025). Therefore, in addition to two experiments investigating the conversion of glucose to fructose, further experiments were conducted for the conversion of fructose to glucose. As the final candidate model, the GGSr identified a reversible kinetic equation with an apparent uncompetitive inhibition contribution according to Eq. (2a), where Eqs. (2b) and (2c) represent other candidates. The estimated parameters are $v_{\text{Glc}} = 1.48 \pm 2.28 \text{ mM min}^{-1}$, $v_{\text{Fru}} = 2.08 \pm 2.23 \text{ mM min}^{-1}$, $K_{M,\text{Glc}} = 241 \pm 342 \text{ mM}$, $K_{M,\text{Fru}} = 341 \pm 314 \text{ mM}$ and

Table 1

Results of model validation for an irreversible Michaelis-Menten kinetic according to Eq. (1). Automatically selected model in italic. Models 1 and 2 fit the data almost identically (the Hill exponent in Model 2 converges to about 1); their AICc difference reflects only the parsimony penalty for Model 2's extra parameter, not a computational error.

Parameter	Output	Input	
$v_{\text{max fwd}} / \text{mM min}^{-1}$	1.6977 ± 0.0037	1.7000	
K_1 / mM	249.5060 ± 0.4669	250.000	
#	AICc	R ²	Equation
1	-30,826.83	0.9999	$v = (V_f^* C_1) / (1 + C_1/K_1)$
2	30,825.49	0.9999	$v = (V_f^* C_1^n) / (1 + C_1^n/K_1^n)$
3	16,382.88	0.9770	$v = (V_f^* C_1^n) / (1)$

Table 2

Results of model validation for a reversible Michaelis-Menten kinetic with uncompetitive product inhibition. Automatically selected model in *italic*.

Parameter	Output	Input
V _{max,fwd} / mM min ⁻¹	1.6917 ± 0.0214	1.7000
K ₁ / mM	0.9830 ± 0.0150	1.0000
V _{max,rev} / mM min ⁻¹	247.8192 ± 2.7164	250.0000
K ₂ = K _{ic} / mM	341.6503 ± 4.7259	350.0000
K _{iu} / mM	386.4266 ± 6.7615	380.0000
#	AICc	R2
1	50,748.41	0.9982
2	-50,710.87	0.9982
Equation		
$v = (V_f \cdot C_1^n - V_r \cdot C_2) / (1 + C_1^n / K_1^n + C_2 / K_2 + C_1 \cdot C_2 / (K_1 \cdot K_{iu}))$		
$v = (V_f \cdot C_1 - V_r \cdot C_2) / (1 + C_1 / K_1 + C_2 / K_{ic} + C_1 \cdot C_2 / (K_1 \cdot K_{iu}))$		

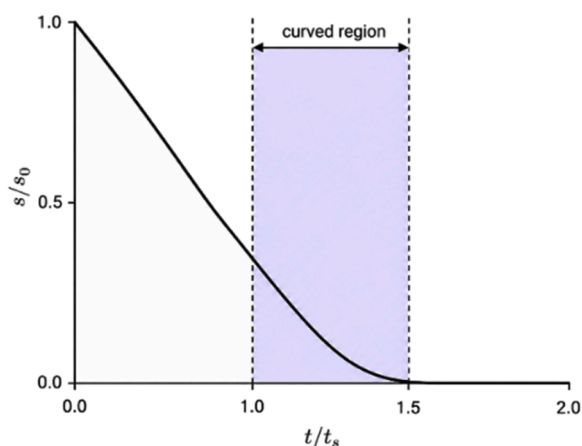


Fig. 3. Substrate concentration over time for an irreversible reaction. The colored region indicates the “curved region”, which is necessary to obtain data for sufficient parameter estimation, adapted from (Stroberg and Schnell, 2016).

