



Discovering Closed-Form Pharmacokinetic Relationships with Symbolic Regression to Improve Interpretability, Mechanistic Plausibility, and Extrapolation Beyond Black-Box Prediction

Nguyen Thanh Huy^{1*}, Pham Quang Minh¹, Le Thi Bich², Tran Van Nam³

¹Department of Symbolic Regression for Pharmacokinetics, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam.

²Department of Mechanistic Plausibility and Interpretability, Faculty of Pharmacy, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam.

³Department of Closed-Form Relationship Discovery, Faculty of Pharmacy, Hue University, Hue, Vietnam.

ABSTRACT

Machine-learning models can predict pharmacokinetic quantities and concentration–time profiles with increasing flexibility, yet prediction alone does not reveal whether the learned relation is dimensionally coherent, pharmacologically plausible, stable under perturbation, or transferable beyond the training domain. This methodological framework examines symbolic regression as a route to explicit pharmacokinetic equations while treating interpretability as a constrained scientific objective rather than a by-product of algebraic compactness. The proposed approach separates equation discovery from mechanistic attribution and requires candidate expressions to preserve measurement level, units, temporal structure, biological directionality where justified, and the distinction between molecular, physiological, and population-level variables. Mechanistic knowledge is introduced as an admissibility constraint or structural prior rather than as proof that the recovered expression is causal. The framework further proposes multi-criterion model selection based on predictive adequacy, parsimony, dimensional validity, stability, mechanistic plausibility, uncertainty, and performance under predefined extrapolation tests. Black-box predictors, population pharmacokinetic models, physiologically based pharmacokinetic models, and hybrid scientific-machine-learning approaches remain necessary comparators because symbolic transparency does not guarantee superior prediction or extrapolation. The central contribution is therefore a validation-oriented strategy for discovering closed-form relationships that can be inspected, challenged, and rejected. Its scope is methodological: recovered equations remain hypotheses conditioned on data quality, identifiability, chemical and population coverage, model specification, and external validation, and should not be interpreted as mechanistic laws or implementation-ready dosing rules without independent evidence.

Key Words: Symbolic regression, Pharmacokinetics, Scientific machine learning, Interpretability, Mechanistic plausibility, Extrapolation

eIJPPR 2026; 16(4):150-158

HOW TO CITE THIS ARTICLE: Huy NT, Minh PQ, Bich LT, Nam TV. Discovering Closed-Form Pharmacokinetic Relationships with Symbolic Regression to Improve Interpretability, Mechanistic Plausibility, and Extrapolation Beyond Black-Box Prediction. Int J Pharm Phytopharmacol Res. 2026;16(4):150-8. <https://doi.org/10.51847/SqnHNfsKpq>

INTRODUCTION

Machine learning has expanded the range of pharmacokinetic tasks that can be approached directly from concentration data, covariates, and other predictors,

but accuracy does not remove persistent problems of data quality, interpretation, validation, and generalization [1]. For pharmacokinetics, these problems are unusually consequential because a prediction may be numerically acceptable while obscuring which variables drive

Corresponding author: Nguyen Thanh Huy

Address: Department of Symbolic Regression for Pharmacokinetics, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam.

E-mail: huy.nguyen@hup.edu.vn

Received: 21 May 2026; **Revised:** 26 August 2026; **Accepted:** 27 August 2026

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



exposure, whether the relation respects units, or whether it is likely to fail when dose, population, chemical series, or sampling design changes. The methodological goal developed here is therefore narrower than “interpretable artificial intelligence”: it is the discovery of explicit mathematical relationships whose scientific admissibility can be inspected independently of their predictive error. Fit-for-purpose use of artificial intelligence in quantitative drug development requires transparent validation, attention to context of use, and explicit treatment of generalizability [2]. Those requirements motivate a distinction between an equation that is readable and an equation that is scientifically defensible. A compact expression can still exploit confounding, encode an accidental scaling relation, or extrapolate catastrophically. Conversely, a complex conventional pharmacometric model may remain scientifically useful because its parameters, compartments, and assumptions are tied to an established modeling purpose. Closed form is therefore proposed here as an additional representational advantage, not as a replacement criterion for pharmacometric validity.

Explainable-machine-learning workflows can accelerate covariate identification and help inspect functional relationships in population modeling, but their explanations remain conditional on the fitted predictive model [3]. This is different from searching directly for the mathematical form of a relationship. The distinction matters because post-hoc attribution asks what a model used, whereas symbolic equation discovery asks whether a compact candidate relation can itself serve as the model. Neither operation establishes mechanism, yet the latter makes algebraic structure, variable interactions, asymptotic behavior, and dimensional consistency directly examinable.

