Sisyphus: A Topology-Compiled Physiologically Based Pharmacokinetic Platform with Structure-Only Input and Bayesian Parameter Refinement

Jae Min Yoon
Independent Researcher
Correspondence: jaemin6013@gmail.com

Abstract

Physiologically based pharmacokinetic (PBPK) platforms have achieved regulatory acceptance for drug–drug interaction and pediatric dose optimization, yet the dominant implementations require experimentally measured absorption, distribution, metabolism, and excretion (ADME) parameters as input, limiting their utility in early drug discovery when only molecular structure is available. Machine learning approaches that predict pharmacokinetic endpoints directly from structure address this constraint but lack mechanistic interpretability and cannot extrapolate to untested scenarios such as drug–drug interactions or alternative dosing regimens. Sisyphus, an open-source PBPK platform that represents the human body as a typed directed multigraph, automatically compiles ordinary differential equation (ODE) systems from graph topology, and accepts a simplified molecular-input line-entry system (SMILES) string as the sole molecular input, is presented here. Five quantitative structure–activity relationship models predict ADME parameters from molecular structure; these are propagated through a whole-body PBPK model with native Monte Carlo uncertainty quantification, and a four-track meta-learner dynamically weights mechanistic-engine, direct machine-learning, clearance-regression, and volume-of-distribution-based predictions on a per-compound basis.

On a held-out set of 107 drugs (healthy volunteers, single oral dose, immediate-release, fasted state), the platform achieved a population-level geometric mean absolute fold error (AAFE) of 2.698 for maximum plasma concentration (Cmax), with 46.7% of predictions within 2-fold of observed values (bootstrap 95% confidence interval [CI] [2.32, 3.17]; in-domain subset AAFE 2.760 [2.30, 3.34], N = 79, excluding applicability-domain-flagged compounds). All reported benchmarks were generated from a fresh clone of the public repository with no proprietary or environment-specific data artifacts loaded, so the reported accuracy is structure-only and independently reproducible. Forty-one enumerated optimization attempts—spanning alternative molecular representations, training-data expansion, in vitro–in vivo extrapolation (IVIVE) chain bypass, bond-dissociation-energy reactivity features, residual-correction models, post-hoc meta-learners, and orthogonal prediction sources—failed to improve downstream accuracy, while two strategies improved it and are retained in the pipeline. These results indicate that the accuracy ceiling originates from intrinsic clearance prediction quality (coefficient of determination $R^2 \approx 0.24$ on the Therapeutics Data Commons Hepatocyte_AZ dataset under scaffold-split cross-validation) rather than from model architecture or data scale.

Substituting experimentally measured ADME parameters for predicted values in a 10-compound subset produced a modest, mixed improvement in engine AAFE (2.81 to 2.69; 6 of 10 compounds improved), indicating that ADME input quality is one contributor to prediction error rather than establishing that the engine is fully input-limited. The platform additionally supports multi-dose regimen simulation, Bayesian posterior parameter refinement that reduced prediction uncertainty by an average of 78% from a single observation across five compounds, drug–drug interaction prediction, and pharmacodynamic modeling.

Prospective validation on 28 small-molecule new molecular entities (NMEs) approved in 2024–2025 yielded AAFE 3.21 (N = 28; in-domain 3.20, N = 16), modestly worse than the retrospective holdout; the gap was localized to systematic under-prediction of oral bioavailability for novel chemotypes rather than to clearance. Sisyphus addresses a capability not met by existing tools: a single platform spanning structure-only screening through observation-informed parameter refinement, in which mechanistic parameters retain physical meaning throughout the pipeline.

Introduction

Physiologically based pharmacokinetic modeling has become a central methodology in drug development and regulatory decision-making [1,2]. The dominant commercial platforms—Simcyp [15], GastroPlus [16], and PK-Sim [14]—integrate in vitro ADME data with physiological parameters to simulate drug disposition in virtual populations. Regulatory acceptance is evidenced by more than 40 PBPK-based label recommendations approved by the U.S. Food and Drug Administration (FDA) between 2018 and 2024, predominantly for drug–drug interaction and pediatric dose optimization [3].

Despite this success, a structural limitation persists: each of these platforms requires experimentally measured ADME parameters as input. Intrinsic clearance (CLint), fraction unbound in plasma (fu,p), effective permeability (Peff), and related in vitro measurements are prerequisite inputs. In early drug discovery, when thousands of candidate structures are evaluated computationally before any are synthesized, these measurements do not exist. The result is a gap between structure-based virtual screening and mechanism-based pharmacokinetic simulation that current tools do not bridge.

Machine learning approaches have partially addressed this gap [11,12]. Direct prediction models map molecular descriptors to pharmacokinetic endpoints (Cmax, area under the curve, clearance) without physiological modeling. However, these models are interpolative: they cannot mechanistically extrapolate to drug–drug interactions, special populations, or alternative dosing regimens because they lack the physiological structure that enables such predictions. They also cannot incorporate new experimental observations to refine individual predictions, as they lack updatable mechanistic parameters.

This work addresses a specific research question: can a mechanistic, whole-body PBPK simulation be driven entirely from molecular structure—with no measured ADME input—while retaining the physical interpretability and extensibility that distinguish PBPK from direct machine learning, and what accuracy ceiling does such a structure-only configuration impose? To investigate this question, Sisyphus was developed to accept a single SMILES string as input, predict all required ADME parameters from molecular structure using trained quantitative structure–activity relationship (QSAR) models, and propagate these predictions through a whole-body PBPK model with full uncertainty quantification. The same mechanistic framework that enables structure-only screening also supports progressive incorporation of measured data as it becomes available during development, and ultimately individual-level parameter refinement through Bayesian posterior estimation. The central architectural decision—representing the body as a typed directed multigraph from which ODE systems are compiled automatically—enables extensibility to new organs, routes of administration, and patient populations without modification of the simulation engine.

This paper describes the design and validation of Sisyphus, reports accuracy benchmarks against a curated holdout set of 107 drugs, presents a systematic evaluation of optimization strategies including both positive and negative results, demonstrates the platform’s multi-dose and Bayesian-refinement

capabilities, and provides evidence that measured ADME inputs substantially improve engine accuracy. All reported accuracy figures were generated from a fresh clone of the public repository, with no proprietary or environment-specific data artifacts loaded (see Methods and Limitations).