$K_{IU} = 275 \pm 463$ mM. All other parameter values and calculated goodness-of-fit metrics of the top 3 model candidates are summarized in Table 3. Candidates 2 and 3 achieve marginally lower AICc and slightly higher R2 than the selected model (Table 3) but are rejected because their fitted Hill coefficients (0.80 and 0.79) fall below the 1.6

admissibility threshold for physically plausible positive cooperativity (Weiss, 1997). Per our selection hierarchy, Table 3 lists them only as rejected alternatives for transparency.

$$v = \frac{V_{Glc} C_{Glc} - V_{Fru} C_{Fru}}{1 + \frac{C_{Glc}}{K_{M,Glc}} + \frac{C_{Fru}}{K_{M,Fru}} + \frac{C_{Glc} C_{Fru}}{K_{M,Glc} K_{iu}}} \quad (2a)$$

$$v = \frac{V_{Glc} C_{Glc}^n - V_{Fru} C_{Fru}}{1 + \frac{C_{Glc}^n}{K_{M,Glc}^n} + \frac{C_{Fru}}{K_{M,Fru}} + \frac{C_{Glc} C_{Fru}}{K_{M,Glc} K_{iu}}} \quad (2b)$$

$$v = \frac{V_{Glc} C_{Glc}^n - V_{Fru} C_{Fru}}{1 + \frac{C_{Glc}^n}{K_{M,Glc}^n} + \frac{C_{Fru}}{K_{M,Fru}}} \quad (2c)$$

The values obtained are in good agreement with literature values. For comparable reaction conditions but with a higher temperature, $K_{M,Fru}$ values in the range of 86 – 335 mM, $K_{M,Glc}$ values in the range of 165 – 219 mM and v_{Glc} in the range of 0.22 – 0.27 U g⁻¹, v_{Fru} in the range of 0.32 – 1.51 U g⁻¹ have been reported (Waluga et al., 2025). Using slightly other reaction conditions in respect to the buffer composition, $K_{M,Fru}$ values of 450 – 822 mM, $K_{M,Glc}$ of 497 – 700 mM and v_{Glc} of 0.29 – 0.49 U g⁻¹, v_{Fru} of 0.19 – 0.80 U g⁻¹ have been reported (Converti and Del Borghi, 1997; Dehkordi et al., 2009). Furthermore, other values for example of $K_{M,Glc}$ in the range of 180–1900 mM have been reported in the literature, summarized by (Waluga et al., 2023). However, these values are only comparable to a limited extent, as the enzymes studied originate from different organisms and the reaction conditions varied. The K_{iu} can partly be rationalized by the reversible character of the reaction, as both substrate and product are able to access the active site (Nam, 2022). At elevated concentrations, additional binding of either substrate or product to the enzyme–substrate complex may lead to the formation of non-productive complexes, resulting in apparent inhibition. Therefore, the additional denominator term should not necessarily be interpreted as evidence for a specific elementary inhibition mechanism, but rather as an apparent substrate/product interaction term that effects the kinetic description under high-concentration conditions.

The present grammar fits the forward and reverse rate parameters independently, without enforcing the Haldane relationship between them. As a post-hoc check, the fitted rate law's equilibrium condition implies a K_{eq} of approximately 0.71 at 45 °C, which in good agreement with reported values of K_{eq} 0.72 and 0.85 at 50 °C and 55 °C, respectively (Dehkordi et al., 2009).