Symbolic regression searches explicit mathematical expressions and has been developed as an approach to interpretable scientific discovery, while remaining vulnerable to noise, search-space explosion, complexity choices, and validation failure [4]. We therefore propose a pharmacokinetic use of symbolic regression in which interpretability is conditional on four separable tests: the expression must be readable, dimensionally admissible, pharmacologically plausible within its stated context, and demonstrably stable under relevant validation and extrapolation challenges. These requirements define the methodological problem addressed in the sections that follow.

Symbolic regression and closed-form discovery

Unlike regression with a prespecified equation, symbolic regression searches over mathematical structures as well as their constants. Physics-inspired approaches show that decomposition, symmetry, separability, and other

structural properties can make recovery of compact equations tractable on benchmark physical relationships [5]. The relevance to pharmacokinetics is methodological rather than evidential: PK data do not inherit the clean generative structure of those benchmarks. Nevertheless, searching mathematical form creates a possibility unavailable to a fixed predictor—candidate expressions can be inspected for limiting behavior, monotonicity, parameter meaning, and dependence on variables before they are accepted.

Pharmacometrics now provides a closer precedent. Symbolic neural networks have been used to automate identification of human-readable covariate functions in a disease-progression model [6]. That result supports the feasibility of learning explicit functional relations inside a pharmacometric setting, but it does not establish closed-form PK law discovery, causal covariate effects, or superiority over conventional covariate modeling. We therefore treat symbolic components as generators of hypotheses about functional form. Their output becomes scientifically useful only after the recovered relation is re-evaluated against the data-generating context and competing formulations.

Physical and dimensional restrictions can also be incorporated into symbolic model discovery so that the search favors scientifically admissible expressions rather than merely low-error ones [7]. For PK, the proposed extension is to constrain the operator library and candidate construction by units, positivity where biologically required, and known structural relationships when justified. These constraints should prune impossible equations without forcing the search toward a predetermined mechanistic answer. The desired output is thus not the “truest-looking” equation, but a small set of explicit candidates that survive empirical, dimensional, and pharmacological challenge.

Pharmacokinetic data structure and dimensional constraints

Pharmacokinetic data are not interchangeable scalar labels. Concentration–time observations are indexed by time, dose, sampling design, subject, formulation, and often additional covariates; derived quantities such as exposure summaries depend on how the underlying profile is sampled and analyzed. Deep-learning approaches to noncompartmental analysis illustrate that temporal sampling and the method used to reconstruct or summarize a profile are part of the modeling problem itself [8]. Symbolic discovery should therefore preserve whether a target is an observed concentration, a derived summary, or an estimated model parameter rather than pooling these as equivalent numerical outcomes.

Artificial neural networks can predict concentration–time trajectories without exposing a closed-form pharmacokinetic relationship [9]. This establishes an important comparator: a flexible predictor may fit temporal behavior while providing no algebraic object to test for units or asymptotic plausibility. The proposed symbolic workflow must not gain apparent interpretability by flattening the time dimension. Time, dose, concentration, and subject-level variables should enter with declared units and provenance, and repeated measurements should remain associated with their experimental or clinical hierarchy. Otherwise, a compact equation can be interpretable syntactically while representing a statistically invalid data transformation.

The same level-of-analysis problem arises when molecular structure is linked to in vivo outcomes. Machine learning from chemical structure has predicted PK parameters and exposure curves in a defined rat data set, while the resulting relationships remain bounded by species and chemical space [10]. Accordingly, a symbolic equation connecting molecular descriptors to clearance is not equivalent to a physiological clearance equation, even when both predict the same endpoint. We propose that every candidate relation carry a level tag—molecular, experimental, individual, population, or physiological—and that dimensional closure be evaluated within that level. Cross-level equations may be valuable, but their mechanistic interpretation must remain explicitly limited.

Embedding mechanistic priors and units

Hybrid scientific-machine-learning approaches demonstrate that known pharmacokinetic structure can be retained while learned components represent unresolved dynamics [11]. The evidence is a precedent for architectural separation, not proof that any learned component corresponds to a missing biological

mechanism. For symbolic regression, the analogous proposal is to distinguish fixed structure from discoverable structure. Known mass-balance relationships, dosing inputs, or compartmental connections may be held fixed when justified, while symbolic search is restricted to the unresolved function. This prevents the algorithm from spending search capacity rediscovering established identities and makes any new relation easier to localize and challenge.

Physics-informed machine learning provides a broader vocabulary for incorporating scientific knowledge through governing equations, constraints, architectures, or loss functions [12]. In PK equation discovery, units can serve as a particularly strict prior because dimensional inconsistency is disqualifying even when predictive error is low. Other priors—positivity, boundedness, monotonicity, or conservation—should be imposed only when the biological context warrants them. Incorrect priors can suppress valid behavior, so constraint provenance should be recorded and sensitivity to relaxing each prior should form part of validation.