Methods

Topology-Driven ODE Architecture

The human body is represented as a typed directed multigraph G = (V, E), where vertices V represent physiological compartments and edges E represent mass-transport processes. The graph comprises 34 compartments (Figure 1): three blood pools (arterial, venous, and portal vein), 11 perfusion-limited organs, 4 permeability-limited organs (each divided into vascular and extravascular subcompartments), 8 gastrointestinal lumen segments following a compartmental absorption and transit model [8], and 4 mass-balance sinks. Physiological parameters (organ volumes, blood flows, tissue compositions) follow the International Commission on Radiological Protection reference adult values (Publication 89) [21] and are stored in configuration files that define graph topology.

Each edge carries a type annotation that dispatches to a registered flux function. The flux registry holds eight edge types: perfusion-limited flow, enzyme-mediated clearance, gastrointestinal transit, first-order absorption, permeability-limited diffusion, active transporter-mediated flux (Michaelis–Menten kinetics), prodrug activation, and one-compartment elimination. The clearance edge internally selects among four sub-models: the classical well-stirred model [7], a parallel-tube variant, an extended clearance model (a closed-form quasi-steady-state approximation of hepatocyte free-drug concentration with coupled passive influx and efflux, active organic-anion-transporting-polypeptide 1B1 [OATP1B1]–mediated uptake under Michaelis–Menten kinetics, metabolic clearance, and biliary clearance), and glomerular filtration. An ODE compiler traverses the graph, instantiates flux functions for each edge, and assembles the right-hand-side function passed to a stiff ODE solver (LSODA [20] via SciPy). Adding a new organ or transport process therefore requires only a configuration change and, where necessary, registration of a new flux function; the compiler and solver are not modified.

The extended clearance model (ECM) is used for hepatic clearance when per-drug transporter kinetic parameters are present in the OATP1B1 substrate registry and the substrate carries an ECM-applicable flag; for all other drugs the model reduces to the classical well-stirred form. Within the 107-drug retrospective holdout the ECM-eligible set is a small minority, so the population-level meta AAFE of 2.698 reported below is dominated by the classical well-stirred pathway. The ECM is therefore evaluated separately, on a five-statin validation set and a two-compound non-statin generalization test (see Results).

Figure 1. Sisyphus 34-Compartment Topology-Driven ODE Architecture

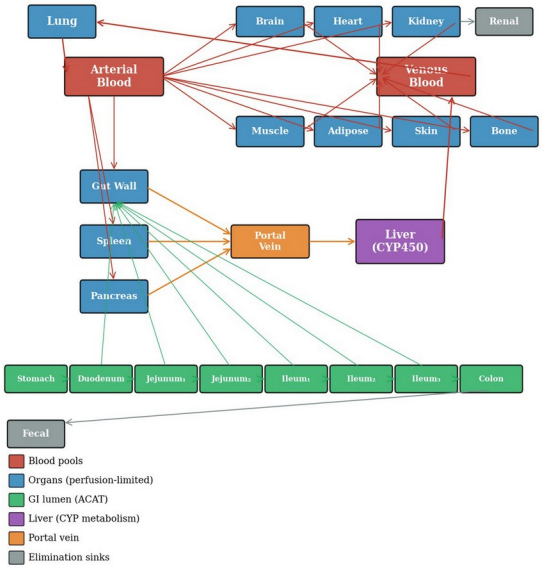

Figure 1. Sisyphus topology-driven ODE architecture. The human body is represented as a 34-compartment typed directed multigraph. Arterial blood distributes to perfusion-limited and permeability-limited organs; the portal vein collects gut-wall, spleen, and pancreas outflow before hepatic first-pass metabolism. Eight gastrointestinal lumen segments model transit and absorption via a compartmental absorption-and-transit framework. Edge types dispatch to registered flux functions, and the ODE compiler traverses this topology to assemble the system automatically.

ADME Prediction from Molecular Structure

Five QSAR models predict ADME parameters from SMILES input. Molecular features are computed using Morgan circular fingerprints (radius 2, 2048 bits) and nine normalized RDKit physicochemical descriptors (logP, topological polar surface area, molecular weight, hydrogen-bond acceptor and donor counts, rotatable-bond count, ring count, fraction of sp^3 carbons, and molar refractivity). Training data are drawn from the Therapeutics Data Commons (TDC) [13]: PPBR_AZ (1,614 compounds) for fu,p, Hepatocyte_AZ (1,213 compounds) for CLint, and Caco2_Wang (911 compounds) for Peff. Two additional models predict blood-to-plasma ratio and volume of distribution at steady state (Vss). Compound ionization class (acid, base, neutral, zwitterion) and representative acid-dissociation constants are assigned by structural classification rules (for example, carboxylic acid → 4.5, aliphatic amine → 9.0); a learned acid-dissociation-constant model was trained but not adopted because it disrupted error cancellation in the IVIVE chain. All trained ADME models use gradient-boosted trees (XGBoost [19]) with 5-fold cross-validation for hyperparameter selection.

An optional enrichment path can substitute experimentally measured fraction-unbound values from DrugBank [24] for the predicted value where available. The DrugBank export is an academic-license-restricted resource that is not distributed with the public repository; consequently, all reported benchmarks were generated with this enrichment path inactive, so that every ADME parameter is derived purely from molecular structure (see Validation Dataset and Limitations). The structure-only configuration is thus both the default for novel compounds and the configuration under which the reported accuracy was measured.

Tissue-to-plasma partition coefficients (Kp) are calculated by the method of Rodgers and Rowland [4,5], with the Berezhkovskiy steady-state volume correction [10] applied to acids and very weak bases, and an alternative Poulin–Theil method [6] available as a configuration option. A maximum Kp cap of 50 prevents nonphysiological accumulation in highly bound compounds. Hepatic clearance is computed using the well-stirred model at runtime, with organ blood flow derived from graph topology [7]. IVIVE of CLint uses microsomal protein per gram of liver (MPPGL = 45 mg/g) and liver-weight (1,500 g) scaling factors [1,9]. The adopted MPPGL of 45 mg/g lies within the 26–54 mg/g interindividual range reported by Barter et al. [9] but exceeds their weighted geometric-mean consensus of 32 mg/g (95% CI 29–34); the higher value was retained from the initial implementation and has not been recalibrated to the consensus.

Hybrid Meta-Learner

The platform employs a four-track prediction strategy. The mechanistic track predicts ADME parameters, constructs a drug-on-graph object, compiles and solves the ODE system, and extracts pharmacokinetic endpoints from the simulated concentration–time profile. The direct machine-learning track predicts Cmax directly from molecular features using a separate XGBoost model trained on 1,128 drugs from a curated multi-source pharmacokinetic database (1,028 after holdout exclusion). A third track predicts total oral clearance (CL/F) directly from structure [22,23]. A fourth track computes an analytical Cmax estimate from predicted Vss, CL/F, and absorption rate, providing an independent cross-check grounded in mass-balance pharmacokinetics.