3.3. Extended discussion

3.3.1. Relation to ALAMO and other symbolic regression approaches

ALAMO and similar parametric surrogate-modeling approaches

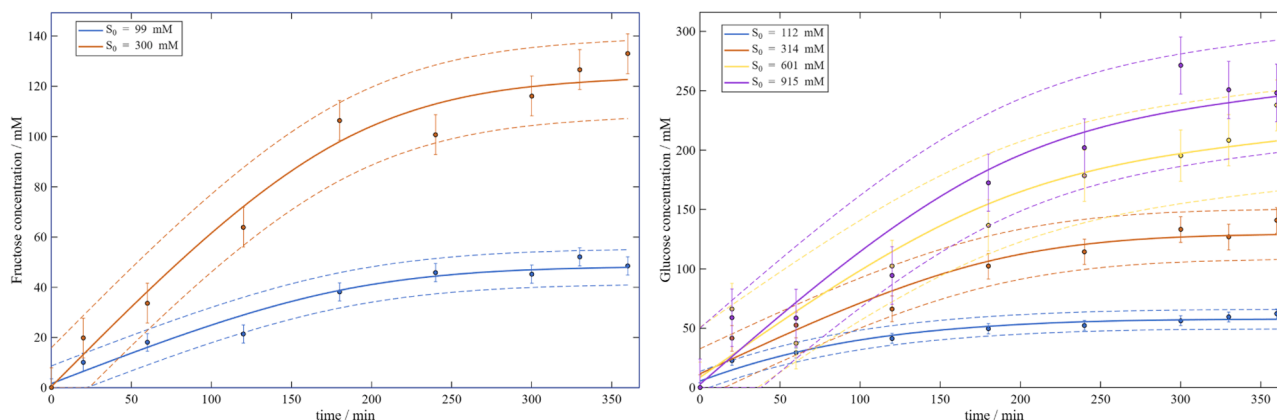


Fig. 4. Representative progress curves of the reversible glucose–fructose isomerization catalyzed by 5.55 g L⁻¹ Sweetzyme IT Extra at 45 °C in 0.05 M MgCl₂, 0.01 M Tris–HCl buffer, pH 7.5. The left panel shows the forward reaction (glucose → fructose) and the right panel shows the reverse reaction (fructose → glucose), each for varying initial substrate concentrations S_0 . Markers denote measured product concentrations, solid lines represent constrained spline fits, and dashed lines represent 95% confidence intervals. Error bars represent experimental error.

Table 3

Estimated kinetic parameters and model statistics for the isomerase reaction. Specific activities are reported relative to the enzyme concentration. Candidates 2–3 (Hill-type alternatives) are rejected by the $n \geq 1.6$ physical-admissibility criterion despite favorable AICc/R2; see main text for the full selection hierarchy.

	decision	V_{Glc}	V_{Fru}	$K_{\text{M,Glc}}$	$K_{\text{M,Fru}}$	K_{iu}	n	AIC	R ²
1	Selected	1.48 mM min ⁻¹ (0.375 U g ⁻¹)	2.08 mM min ⁻¹ (0.267 U g ⁻¹)	241 mM	341 mM	275 mM	-	-149	0.97
2	Rejected	6.89 mM min ⁻¹	1.30 mM min ⁻¹	389 mM	186 mM	1681 mM	0.80	-153	0.98
3	Rejected	6.94 mM min ⁻¹	1.96 mM min ⁻¹	360 mM	269 mM	-	0.79	-155	0.98

solve a one-shot optimal-subset-selection problem: a fixed, pre-enumerated library of basis functions is weighted via a global mixed-integer nonlinear program, and a single fitness metric derived from information theory determines the retained terms. GGSR instead incrementally grows candidate rate-law expressions through a multi-phase adaptive beam search: (i) a mechanistic grammar restricts candidate terms to physically admissible enzymatic building blocks from the outset, rather than admitting an unconstrained combination of generic basis functions; (ii) a perturbation-based identifiability ("sloppiness") gate is applied during the search itself, pruning branches with poorly constrained parameter combinations before they can propagate; (iii) a dedicated phenomenological shape-test phase evaluates Hill-type cooperativity separately, gated by an explicit physical-admissibility rule ($n \geq 1.6$) rather than by the information criterion alone; and (iv) a lineage-wide leave-one-reactor-out cross-validation step can downgrade an AICc-preferred model back to a simpler mechanistic ancestor if the latter generalizes better. This staged, gated search architecture is the main methodological distinction from ALAMO's single-shot subset selection, in addition to the enzyme-kinetics-specific vocabulary of the grammar itself.

3.3.2. Relation to SINDy

Sparse identification of nonlinear dynamics (SINDy) and related approaches identify a sparse combination of terms from a generic candidate-function library (typically low-order polynomials and trigonometric functions) that reconstructs the governing ordinary differential equations of a dynamical system, usually via L1-regularized (LASSO-type) or sequential thresholded least-squares regression. By construction, SINDy does not restrict candidate terms to mechanistically interpretable building blocks, and it targets the full system of state equations rather than a single algebraic constitutive relation. GGSR instead targets the rate law itself, constrains candidate terms a priori to physically admissible enzymatic building blocks via a grammar, and uses a staged search combining AICc-based ranking, an identifiability gate, and cross-validation, rather than a single sparsity-promoting regression step. SINDy is well suited to discovering unknown governing equations with comparatively few prior structural assumptions, provided a suitable term library is available. GGSR trades some of that generality for stronger mechanistic-interpretability guarantees, at the cost of requiring more domain-specific grammar design. The two approaches are complementary rather than competing: a SINDy-type sparse regression could, for example, be used to screen for unexpected additional terms outside the current enzymatic grammar, which could then be incorporated as new grammar building blocks.