PBPK modeling provides a complementary example in which explicit physiological structure can be combined with machine learning for parameter estimation or component emulation [13]. This modularity suggests that symbolic regression need not compete with mechanistic PK models at the whole-model level. It can instead operate at defined interfaces: learning a covariate function, an empirical scaling term, or an unresolved input–output relation while leaving validated physiological structure intact. The resulting equation would remain an empirical component until independently supported.

These constraints are consolidated in **Table 1** as assumptions that must remain visible before a discovered closed-form relationship is interpreted scientifically.

Table 1. Assumptions conditioning scientific interpretation of a discovered pharmacokinetic equation

Assumption	Risk if violated	Proposed methodological remedy	Evidence status
Temporal observations and derived PK summaries remain distinguishable	A compact relation may encode sampling or summarization artifacts	Preserve time, dose, sampling, and target provenance before symbolic search	Evidence-grounded; remedy proposed
Predictors and outcomes occupy a declared level of analysis	Molecular association may be misread as physiological mechanism	Tag molecular, experimental, individual, population, and physiological variables explicitly	Evidence-grounded; taxonomy proposed
Candidate expressions are dimensionally closed	Low-error equations may be physically meaningless	Reject unit-inconsistent operators and transformations during search	Evidence-grounded; implementation proposed
Mechanistic priors are valid for the intended context	Incorrect constraints can suppress admissible relationships	Record each prior, test relaxed alternatives, and separate fixed from discoverable structure	Evidence-grounded; procedure proposed
Mechanistic modules and empirical relations remain distinguishable	Learned terms may acquire unjustified biological meaning	Confine symbolic discovery to defined interfaces and retain empirical labels until independently validated	Evidence-grounded; framework synthesis proposed

Discovery of relationships across absorption, distribution, and clearance

The value of symbolic regression is greatest when the target relationship is specified before equation search

begins. Human pharmacokinetic endpoints can be predicted from molecular and early experimental information, and uncertainty-aware models can support drug-design decisions within a defined chemical space [14]. That evidence supports learnability, not the existence of a universal analytical law linking structure to exposure. We therefore propose that symbolic discovery distinguish endpoint equations from cross-endpoint relations. An equation predicting clearance from molecular or in vitro variables should be labelled an empirical clearance relation; an equation linking clearance, volume, and exposure through an established identity occupies a different mechanistic level.

Multi-property and multi-species modeling can exploit statistical dependence among pharmacokinetic tasks while representing predictive uncertainty [15]. For symbolic regression, however, shared predictive signal should not automatically be converted into a shared physiological equation. Species, route, formulation, and measurement context may change both coefficients and functional form. Cross-species symbolic candidates should therefore be retained only when the transfer question is explicit and the equation is challenged against species-held-out data or independent physiological expectations.

Concentration-time prediction adds a third relationship class: trajectory structure. Recent work shows that machine-learning models can predict plasma concentration-time profiles under defined validation settings [16]. A symbolic trajectory model should preserve time and dosing as explicit variables or act inside a declared dynamical structure rather than reducing an entire profile to a single unconstrained scalar mapping. The proposed discovery space thus contains three separable targets—endpoint relations, cross-endpoint relations, and trajectory relations. Absorption, distribution, and clearance may contribute to all three, but the manuscript does not assume that one discovered expression can identify their underlying mechanisms. Each candidate remains conditional on the variables observed, the scale analyzed, and the experimental or clinical context represented.

Model selection for parsimony and plausibility

Once symbolic search produces multiple equations with similar predictive performance, model selection becomes the central scientific step. Recent symbolic-regression methodology explicitly treats parsimony as an optimization objective rather than an informal preference [17]. In pharmacokinetics, however, minimum algebraic complexity should not be equated with maximum credibility. A very short expression may omit a relevant process, exploit correlation among covariates, or behave implausibly outside the observed range. Parsimony is therefore proposed as one criterion within a constrained selection problem, not as the final objective.

Population-pharmacokinetic model-selection methods differ substantially in assumptions, workflow, and susceptibility to overfitting, and no universally optimal selector is established [18]. Symbolic regression inherits the same problem in amplified form because it searches both structure and coefficients. Selection should therefore separate predictive adequacy from scientific admissibility. First, candidate equations that fail prespecified predictive checks are removed. Second, dimensionally invalid or contextually impossible expressions are rejected. Third, remaining equations are compared for complexity, stability, and dependence on weakly supported terms. Only then should pharmacological plausibility be considered.