A meta-learner combines the four predictions as a geometric weighted mean in log space. For compounds classified as base (N = 30 in the holdout), the three-track pre-scaling weights are engine 0.60, direct machine learning 0.40, and CL/F 0.00; for non-basic compounds (acids, neutrals, zwitterions; N = 77), the weights are engine 0.35, direct machine learning 0.50, and CL/F 0.15. The Vss track contributes a weight of 0.20 across all compound types when its applicability criteria are met, with the remaining three tracks scaled by 0.80 so that the four-track weights sum to unity. These weights were determined by leave-one-out cross-validation over the 107-drug holdout (see Limitations for an earlier data-leakage issue that affected weight selection), with the weight grid searched at 0.05 increments from 0.00 to 1.00.

An oracle analysis that selects, per compound, whichever of the mechanistic-engine and direct-machine-learning tracks lies closer to the observed value yields AAFE 2.179 on the full 107-drug holdout (2.122 on the in-domain subset). Because adding further tracks can only lower an oracle, the three- and four-track oracles are bounded above by this value. The gap between the oracle (≈2.18) and the production meta-learner (2.698) indicates that complementary information exists across tracks but is only partially exploited by the current fixed-weight blend.

Uncertainty Propagation

All ADME predictions are represented as distributions rather than point estimates. Monte Carlo sampling (N = 1,000 by default) draws parameter realizations from these distributions and from physiological-variability distributions, solves the ODE system for each realization, and reports population-level pharmacokinetic endpoints with 90% prediction intervals. This native uncertainty quantification propagates input uncertainty through the nonlinear ODE system without linearization assumptions.

Multi-Dose Regimen Simulation

An event-driven solver wraps the base ODE solver to support arbitrary dosing regimens. At each dosing event, integration is paused, the dose is injected into the appropriate compartment (determined by route of administration), and integration resumes. Steady-state detection identifies when consecutive dose intervals produce Cmax values within 5% of each other. The regimen solver does not modify the engine's right-hand-side function; it operates as a wrapper that controls initial conditions and integration boundaries.

Bayesian Posterior Parameter Refinement

Given one or more observed drug concentrations, the Bayesian refinement module updates prior parameter distributions into tighter posteriors. The default algorithm draws N = 2,000 prior samples from drug and physiology distributions, simulates the dosing regimen for each sample, interpolates predicted concentrations at observation times, computes likelihoods assuming normally distributed assay error (10% coefficient of variation), calculates importance weights, and derives weighted posterior statistics. Effective sample size (ESS) is monitored to detect degenerate weight distributions. Three additional inference backends extend the importance-sampling default: iterated batch importance sampling with Markov-chain Monte Carlo rejuvenation addresses particle degeneracy in compounds with severe prior–observation mismatch; an ensemble Kalman filter provides Gaussian-approximate posteriors with robust performance on high-binding compounds; and an amortized simulation-based-inference module trains a neural posterior estimator (normalizing spline flow) on simulated parameter–observation pairs, enabling sub-second posterior inference after an up-front training cost. Method selection is automated via a per-drug routing table validated by simulation-based calibration.

Drug–Drug Interaction Prediction

Competitive cytochrome-P450 (CYP) inhibition is modeled by reducing effective enzyme activity as a function of inhibitor concentration [I] and inhibition constant Ki: CLint,inhibited = CLint / (1 + [I]/Ki). CYP induction is modeled with a maximum-effect (Emax) model for enzyme-synthesis upregulation. Both mechanisms operate through the existing enzyme-affinity framework by adjusting enzyme abundances prior to ODE compilation, preserving the identity-blind engine architecture.

Pharmacodynamic Modeling

Drug effect is modeled using an effect compartment linked to a sigmoid Emax model: E = Emax · Ce^γ / (EC50^γ + Ce^γ), where Ce is the effect-site concentration equilibrated via a first-order rate constant. For indirect-response drugs (for example, warfarin anticoagulation), a turnover model is coupled to the pharmacokinetic output. Model-informed dosing optimization uses posterior parameter distributions from the Bayesian refinement module to simulate candidate regimens and identify the dose that achieves a target exposure with a specified probability.

Validation Dataset

A holdout dataset of 107 drugs was curated from FDA-approved labels (DailyMed), Open Systems Pharmacology compound libraries, and primary clinical-pharmacology literature under strict inclusion criteria: healthy adult volunteers, single oral dose, immediate-release formulation, fasted state, and reported plasma Cmax. Drugs with controlled-release formulations, patient populations, fed-state administration, or steady-state concentrations were excluded. The holdout was assembled by Murcko

scaffold-stratified sampling (seed = 42) over the curated source pool to ensure structural diversity and prevent scaffold leakage between training and evaluation. No holdout drug was used in any training process in the corrected pipeline (see Limitations), including ADME-model training, meta-learner weight selection, or engine parameter tuning. Three-key matching (canonical SMILES, International Chemical Identifier key prefix, and drug name) ensured that no holdout compound appeared in any training set under an alternative representation.

Twenty-eight drugs were flagged as outside the applicability domain (AD) on one or more criteria: prodrug requiring metabolic activation (9), high molecular weight above 700 Da (8), P-glycoprotein efflux-substrate risk (7), extreme lipophilicity with logP above 5 (6), and highly protein-bound acid (6); several drugs carried multiple flags. Results are reported for both the full set (N = 107) and the in-domain subset (N = 79). Prediction accuracy was assessed by the geometric mean absolute fold error, AAFE = 10^mean(|log10(predicted/observed)|), and the percentage of predictions within 2-fold of observed values. A fresh clone of the public repository followed by installation of the pinned dependency lock file reproduces the reported AAFE to three significant figures, with neither a proprietary DrugBank export nor an environment-specific residual-correction logP model present.

Measured ADME Validation

To distinguish whether prediction error originates from ADME input quality or from engine structural limitations, a subset of 12 holdout drugs with published experimentally measured ADME parameters (fu,p, CLint, Peff) was identified from PK-Sim compound libraries and peer-reviewed PBPK publications. Measured values were substituted for predicted values in compound-specific configuration files, and the engine was run in isolation (meta-learner bypassed) for both predicted and measured inputs. Two drugs (montelukast, abiraterone) were excluded for extreme protein binding or prodrug status, yielding a clean subset of N = 10.