3.3.3. Extension toward homogeneous and heterogeneous catalysis

The overall search and validation architecture of GGSR - the beam search, AICc-based ranking, the identifiability gate, the phenomenological shape-test phase, and lineage-wide cross-validation - is not intrinsically specific to enzyme kinetics and could in principle be reused for other saturation-kinetics formalisms. What is enzyme-specific in the present implementation is the vocabulary of grammar building blocks itself: Michaelis-Menten-type saturation terms, competitive/uncompetitive/mixed inhibition terms derived from rapid-equilibrium or quasi-steady-state assumptions for a single active site, and Hill-type cooperativity terms for multi-site binding. Extending the framework toward homogeneous or heterogeneous catalysis would require replacing this

vocabulary with mechanistically appropriate building blocks for the relevant catalytic cycle, for example, Langmuir-Hinshelwood-Hougen-Watson-type adsorption/desorption and surface-reaction terms for heterogeneous catalysis, or elementary-step rate expressions derived from a proposed catalytic cycle for homogeneous catalysis, while the overall search and validation architecture could largely be retained. As for multi-substrate enzymatic reactions, a key open challenge for such an extension is managing the combinatorial growth of admissible mechanistic terms as the number of catalytic-cycle steps and participating species increases.

4. Summary and outlook

In the present study, the concept of GGSR was successfully applied for the first time to the generation of a kinetic model for an enzymatic reaction. In addition to the development and validation of the workflow using synthetic datasets, a case study based on experimental data from an isomerase-catalyzed reaction was also presented.

At the current stage, the implemented grammar is limited to the simple case of a uni-uni reaction system. Future work must therefore extend the grammar toward multi-substrate reactions, as they are more common in nature and of industrial relevance (Cornish-Bowden, 2012). This will substantially increase the number of possible mechanistic interactions and inhibition terms, as all participating components and their potential interactions need to be considered. Due to this rapidly increasing combinatorial complexity, the integration of dynamic grammar-pruning heuristics or machine-learning-guided beam searches will become necessary to manage the computational effort while preserving the deterministic, physically constrained nature of the discovery process. Further necessary extensions include a rigorous identifiability analysis, starting with a normalized parameter correlation matrix derived directly from the covariance matrix already estimated via the fit Jacobian, and extending to a full profile-likelihood analysis, combined with a total quasi-steady-state (tQSSA) fitting layer, since sQSSA-based fitting is inherently limited here (Hong et al., 2022), a systematic noise-level sensitivity study, hard enforcement of the Haldane relationship, automated canonicalization of algebraically equivalent candidates, and a systematic procedure for activating grammar modules for less-characterized reaction systems. Nevertheless, the present work represents an important first step toward the future automated generation of mechanistically interpretable kinetic models using symbolic regression techniques.

Data availability statement

All code is available under <https://collaborating.tuhh.de/v-4/psi-public/tools4progresscurveanalysis>

CRediT authorship contribution statement

Thomas Waluga: Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis, Conceptualization. **Jorge Adrian Peña Ayala:** Writing – review & editing, Visualization, Software, Investigation, Formal analysis, Data curation. **Miriam Edel-Teichmann:** Writing – review & editing, Data curation. **Mirko Skiborowski:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Financial support from the Collaborative Research Center (CRC) “SMART Reactors for Future Process Engineering” funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project number 503850735 — at the Hamburg University of Technology (TUHH), is gratefully acknowledged.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.dche.2026.100346](https://doi.org/10.1016/j.dche.2026.100346).