Domain knowledge can be incorporated directly into symbolic-regression search and used to restrict candidate functions toward scientifically plausible relations [19]. Yet plausibility can itself become a source of confirmation bias. An equation should not be preferred merely because it resembles a familiar one-compartment or allometric relationship. The proposed criterion is stronger: a candidate must state which prior it satisfies, whether that prior was enforced or merely observed after fitting, and whether alternative admissible expressions fit comparably well. Mechanistic familiarity without discriminating evidence is insufficient.

This produces a workflow in which equation discovery and equation acceptance are deliberately separated. **Table 2** summarizes the manuscript-derived analytic stages used to move from heterogeneous PK observations to a ranked, challengeable candidate relation.

Table 2. Proposed analytic workflow for ranking symbolic pharmacokinetic relationships after equation generation

Workflow stage	Question asked	Decision signal	Main scientific risk
Context registration	What was measured, at what level, with which units and sampling structure?	Complete provenance and dimensional specification	Mixing endpoints, scales, or derived quantities
Candidate generation	Which explicit forms fit within the allowed operator and prior set?	Reproducible search yielding multiple viable expressions	Search-library bias or premature convergence
Admissibility screening	Is the expression dimensionally and contextually possible?	Unit closure and justified structural constraints	Accepting numerically accurate but impossible equations
Predictive screening	Does the candidate retain adequate fit on held-out data?	Prespecified external or resampled predictive checks	Selecting elegance despite poor prediction
Parsimony assessment	Can similar performance be achieved with fewer unsupported terms?	Stable accuracy-complexity trade-off	Treating shortest form as mechanistic truth

Plausibility review	What prior does the relation satisfy, and what alternatives remain?	Explicit rationale plus competing admissible forms	Confirmation bias toward familiar PK equations
----------------------------	---	--	--

note: All stages constitute the proposed framework. Evidence supporting the separation of parsimony, model selection, and domain-informed plausibility is provided in the preceding section; no numerical acceptance threshold is implied.

Extrapolation beyond the training domain

Extrapolation is the strongest test of whether an interpretable equation captures a transferable relationship rather than a local regularity. A recent first-in-human comparison of standard PBPK, high-throughput PBPK, and machine learning showed that these approaches can have complementary strengths under different information settings [20]. The relevant lesson is not that mechanistic models invariably extrapolate better, but that translational prediction changes the task: assumptions that are harmless during interpolation can become dominant when human physiology, dose, or chemical properties depart from the training distribution.

Chemical-space locality creates a parallel problem. Local and global machine-learning models for ADME properties can behave differently across chemical neighborhoods, demonstrating that performance depends on where a query compound lies relative to the available data [21]. A symbolic equation is not exempt from this dependence.

Closed form makes its asymptotic behavior visible, but visibility does not make that behavior correct. An algebraically smooth expression may become biologically absurd when molecular descriptors, doses, covariates, or physiological variables move beyond the region that determined its coefficients.

We therefore propose that every symbolic PK relationship be accompanied by an extrapolation profile defined before interpretation. The profile specifies which shifts were actually tested: chemical series, species, population, dose or regimen, route or formulation, sampling pattern, and relevant physiological state. Interpolation performance and extrapolation performance should be reported separately. Failure under one shift need not invalidate the equation for another context, but it narrows its admissible claim.

Figure 1 synthesizes this argument by placing an explicit candidate equation inside a distributed field of data-level, constraint, and transfer challenges rather than treating closed form as the endpoint of discovery.