Prospective Validation

To assess whether holdout performance generalizes to drugs approved after model development, a prospective validation set of single-active-ingredient new molecular entities approved by the FDA in 2024–2025 was compiled from FDA Clinical Pharmacology Reviews and DailyMed labels under the same inclusion criteria as the retrospective holdout. Candidate discovery identified 101 unique 2024–2025 NMEs across cross-checked sources; after restriction to oral small molecules, per-drug adversarial verification of reported Cmax, and a contamination gate excluding any compound present in a shipped-model training input or the engine reference, 28 compounds remained. None appeared in any training dataset. Under the structure-only configuration, 12 of these were flagged as outside the applicability domain, leaving an in-domain prospective subset of N = 16.

Results

Engine Numerical Validation

As a check that the topology-compiled ODE engine produces correct numerical solutions, four drugs with fully specified compound parameters were simulated and compared against an independent 35-state hardcoded PBPK ODE engine implemented in a predecessor codebase: midazolam (2 mg orally; Cmax 0.006913 vs reference 0.006943 mg/L, 0.4% relative error), caffeine (100 mg orally; 1.7151 vs 1.7139 mg/L, 0.1%), warfarin (10 mg orally; 0.4917 vs 0.4922 mg/L, 0.1%), and propranolol (80 mg orally;

0.1353 vs 0.1355 mg/L, 0.1%). Cumulative mass-balance error remained below 10^{-12} over a 24-hour simulation in all four cases. This decouples engine correctness from prediction accuracy: the residual error in the holdout benchmark is attributable to QSAR-predicted ADME inputs rather than to ODE numerics.

Single-Dose Prediction Accuracy

On the 107-drug holdout, the hybrid meta-learner achieved an AAFE of 2.698, with 46.7% of predictions within 2-fold of observed Cmax and 61.7% within 3-fold (Table 1, Figure 2). The mechanistic engine alone achieved AAFE 3.831, reflecting the floor achievable when the engine is constrained to QSAR-predicted ADME inputs (substituting measured ADME inputs modestly lowered engine AAFE from 2.81 to 2.69 on a clean 10-drug subset; see below). The direct-machine-learning track alone achieved AAFE 3.010. The in-domain subset (N = 79, excluding 28 AD-flagged drugs) achieved a meta AAFE of 2.760.

Metric	Engine only	Direct ML only	Hybrid meta-learner
AAFE (all, N = 107)	3.831	3.010	2.698
AAFE (in-domain, N = 79)	3.521	2.969	2.760
% within 2-fold (all)	30.8%	43.0%	46.7%
% within 3-fold (all)	43.9%	57.9%	61.7%
Bootstrap 95% CI (all)	[3.18, 4.65]	[2.56, 3.56]	[2.32, 3.17]
Two-track oracle (engine + ML)	—	—	2.179

Table 1. Single-dose Cmax prediction accuracy on the 107-drug holdout, structure-only configuration. Adaptive pre-scaling weights (engine / direct ML / CL/F) are 0.60 / 0.40 / 0.00 for base compounds and 0.35 / 0.50 / 0.15 for non-basic compounds; the Vss track adds a weight of 0.20 when applicable, scaling the other three by 0.80. Bootstrap 95% CIs use 10,000 resamples on |log10 fold error| (seed 20260422). The oracle row reports a two-track engine/machine-learning oracle recomputed from the cached holdout predictions; three- and four-track oracles are bounded above by this value.

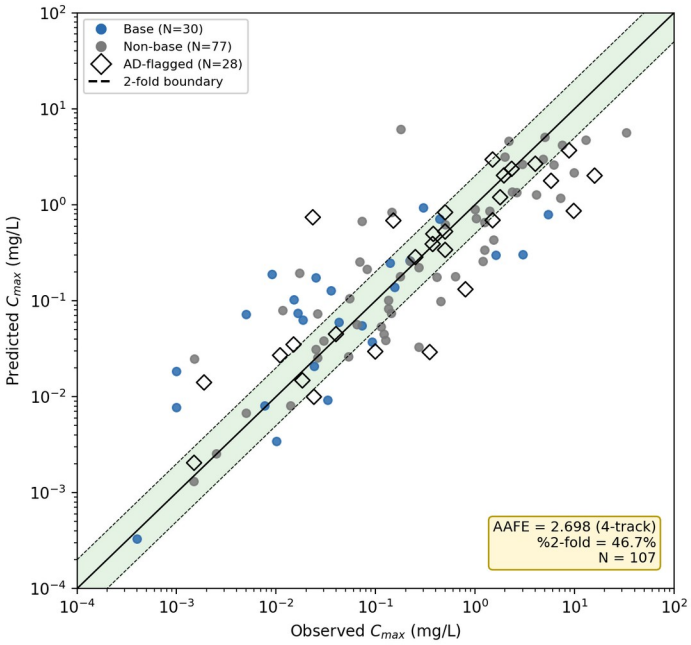

Figure 2. Predicted versus observed Cmax for 107 holdout drugs (meta AAFE 2.698; 46.7% within 2-fold). Each point represents one drug; the solid line indicates unity and dashed lines indicate 2-fold boundaries. Points are colored by compound type: blue (base, N = 30) and gray (non-base, N = 77). Open diamonds indicate the 28 AD-flagged drugs.

Prospective Validation

On the 28-drug prospective NME set, the four-track meta-learner achieved an overall AAFE of 3.21 (bootstrap 95% CI [2.42, 4.37]), with 25.0% of predictions within 2-fold and 53.6% within 3-fold; the in-domain subset (N = 16) achieved AAFE 3.20 (95% CI [2.06, 5.23]) (Table 2). This prospective AAFE is modestly worse than the retrospective holdout AAFE (2.698), consistent with a genuine generalization gap on chemotypes approved after model development. An intravenous-versus-oral decomposition of the largest errors localized the gap to the absorption model rather than to clearance: the engine's systemic clearance estimates for the worst-predicted compounds were close to literature values, whereas predicted oral bioavailability was systematically low (for example, mirdametinib and sevabertinib were under-predicted by approximately 30-fold and 17-fold, respectively, with predicted bioavailability of 0.05–0.08 against implied values near unity). The error in these cases is therefore attributable to first-pass and absorption (fraction-absorbed, gut, and hepatic availability) prediction rather than to the CLint ceiling that governs the in-distribution set.