References

- Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Stat. Surv.* 4 (none).
- Boodhoo, K., Flickinger, M.C., Woodley, J.M., Emanuelsson, E., 2022. Bioprocess intensification: a route to efficient and sustainable biocatalytic transformations for the future. *Chem. Eng. Process. - Process Intensif.* 172, 108793.
- Burnham, K.P., Anderson, D.R., 2004. Multimodel inference. *Sociol. Methods Res.* 33 (2), 261–304.
- Converti, A., Del Borghi, M., 1997. Simultaneous effects of immobilization and substrate protection on the thermodynamics of glucose isomerase activity and inactivation. *Enzyme Microb. Technol.* 21 (7), 511–517.
- Cornish-Bowden, A., 2012. *Fundamentals of Enzyme Kinetics*. Wiley-Blackwell, Weinheim, p. 498, 4th ed.
- Cozad, A., Sahinidis, N.V., 2018. A global MINLP approach to symbolic regression. *Math. Program.* 170 (1), 97–119.
- Cozad, A., Sahinidis, N.V., Miller, D.C., 2014. Learning surrogate models for simulation-based optimization. *AIChE J.* 60 (6), 2211–2227.
- Crochepierre L, Boudjeloud-Assala L, Barbesant V. A reinforcement learning approach to domain-knowledge inclusion using grammar guided symbolic regression 2022.
- Dehkordi, A.M., Tehrani, M.S., Safari, I., 2009. Kinetics of glucose isomerization to fructose by immobilized glucose isomerase (Sweetzyme IT). *Ind. Eng. Chem. Res.* 48 (7), 3271–3278.
- Goebel, R., Glaser, T., Skiborowski, M., 2020. Machine-based learning of predictive models in organic solvent nanofiltration: solute rejection in pure and mixed solvents. *Sep. Purif. Technol.* 248, 117046.
- Goebel, R., Skiborowski, M., 2020. Machine-based learning of predictive models in organic solvent nanofiltration: pure and mixed solvent flux. *Sep. Purif. Technol.* 237, 116363.
- Hoai NX, McKay RI, Essam D, Chau R. Solving the symbolic regression problem with tree-adjunct grammar guided genetic programming: the comparative results: 2026 1326–31.
- Hong, H., Choi, B., Kim, J.K., 2022. Beyond the Michaelis-Menten: bayesian inference for enzyme kinetic analysis. *Methods Mol. Biol.* 2385, 47–64.
- Hurvich, C.M., Tsai, C-L, 1989. Regression and time series model selection in small samples. *Biometrika* 76 (2), 297–307.
- Lemons, S., Linares López, C., Holte, R.C., Ruml, W., 2022. Beam search: faster and monotonic. *ICAPS* 32, 222–230.
- Lima, F.A.R., MG, de Moraes, Barreto, A.G., Secchi, A.R., Grover, M.A., de Souza, MB, 2025. Applications of machine learning for modeling and advanced control of crystallization processes: developments and perspectives. *Digit. Chem. Eng.* 14, 100208.
- Markus, M., Plesser, T., 1981. Progress curves in enzyme kinetics: design and analysis of experiments. editor. In: Endrenyi, L. (Ed.), *Kinetic Data Analysis*. Springer US, Boston, MA, pp. 317–339.
- McKay, B., Willis, M., Barton, G., 1997. Steady-state modelling of chemical process systems using genetic programming. *Comput. Chem. Eng.* 21 (9), 981–996.
- Nam, K.H., 2022. Glucose isomerase: functions, structures, and applications. *Appl. Sci.* 12 (1), 428.
- Narayanan, H., Cruz Bournazou, M.N., Guillén Gosálbez, G., Butté, A., 2022. Functional-hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. *Chem. Eng. J.* 430, 133032.
- Neumann, P., Cao, L., Russo, D., Vassiliadis, V.S., Lapkin, A.A., 2020. A new formulation for symbolic regression to identify physico-chemical laws from experimental data. *Chem. Eng. J.* 387, 123412.
- Reissmann, M., Fang, Y., Ooi, A.S.H., Sandberg, R.D., 2025. Constraining genetic symbolic regression via semantic backpropagation. *Genet. Program. Evolvable Mach.* 26 (1).
- Stroberg, W., Schnell, S., 2016. On the estimation errors of KM and V from time-course experiments using the Michaelis-Menten equation. *Biophys. Chem.* 219, 17–27.
- Waluga, T., Klein, M., Skiborowski, M., 2023. On the use of the adsorption energy distribution for the analysis of competing substrate kinetics. *Ind. Eng. Chem. Res.* 62 (1), 355–364.
- Waluga, T., Nagshi, P., 2026. On over-parameterisation and parameter estimation of enzyme kinetics. *Appl. Chem.* 6 (2), 25.
- Waluga, T., Ohde, D., Ansorge-Schumacher, M., Benfer, R., Bluemke, W., Briesen, H., et al., 2026. Bioreactive separation technology: a retrospective and perspective. *Chem. Eng. Process. - Process Intensif.* 224, 110787.
- Waluga, T., von Ziegner, F., Skiborowski, M., 2025. Analytical and numerical approaches to the analysis of progress curves: a methodological comparison. *Process Biochem.* 151, 1–13.
- Wang, Y., Wagner, N., Rondinelli, J.M., 2019. Symbolic regression in materials science. *MRS. Commun.* 9 (3), 793–805.
- Weiss, J.N., 1997. The Hill equation revisited: uses and misuses. *FASEB J.* 11 (11), 835–841.
- Wilson, Z.T., Sahinidis, N.V., 2017. The ALAMO approach to machine learning. *Comput. Chem. Eng.* 106, 785–795.