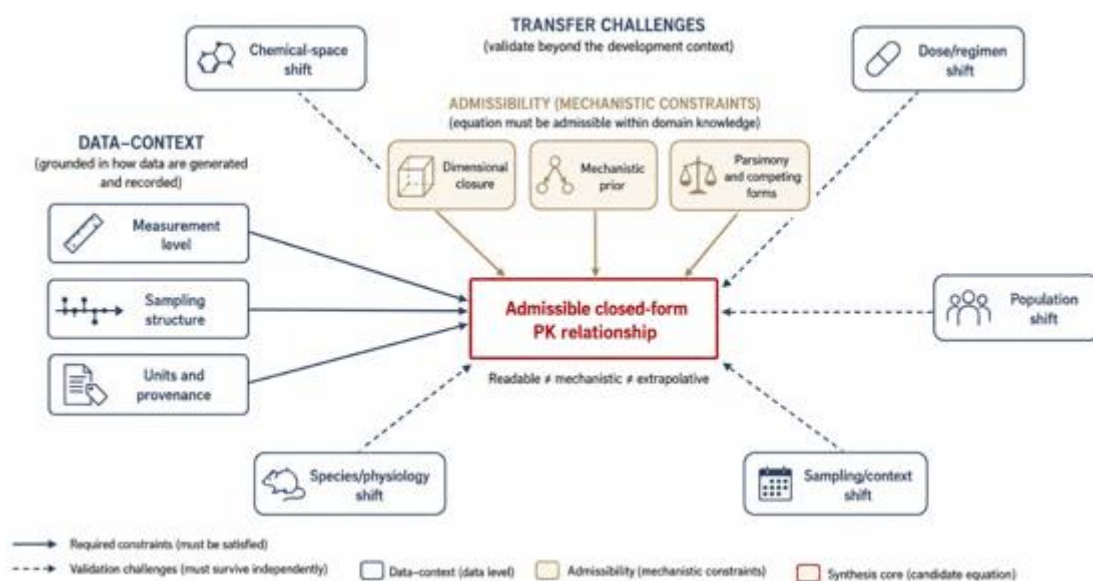


Figure 1. Closed-form pharmacokinetic discovery as a constraint-and-transfer problem

Comparison with black-box predictors

A meaningful comparison with black-box prediction must avoid making interpretability the only criterion on which symbolic regression can win. Direct comparisons among scientific machine learning, population pharmacokinetics, and classical machine-learning methods show that relative performance can depend on the drug, data regime, and modeling assumptions [22]. Symbolic regression should therefore be benchmarked on the same inputs, targets, splits, and missing-data rules as competing approaches. If

a black-box model predicts materially better, that difference remains scientifically relevant even when its internal mapping is harder to inspect.

Recent comparative work on plasma concentration-time prediction likewise reinforces that model ranking depends on representation, endpoint, and validation design [23]. For symbolic methods, conventional predictive metrics should be complemented by equation-specific criteria: complexity, dimensional validity, coefficient and structural stability across resampling, and behavior under

predefined shifts. These additions do not compensate for poor prediction; they test a different proposition—whether a model that is sufficiently predictive also offers a more inspectable scientific representation.

Across data-driven PK studies, model rankings depend on model class, endpoint, and validation regime rather than a universal hierarchy [9, 22, 23]. The proposed framework therefore treats black-box models as active comparators rather than straw alternatives. They can reveal when an explicit equation sacrifices excessive predictive

information, identify nonlinearities that the symbolic operator set cannot capture, and provide an upper or lower benchmark for selected tasks. Conversely, symbolic models provide direct access to algebraic form and limiting behavior that black-box systems may not expose.

Table 3 converts this comparison into an applicability-domain matrix: the relevant question is not which model family is intrinsically superior, but which transfer claim is supported after each model has been challenged on the same domain boundary.

Table 3. Applicability-domain matrix for interpreting symbolic and non-symbolic pharmacokinetic predictions

Transfer domain	What must be tested	What closed form makes inspectable	What remains unproven
New chemical series	Performance away from training chemistry	Descriptor dependence and asymptotic form	Biological transfer to unfamiliar chemistry
New dose or regimen	Prediction under altered input pattern	Dose/time dependence and limiting behavior	Linearity or unchanged kinetics outside tested range
New population	Performance across covariate distributions	Explicit covariate functional form	Transportability across unrepresented physiology
New species or physiology	Cross-system prediction	Where species or physiological terms enter	Conservation of mechanism across systems
New sampling context	Robustness to schedule, sparsity, or observation pattern	Dependence on time and observed variables	Immunity to measurement or reconstruction artifacts
Matched in-domain benchmark	Equal-input comparison with alternative models	Accuracy-complexity trade-off	Superiority beyond the benchmarked task

note: The matrix is a proposed interpretation aid derived from the preceding extrapolation and model-comparison arguments. “Inspectable” does not mean validated, causal, or transferable.

The mechanistically constrained symbolic regression framework

Hybrid pharmacometric-machine-learning guidance emphasizes transparent separation between mechanistic and learned components and fit-for-purpose validation [24]. Building on that principle, the framework proposed here treats symbolic regression as a bounded discovery module rather than an autonomous replacement for pharmacometrics. Its input is not simply a matrix of predictors and outcomes; it is a registered modeling context containing variable definitions, units, level of analysis, sampling structure, intended prediction target, and the transfer question the equation will eventually face. A pre-search analysis plan should also define the comparator models, validation split, permissible transformations, and intended interpretation level. Fixing these choices before examining preferred equations reduces the risk that constraints, operator sets, or complexity penalties are tuned retrospectively to make a particular symbolic form appear scientifically compelling. The second component is a constraint registry. Each restriction imposed on the search—dimensional homogeneity, positivity, allowable operators, fixed structural identities, monotonicity, conservation, or parameter bounds—is declared with its scientific rationale.