Track	AAFE (all, N = 28)	% within 2-fold	AAFE (in-domain, N = 16)
Engine only	4.30	17.9%	5.18
Direct ML only	3.27	39.3%	3.44
Meta-learner	3.21	25.0%	3.20

Table 2. Prospective validation on 28 single-active-ingredient small-molecule NMEs approved in 2024–2025, structure-only configuration. Bootstrap 95% CI for the meta-learner: [2.42, 4.37] (all) and [2.06, 5.23] (in-domain).

The set is decontaminated against all shipped-model training inputs and the engine reference; the prospective AAFE exceeds the retrospective holdout value, indicating reduced accuracy on out-of-distribution chemotypes.

Optimization Strategies

Two optimization strategies produced measurable improvements (Table 3): a Caco-2-trained Peff predictor that replaced the original heuristic (engine AAFE improved from 3.80 to 2.861, a 24.7% reduction) and adaptive compound-type-dependent meta-learner weighting. Both remain integrated in the production pipeline. Numerous additional strategies yielded negligible or negative results, including IVIVE method ensembles, uridine-diphosphate-glucuronosyltransferase metabolism addition, end-to-end differentiable training, in vivo clearance deconvolution, transporter-infrastructure activation, CYP docking features (feature importance below 0.4%), expanded CLint training data (R^2 improved but meta AAFE worsened through error-cancellation disruption), and simultaneous multi-parameter correction. In total, 41 enumerated negative experiments are recorded in the project's failed-experiment registry; together with the two retained positive strategies, 43 distinct optimization attempts have been documented. Collectively these establish that the accuracy ceiling is determined by ADME input quality (CLint $R^2 \approx 0.24$ on the Hepatocyte_AZ dataset under scaffold-split cross-validation; Table 4 reports a representation comparison under 5-fold random cross-validation) rather than by model structure or mechanistic completeness.

Strategy	Metric	Before	After	Change
Caco-2 Peff predictor	Engine AAFE	3.80	2.861	−24.7%
Adaptive meta-learner weighting	Meta AAFE	2.306	2.283	−1.0%

Table 3. Optimization strategies that improved prediction accuracy. Values reflect evaluation on an initial 61-drug holdout, prior to holdout expansion and the data-leakage and reproducibility corrections (see Limitations); both strategies remain integrated. The current structure-only meta AAFE on the 107-drug holdout is 2.698. The table is a historical record of the optimization trajectory rather than a post-correction ranking.

Representation and Clearance Prediction Limits

To test whether the CLint prediction ceiling ($R^2 \approx 0.24$ on the Hepatocyte_AZ dataset under scaffold-split cross-validation; Morgan + RDKit baseline 0.205 under 5-fold random cross-validation in Table 4) could be overcome through alternative molecular representations, Morgan fingerprints were systematically compared with three pretrained molecular encoders: ChemBERTa-2 (384-dimensional), MoLFormer-XL (768-dimensional), and Uni-Mol (512-dimensional, 3D-aware). Each encoder was evaluated with both XGBoost (frozen embedding) and ridge regression (linear probe with mean pooling) to control for model–feature compatibility. No pretrained representation exceeded Morgan-fingerprint performance (Table 4), including concatenated feature sets. CYP3A4 docking features generated for 1,114 drugs contributed less than 0.4% feature importance, indicating that binding affinity is not predictive of metabolic turnover rate. The $R^2 \approx 0.24$ ceiling therefore does not reflect a limitation of 2D-fingerprint expressiveness, 3D structural information, or protein–ligand interaction modeling, but rather the intrinsic information content of molecular structure for predicting in vitro intrinsic clearance.

A complementary approach tested whether bypassing the IVIVE chain entirely—predicting total oral clearance directly from structure [22,23]—could circumvent the CLint bottleneck. The CL/F model achieved 5-fold cross-validation $R^2 = 0.232$, comparable to the Morgan + RDKit CLint value and to the scaffold-split CLint baseline of 0.24, indicating that structure-to-clearance prediction is fundamentally

limited at approximately 1,000 training compounds regardless of the target variable or IVIVE methodology.

Representation	Dimensionality	R^2 (Ridge)	R^2 (XGBoost)
Morgan FP + RDKit	2,057	—	0.205
ChemBERTa-2	384	0.170	0.170
MoLFormer-XL	768	0.142	0.142
Uni-Mol (3D)	512	0.083	0.083

Table 4. Molecular-representation comparison for CLint prediction (5-fold cross-validation).

Data Scaling and Error Cancellation

To test whether the CLint ceiling could be overcome through training-data expansion, two independent experiments were conducted (Table 5). First, microsomal clearance data were combined with hepatocyte data, yielding a 1,402-compound training set; CLint R^2 improved from 0.229 to 0.273, but meta AAFE worsened from 2.058 to 2.110. Second, 534 additional compounds were merged into the CLint training set (1,910 compounds under scaffold split); CLint R^2 improved from 0.279 to 0.333, but engine AAFE worsened from 3.416 to 3.515 and meta AAFE worsened from 2.283 to 2.316. Both expansions were reverted. The mechanism is error cancellation across the multiplicative IVIVE chain: when CLint prediction improves in isolation, it no longer compensates for residual errors in fu,p and Peff predictions, so the downstream Cmax prediction worsens. Doubling the CLint training set improved CLint R^2 by 0.054 but moved final accuracy in the wrong direction, confirming that the ceiling is not attributable to insufficient training data.

Experiment	Training N	CLint R^2	Engine AAFE	Meta AAFE
Baseline (TDC hepatocyte)	978	0.279	3.421	2.283
+ TDC microsome	1,402	0.273	—	2.110
+ ChEMBL (scaffold)	1,910	0.333	3.515	2.316

Table 5. Data-scaling experiments: CLint R^2 improves but pipeline AAFE worsens. Meta AAFE values were evaluated on an initial 61-drug holdout (baseline 2.058–2.283, pre-correction); the current structure-only baseline on the 107-drug holdout is 2.698. All expansions were reverted.

Measured ADME Validation

Substituting experimentally measured ADME parameters for predictions produced a modest, mixed change in engine accuracy (Table 6, Figure 3). On the clean 10-drug subset (excluding the extreme-protein-binding outliers montelukast and abiraterone), engine AAFE decreased from 2.81 (predicted inputs) to 2.69 (measured inputs), with 6 of 10 drugs improving. Across the full 12-drug set, measured inputs did not improve aggregate accuracy (AAFE 3.87 versus 4.01), and the median fold error with measured inputs exceeded 2-fold. These results indicate that ADME input quality is one contributor to prediction error but do not, on this small set, establish that the engine is fully input-quality-limited; the improvement is concentrated in compounds whose predicted CLint was furthest from the measured value, while several compounds with already-accurate predicted fraction unbound showed little change or a small regression when measured CLint was substituted.