Integration of machine learning with PBPK and QSAR provides precedent for modular modeling in which data-driven and mechanistic components operate at explicit interfaces [25]. The proposed symbolic module follows the same principle: a known PK relationship can remain fixed while an unresolved covariate, scaling, or input-output function is searched symbolically. This makes the learned component localizable and prevents a successful fit from being mistaken for wholesale reconstruction of physiology.

Candidate generation is followed by three gates. The admissibility gate removes dimensionally invalid or contextually impossible expressions. The comparative gate evaluates predictive adequacy against conventional pharmacometric, black-box, or hybrid alternatives. The transfer gate challenges surviving candidates under prespecified shifts. Continuous-time neural models are relevant comparators because they can learn flexible longitudinal PK/PD dynamics without yielding simple symbolic expressions [26]. Their inclusion prevents the evaluation from conflating continuous-time flexibility with symbolic transparency.

Finally, interpretation is assigned only after validation. We propose three labels: descriptive equation for a reproducible empirical mapping within its data domain;

structurally informed equation when justified priors constrain its form; and mechanistic candidate when the relation additionally survives targeted experiments or independent mechanistic tests capable of falsifying it. These labels are deliberately asymmetric: increased readability can move an equation from opaque to inspectable, but only independent evidence can strengthen the mechanistic claim. The framework therefore uses symbolic regression to generate scientifically examinable hypotheses while preserving the option to reject them.

Validation, identifiability, and failure modes

Predictive success is insufficient when coefficients or functional forms are given pharmacological meaning. Hybrid neural ordinary differential-equation studies demonstrate that parameter identifiability and robustness can remain separate problems even when observed dynamics are fitted well [27]. Symbolic regression adds structural identifiability questions: distinct expressions may approximate the same observations, algebraically equivalent forms may appear different, and correlated predictors may exchange roles across resampled data. A reported equation should therefore be accompanied by evidence that its structure, not only its predictions, is reasonably stable. Algebraically equivalent expressions should also be canonicalized where possible before stability is judged; otherwise superficial changes in notation can be mistaken for structural disagreement, while genuinely different equations with similar predictions can be overlooked.

Population implementation introduces additional uncertainty. Low-dimensional neural ODEs can incorporate inter-individual variability in established pharmacometric software, while estimation stability and variability modeling remain explicit considerations [28]. For symbolic PK relations, population validation should similarly distinguish fixed functional structure from subject-level variability and should test whether the discovered form persists when influential individuals, covariate ranges, or sampling patterns change.

Hybrid dynamical models require both predictive checks and parameter/structural identifiability assessments [11, 27, 28]. We extend this principle into a proposed validation hierarchy. First, simulation experiments with known generating equations can test recovery failure and operator-library bias. Second, repeated search across resamples can quantify structural recurrence and coefficient stability. Third, external data should challenge the equation within its stated domain. Fourth, prespecified extrapolation tests should probe the exact transfer claim. Finally, any mechanistic interpretation should face experiments or perturbations that distinguish the candidate relation from plausible alternatives.

Failure is informative at each level. Persistent predictive error rejects the candidate as a predictor; dimensional inconsistency rejects it as scientifically admissible; unstable structure weakens interpretability; non-identifiability prevents strong parameter meaning; and extrapolation failure narrows the applicability domain. None of these outcomes requires abandoning symbolic regression. They define what the recovered equation is allowed to claim.

Limitations

Symbolic regression does not remove uncertainty by converting predictions into equations. Uncertainty estimates can themselves degrade under distribution shift, so calibration within the training domain cannot be assumed to remain valid during extrapolation [29]. Moreover, the selected operator library, complexity penalty, priors, data preprocessing, and search algorithm all shape which expressions can be discovered. A reproducible equation may therefore reflect the modeling language offered to the algorithm as much as an underlying biological relationship.

Applicability-domain analysis is essential when molecular or learned predictors are transported beyond supported chemical space [30]. The same principle applies to dose, population, species, formulation, and sampling domains, although the appropriate domain representation will differ. Observational PK data also contain collinearity, sparse sampling, measurement error, treatment selection, and latent physiological variation that can support several competing equations. Sparse coverage may also make a seemingly simple domain boundary unstable when the observed covariate distribution is narrow.

Uncertainty, dataset shift, and applicability-domain limitations jointly constrain transfer of learned PK relationships [21, 29, 30]. Consequently, the proposed framework cannot guarantee that symbolic models will outperform black-box predictors, recover true mechanisms, or extrapolate better than PBPK or population models. Its contribution is a more explicit procedure for exposing these failures. Prospective evaluation on diverse PK tasks is required before any framework-specific performance or decision advantage can be established.