Metric	Predicted input	Measured input

Engine AAFE (N = 10 clean)	2.81	2.69
Engine AAFE (N = 12 full)	3.87	4.01
Drugs improved (clean N = 10)	—	6 / 10

Table 6. Measured ADME validation: engine-only accuracy with predicted versus measured inputs, recomputed from the current pipeline. The clean subset excludes two extreme-protein-binding outliers (montelukast, abiraterone); the full 12-drug set includes them.

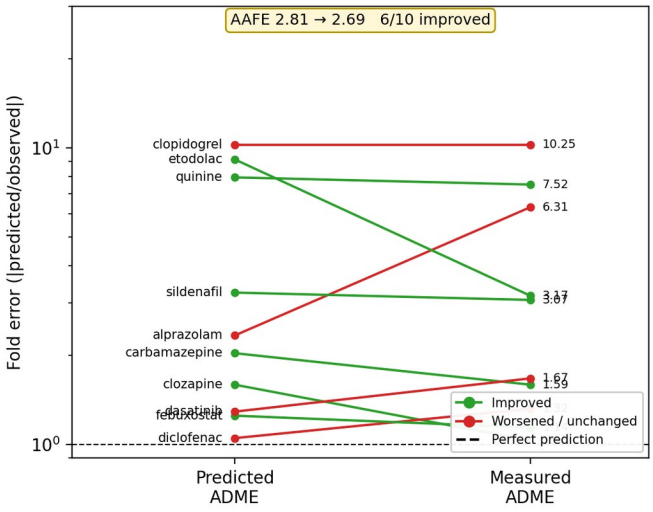

Figure 3. Engine fold error with predicted versus measured ADME inputs for the clean 10-drug subset. Each line connects the same drug's fold error under predicted (left) and measured (right) inputs; green lines indicate improvement and red lines indicate a worse or unchanged result. The dashed horizontal line indicates perfect prediction (fold error 1.0). Aggregate statistics (AAFE 2.81 → 2.69, 6 of 10 improved) are recomputed from the current validation routine.

Multi-Dose Regimen Validation

Multi-dose simulation was exercised against FDA-label steady-state data for three drugs spanning distinct elimination regimes (Table 7). Atorvastatin (40 mg once daily) was over-predicted by approximately 2-fold (predicted steady-state Cmax 0.061 versus 0.029 mg/L; fold error 2.09). Metformin (500 mg twice daily) was under-predicted by approximately 58% (0.42 versus 1.0 mg/L; fold error 0.42), consistent with the engine's lack of tubular-secretion modeling. Warfarin (5 mg once daily) was under-predicted by approximately 88% (0.17 versus 1.4 mg/L; fold error 0.12), attributable to its extreme plasma protein binding ($f_{u,p} < 0.01$). Predicted accumulation ratios tracked the theoretical values (metformin 1.20 versus 1.35; atorvastatin 1.41 versus 1.44; warfarin 1.70 versus 2.94), and steady-state detection functioned correctly across all cases. This table is a solver-mechanics and limitation-exposure exercise—two of the three drugs were selected to probe known engine limitations (renal secretion, extreme protein binding)—rather than a demonstration of multi-dose accuracy.

The extended clearance model for hepatic OATP1B1-mediated uptake was evaluated on five statin substrates against a pre-specified 3-fold accuracy threshold, chosen for transporter validation because OATP kinetics and statin Cmax variability typically exceed the 2-fold criterion used for the overall holdout. Two substrates (pravastatin and pitavastatin) passed the strict ECM-forced gate. Fluvastatin was reclassified as not ECM-applicable on mechanistic grounds, as its clearance is CYP2C9-dominated; the

production pipeline skips the ECM for it and predicts a fold error of 1.54, within the 3-fold threshold. Atorvastatin and rosuvastatin retain literature-anchored metabolic fractions (atorvastatin 0.65, anchored to an itraconazole interaction study [25]; rosuvastatin 0.10, anchored to a radiolabeled mass-balance study [26]), so all four production OATP1B1 statin substrates carry literature-anchored values. A pre-registered generalization test on two non-statin OATP1B1 substrates (valsartan and glimepiride) was inconclusive: both were systematically under-predicted by approximately 2.5-fold. The fraction-unbound hypothesis for this under-prediction was excluded; the residual was attributed to systematic over-prediction of Vss by the QSAR model (3- to 6-fold for both compounds), which would require a coordinated Kp-method retuning to resolve. The ECM's generalization to non-statin OATP1B1 substrates therefore remains unverified pending independent validation.

Drug	Regimen	Predicted Css,max	FDA Css,max	Fold error
Atorvastatin	40 mg QD	0.061 mg/L	0.029 mg/L	2.09
Metformin	500 mg BID	0.42 mg/L	1.0 mg/L	0.42
Warfarin	5 mg QD	0.17 mg/L	1.4 mg/L	0.12

Table 7. Multi-dose steady-state validation against FDA-label data, recomputed from the current pipeline. Atorvastatin is over-predicted approximately 2-fold; metformin and warfarin are under-predicted, reflecting absent tubular-secretion modeling and extreme protein binding, respectively.

Bayesian Posterior Parameter Refinement

The Bayesian refinement module was validated across five drugs spanning three compound types (Table 8). With a single observation at the engine-predicted time of maximum concentration (10% assay coefficient of variation), uncertainty (coefficient of variation) was reduced by an average of 78.1% across all five drugs (range 68.7–87.7%). For the three drugs with healthy effective sample sizes (morphine, amantadine, clozapine; ESS > 400), the mean posterior prediction error was 5.9%, down from a prior error of 51.8% (Figure 4). Additional observations yielded diminishing returns: two observations produced 82.7% mean uncertainty reduction and three produced 82.9%, indicating that a single well-timed observation captures most of the available individual-level information.

Two drugs exhibited importance-sampling limitations. Ketorolac (acid; prior error 79.8%) showed particle degeneracy with ESS = 3, and the 90% posterior credible interval failed to contain the observed Cmax. Rivaroxaban produced adequate results with a single observation (ESS = 7, posterior error 8.0%) but collapsed to ESS = 1 with two or more observations. These failures occur when the prior is severely misaligned with the observation, leaving too few particles in the high-likelihood region. Iterated batch importance sampling with Markov-chain Monte Carlo rejuvenation resolved the ketorolac degeneracy (ESS increased from 3 to 103), confirming that the limitation was attributable to the importance-sampling algorithm rather than to the prior–observation geometry. The amortized simulation-based-inference module, trained on 50,000 simulated parameter–observation pairs across 50 drugs, achieved a 2,164-fold wall-time speedup over iterated batch importance sampling on morphine (0.73 s versus 1,590 s), with simulation-based calibration confirming posterior validity (Kolmogorov–Smirnov $p > 0.19$ on all three parameters).