CONCLUSION

Closed-form pharmacokinetic equations are valuable only when their interpretability survives scientific scrutiny. Symbolic regression creates an opportunity to search explicit functional relationships rather than accepting either fixed equations or opaque predictors, but algebraic readability alone does not confer mechanistic meaning or

extrapolative validity. The framework proposed here therefore couples symbolic discovery to data-level declaration, dimensional constraints, explicit mechanistic priors, multi-criterion model selection, matched comparison with alternative predictors, prespecified extrapolation challenges, and identifiability-aware validation.

This architecture reframes symbolic regression as a hypothesis-generation and model-discrimination tool within pharmacometrics. A candidate expression may remain useful as a descriptive relation even when it cannot support a mechanistic interpretation; failure under a particular transfer test may narrow its domain without invalidating interpolation. Conversely, mechanistic language should be strengthened only when independent evidence distinguishes the relation from alternative explanations. The practical objective is not to replace black-box, population, or physiologically based pharmacokinetic models, but to add a disciplined route for discovering equations that are inspectable enough to be challenged. Prospective, cross-domain evaluation will be needed to determine when that added transparency translates into better scientific inference or more reliable extrapolation.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Huang S, Xu Q, Yang G, Ding J, Pei Q. Machine learning for prediction of drug concentrations: Application and challenges. *Clin Pharmacol Ther.* 2025;117(5):1236-47. doi:10.1002/cpt.3577
- [2] Terranova N, Renard D, Shahin MH, Menon S, Cao Y, Hop CECA, et al. Artificial intelligence for quantitative modeling in drug discovery and development: An Innovation and Quality Consortium perspective on use cases and best practices. *Clin Pharmacol Ther.* 2024;115(4):658-72. doi:10.1002/cpt.3053
- [3] Brooks L, Harun R, Jin JY, Lu J. Shap-Cov: An explainable machine learning based workflow for rapid covariate identification in population modeling. *CPT Pharmacometrics Syst Pharmacol.* 2025;14(8):1322-31. doi:10.1002/psp4.70044
- [4] Makke N, Chawla S. Interpretable scientific discovery with symbolic regression: A review. *Artif Intell Rev.* 2024;57(1):2. doi:10.1007/s10462-023-10622-0
- [5] Udrescu SM, Tegmark M. AI Feynman: A physics-inspired method for symbolic regression. *Sci Adv.* 2020;6(16):eaay2631. doi:10.1126/sciadv.aay2631
- [6] Sundell J, Wahlquist Y, Kjellsson MC, Karlsson MO, Soltesz K. Learning covariate relations in disease progression models using symbolic neural networks. *CPT Pharmacometrics Syst Pharmacol.* 2026;15(3):e70214. doi:10.1002/psp4.70214
- [7] Neumann P, Cao L, Russo D, Vassiliadis VS, Lapkin AA. A new formulation for symbolic regression to identify physico-chemical laws from experimental data. *Chem Eng J.* 2020;387:123412. doi:10.1016/j.cej.2019.123412
- [8] Liu G, Brooks L, Canty J, Lu D, Jin JY, Lu J. Deep-NCA: A deep learning methodology for performing noncompartmental analysis of pharmacokinetic data. *CPT Pharmacometrics Syst Pharmacol.* 2024;13(5):870-9. doi:10.1002/psp4.13124
- [9] Bräm DS, Parrott N, Hutchinson L, Steiert B. Introduction of an artificial neural network-based method for concentration-time predictions. *CPT Pharmacometrics Syst Pharmacol.* 2022;11(6):745-54. doi:10.1002/psp4.12786
- [10] Obrezanova O, Martinsson A, Whitehead T, Mahmoud S, Bender A, Miljković F, et al. Prediction of in vivo pharmacokinetic parameters and time-exposure curves in rats using machine learning from the chemical structure. *Mol Pharm.* 2022;19(5):1488-504. doi:10.1021/acs.molpharmaceut.2c00027
- [11] Valderrama D, Ponce-Bobadilla AV, Mensing S, Fröhlich H, Stodtmann S. Integrating machine learning with pharmacokinetic models: Benefits of scientific machine learning in adding neural networks components to existing PK models. *CPT Pharmacometrics Syst Pharmacol.* 2024;13(1):41-53. doi:10.1002/psp4.13054
- [12] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. *Nat Rev Phys.* 2021;3(6):422-40. doi:10.1038/s42254-021-00314-5
- [13] Chou WC, Lin Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. *Toxicol Sci.* 2023;191(1):1-14. doi:10.1093/toxsci/kfac101
- [14] Komissarov L, Manevski N, Groebke Zbinden K, Schindler T, Zitnik M, Sach-Peltason L. Actionable predictions of human pharmacokinetics at the drug design stage. *Mol Pharm.* 2024;21(9):4356-71. doi:10.1021/acs.molpharmaceut.4c00311

- [15] Zhang Y, Xie Z, Xiao F, Yu J, Fan Z, Sun S, et al. Prediction of multi-pharmacokinetics property in multi-species: Bayesian neural network stacking model with uncertainty. *Mol Pharm.* 2024;21(12):6177-92. doi:10.1021/acs.molpharmaceut.4c00406
- [16] Iwata H, Kageyama M, Handa K. Machine learning prediction and validation of plasma concentration-time profiles. *Mol Pharm.* 2025;22(6):2976-84. doi:10.1021/acs.molpharmaceut.4c01431
- [17] Muthyala MR, Sorourifar F, Peng Y, Paulson JA. SyMANTIC: An efficient symbolic regression method for interpretable and parsimonious model discovery in science and beyond. *Ind Eng Chem Res.* 2025;64(6):3354-69. doi:10.1021/acs.iecr.4c03503
- [18] Karlsen M, Khier S, Fabre D, Marchionni D, Azé J, Bringay S, et al. Covariate model selection approaches for population pharmacokinetics: A systematic review of existing methods, from SCM to AI. *CPT Pharmacometrics Syst Pharmacol.* 2025;14(4):621-39. doi:10.1002/psp4.13306
- [19] Keren LS, Liberzon A, Lazebnik T. A computational framework for physics-informed symbolic regression with straightforward integration of domain knowledge. *Sci Rep.* 2023;13(1):1249. doi:10.1038/s41598-023-28328-2
- [20] Käser S, Bassani D, Parrott NJ, Umehara K, Delporte ML, David-Pierson P. Predicting first-in-human pharmacokinetics: Comparative evaluation of standard PBPK, high-throughput PBPK, and machine learning. *Mol Pharm.* 2026;23(8):4251-62. doi:10.1021/acs.molpharmaceut.6c00429
- [21] Di Lascio E, Gerebtzoff G, Rodríguez-Pérez R. Systematic evaluation of local and global machine learning models for the prediction of ADME properties. *Mol Pharm.* 2023;20(3):1758-67. doi:10.1021/acs.molpharmaceut.2c00962
- [22] Valderrama D, Teplytska O, Koltermann LM, Trunz E, Schmulenson E, Fritsch A, et al. Comparing scientific machine learning with population pharmacokinetic and classical machine learning approaches for prediction of drug concentrations. *CPT Pharmacometrics Syst Pharmacol.* 2025;14(4):759-69. doi:10.1002/psp4.13313
- [23] Jost F, Giegerich C, Grebner C, Matter H, Cordes H. Machine learning meets pharmacokinetics: A comparative analysis of predictive models for plasma concentration-time profiles. *CPT Pharmacometrics Syst Pharmacol.* 2026;15(5):e70226. doi:10.1002/psp4.70226
- [24] Fochesato A, Brooks L, Bazgir O, Pierrillas PB, Jamois C, Lu J, et al. Building hybrid pharmacometric-machine learning models in oncology drug development: Current state and recommendations. *CPT Pharmacometrics Syst Pharmacol.* 2026;15(1):e70113. doi:10.1002/psp4.70113
- [25] Chen X, Lin Z. Integration of machine learning with PBPK and QSAR modeling approaches to facilitate drug discovery and development. *CPT Pharmacometrics Syst Pharmacol.* 2026;15(4):e70228. doi:10.1002/psp4.70228
- [26] Wu Z, Luo P, Chen R, Liu Y, Jian W, Zhou T. Neural controlled differential equation and its application in pharmacokinetics and pharmacodynamics. *CPT Pharmacometrics Syst Pharmacol.* 2026;15(1):e70146. doi:10.1002/psp4.70146
- [27] Giampiccolo S, Reali F, Fochesato A, Iacca G, Marchetti L. Robust parameter estimation and identifiability analysis with hybrid neural ordinary differential equations in computational biology. *NPJ Syst Biol Appl.* 2024;10(1):139. doi:10.1038/s41540-024-00460-3
- [28] Bräm DS, Steiert B, Pfister M, Steffens B, Koch G. Low-dimensional neural ordinary differential equations accounting for inter-individual variability implemented in Monolix and NONMEM. *CPT Pharmacometrics Syst Pharmacol.* 2025;14(1):5-16. doi:10.1002/psp4.13265
- [29] Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. *J Chem Inf Model.* 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
- [30] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. *Chem Soc Rev.* 2020;49(11):3525-64. doi:10.1039/d0cs00098a