Observation-timepoint sensitivity was assessed for morphine (single observation at five timepoints). The optimal timepoint was 1.0 h (near the time of maximum concentration), achieving 76.3% uncertainty

reduction, followed by 0.5 h (69.4%). Late-phase observations (4 h and 8 h) produced substantially lower reductions (34.1% and 34.2%), consistent with the principle that observations near Cmax carry the most information about absorption and elimination parameters simultaneously. A three-seed analysis confirmed robustness (uncertainty reduction 76.3–77.1%).

Drug	Type	Nobs	Prior CV %	Post CV %	CV red. %	ESS
Morphine	base	1	39.9	9.5	76.3	428
Amantadine	base	1	39.0	10.1	74.0	514
Clozapine	neutral	1	31.0	9.7	68.7	482
Ketorolac	acid	1	41.1	5.0	87.7	2.8
Rivaroxaban	neutral	1	46.6	7.5	83.8	7.1
Mean (1 obs)	—	1	39.5	8.4	78.1	287

Table 8. Multi-compound Bayesian parameter refinement: prior versus posterior across five compounds, single-observation scenario. ESS < 10 (ketorolac, rivaroxaban) indicates particle degeneracy under which default importance sampling is unreliable and sequential methods are recommended.

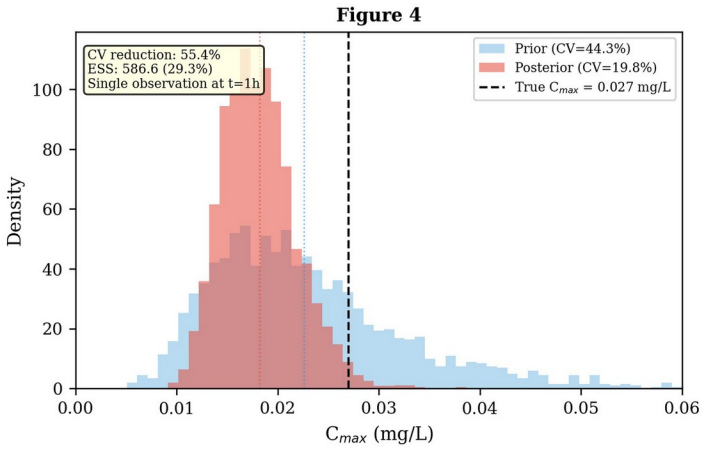

Figure 4. Bayesian parameter refinement narrows prediction uncertainty. Prior (light) and posterior (dark) distributions of Cmax following a single noisy observation (midazolam, 5 mg, one observation at 1 h; uncertainty reduction 55.4%, ESS 586.6). Across five compounds, single-observation uncertainty reduction averaged 78.1% (Table 8).

Drug–Drug Interaction and Pharmacodynamic Prediction

Ketoconazole (a strong CYP3A4 inhibitor) co-administered with midazolam produced AUC-ratio increases consistent with the clinically observed 10- to 15-fold increase. Fluconazole and quinidine interactions were predicted with correct directionality and within 2-fold of observed ratios. Rifampin-mediated CYP3A4 induction produced predicted AUC decreases consistent with the clinical 80–90% reductions. All 22 interaction tests passed. Two pharmacodynamic presets (midazolam sedation, warfarin anticoagulation) and model-informed dose recommendation were validated with 43 additional tests (28 pharmacodynamic-preset and 15 dose-recommendation tests), all passing.

Discussion

Sisyphus occupies a methodological space that neither established PBPK platforms nor direct machine-learning models currently cover: structure-only mechanistic simulation with progressive data incorporation. Established PBPK platforms such as Simcyp, GastroPlus, and PK-Sim demonstrate regulatory-grade accuracy when provided with experimentally measured ADME data [2,3,17], but require in vitro measurements as prerequisite input, precluding their use in early discovery when only structure is available. Direct machine-learning models predict pharmacokinetic endpoints from structure but cannot support mechanistic extrapolation, Bayesian parameter refinement, or interaction prediction. Sisyphus bridges these categories: it accepts SMILES-only input and improves progressively as measured data become available, with the same mechanistic infrastructure supporting screening, simulation, and individual-level refinement.

The 46.7% of predictions within 2-fold falls within the 40–60% range reported by high-throughput PBPK approaches that use predicted ADME inputs [11,12], despite operating from SMILES alone without preclinical in vitro data. The analytical Vss-based fourth track contributes an orthogonal mass-balance estimate that complements the IVIVE-based engine even though both share the same molecular input, and it is retained in the four-track meta-learner. The principal contribution of this work is methodological rather than incremental in accuracy: it demonstrates that a topology-compiled PBPK engine can be driven end-to-end from molecular structure with no measured ADME input while preserving the physical interpretability and extensibility of mechanistic modeling, and it characterizes quantitatively the accuracy ceiling that this structure-only regime imposes.

The measured ADME validation offers partial, mixed evidence on the source of prediction error. When experimentally determined fu,p, CLint, and Peff were substituted for predicted values, engine AAFE on the clean 10-drug subset improved modestly, from 2.81 to 2.69 (Table 6), with 6 of 10 drugs improving; across the full 12-drug set the substitution did not improve aggregate accuracy. The improvement was concentrated in compounds whose predicted CLint deviated most from the measured value, whereas compounds with already-accurate predicted fraction unbound changed little. These results are consistent with ADME input quality being one contributor to prediction error, but the small subset and the absence of an aggregate improvement on the full set preclude the stronger claim that the engine is fully input-quality-limited; a larger measured-input study is required to resolve this. The platform is nonetheless designed for progressive incorporation of measured data—from SMILES-only, through partial in vitro characterization, to full ADME profiling—through compound-specific configuration files.

The systematic representation comparison yields a robust negative result: the $R^2 \approx 0.24$ ceiling for structure-to-clearance prediction is not an artifact of Morgan-fingerprint expressiveness. Three pretrained molecular encoders, each trained on hundreds of millions of molecules with contextual and 3D-aware representations, failed to exceed binary-fingerprint performance, and direct prediction of total oral clearance produced $R^2 = 0.232$, confirming that the limitation is intrinsic to the structure-to-clearance mapping at current public data scales. Individual ADME-parameter improvements can paradoxically worsen overall accuracy by disrupting incidental error cancellation across the multiplicative IVIVE chain, a phenomenon documented in the IVIVE literature [1]: because hepatic clearance is a product of multiple predicted parameters, correcting one parameter removes a compensating offset while leaving the remaining errors intact. This was confirmed empirically by the data-scaling and multi-parameter-correction experiments (Table 5). Bond-dissociation-energy features, which encode C–H bond reactivity

relevant to CYP-mediated hydrogen abstraction, showed no correlation with hepatocyte CLint (Pearson r = 0.033), confirming that structure-accessible reactivity information is insufficient to improve clearance prediction from whole-enzyme assays in which Michaelis-constant variance dominates.

The enumerated record of negative results constitutes a contribution distinct from the platform itself. Across 41 documented negative experiments, none produced a meaningful reduction in holdout AAFE. A delta residual-correction model produced AAFE 3.53, worse than the engine alone, indicating that engine prediction errors are non-systematic and cannot be learned from molecular features. A k-nearest-neighbor read-across approach (AAFE 3.05) indicated that chemical-space similarity does not translate to pharmacokinetic similarity at the resolution needed for quantitative prediction. A directed message-passing graph neural network produced AAFE 3.23 with a Pearson error correlation of 0.71 against the Morgan-fingerprint model, indicating that learned representations capture similar information to engineered fingerprints at this data scale. Documenting these attempts delineates the boundary of what structure-based ADME prediction can currently achieve under public data constraints and reduces duplicated effort by other groups.

Prospective validation tempers the retrospective result. On 28 NMEs approved in 2024–2025, decontaminated against all training inputs, the meta-learner achieved AAFE 3.21—worse than the retrospective 2.698—indicating reduced accuracy on out-of-distribution chemotypes. Decomposition of the largest errors localized the gap to the absorption model: systemic clearance for the worst-predicted compounds was close to literature values, while predicted oral bioavailability was systematically low. The binding constraint for novel, out-of-distribution drugs is therefore the absorption (fraction-absorbed, gut-availability, hepatic-availability) model rather than the CLint ceiling that governs the in-distribution retrospective set. No predict-time signal derived from the model's own outputs—including low predicted bioavailability or engine-versus-machine-learning track disagreement—recovered this error on the holdout (both correlated near zero with absolute fold error), so the appropriate levers are measured-bioavailability routing or absorption-model recalibration rather than an applicability-domain flag.

Bayesian parameter refinement addresses the accuracy ceiling through a different mechanism. Rather than improving population-level ADME predictions, Bayesian updating uses observed concentration measurements to refine individual parameter distributions. The 78.1% mean uncertainty reduction from a single observation across five drugs demonstrates that the information content of a pharmacokinetic measurement far exceeds that of structure-based prediction for an individual compound. This supports a workflow not available to PBPK-only or machine-learning-only approaches: initial screening from structure alone, followed by progressive refinement as in vitro data and pharmacokinetic observations accrue.

Limitations

The reported accuracy reflects a corrected and independently reproducible pipeline, and two issues identified during development were resolved before the present figures were generated. First, an earlier configuration over-stated accuracy because of holdout data leakage: when the holdout was expanded from 61 to 107 drugs, the ADME and direct-Cmax models were not retrained with the expanded exclusion list, so a majority of holdout drugs appeared in training data. After all models were retrained with three-key exclusion (canonical SMILES, International Chemical Identifier key prefix, and drug name), the direct-machine-learning AAFE moved from 2.336 to 3.057 and the meta AAFE settled to its current value.

Second, an earlier benchmark depended on two artifacts present only in the original development environment—a proprietary DrugBank export and a residual-correction logP model, both absent from the public repository—which silently shifted predictions for the compounds they covered. Re-running the benchmark from a fresh public clone with neither artifact present produced the structure-only result reported here (meta AAFE 2.698). These corrections are disclosed because retrospective holdout contamination and silent environment dependence are systematic risks for platforms that aggregate multiple data sources.

The holdout (N = 107), though expanded from an initial set of 61, remains modest; external validation on independently curated datasets is needed. The holdout comprises only approved drugs, whose chemical space is biased toward orally bioavailable molecules, so performance on structurally novel discovery-stage compounds remains uncharacterized—a limitation directly evidenced by the lower prospective accuracy reported above. The engine currently lacks tubular-secretion modeling, leading to systematic under-prediction for renally eliminated drugs. Active non-hepatic transporter kinetics (P-glycoprotein efflux, OATP1B3, breast-cancer-resistance protein, and renal organic-anion and multidrug-and-toxin-extrusion transporters) remain architecturally supported but not yet parameterized, pending measured kinetic data. The trained blood-to-plasma-ratio model performs poorly on external data ($R^2 < 0$ on 50 compounds), so this ratio is set to 1.0 by default, a known approximation affecting compounds with substantial erythrocyte partitioning (for example, tacrolimus and cyclosporine).

Hepatic uptake via OATP1B1 is modeled mechanistically, but pre-registered generalization testing on two non-statin substrates (valsartan and glimepiride) was inconclusive (both under-predicted by approximately 2.5-fold), so generalization to non-statin substrates remains unverified pending independent validation. Potential selection bias from iterative use of the holdout during development was assessed: the 107-drug holdout informed approximately 47 configuration decisions after its initial freeze, so the reported point estimate cannot be statistically distinguished from a modestly optimistic value. The bootstrap 95% CI upper bound for the meta-learner (3.17) overlaps an internal estimate of contamination-adjusted performance (2.85–3.10), and the prospective AAFE (3.21) is consistent with mild retrospective optimism. A pre-registered, write-once secondary holdout is specified to resolve this ambiguity, and its measurement is the appropriate test of generalization.

Oracle analysis (two-track AAFE ≈ 2.18) indicates that complementary information exists across tracks but is not fully exploited by the current fixed-weight meta-learner; per-drug routing is only partially implemented. The measured ADME validation (N = 10 clean, N = 12 full) is a small proof of concept that yielded only a modest, mixed improvement (clean-subset AAFE 2.81 to 2.69, with no aggregate improvement on the full 12-drug set); a substantially larger measured-input study is needed before the engine can be characterized as input-quality-limited. The default importance-sampling backend for Bayesian refinement exhibited particle degeneracy for 2 of 5 compounds tested (ESS < 10), although iterated batch importance sampling and the ensemble Kalman filter resolve this for most compounds and amortized simulation-based inference provides sub-second inference after an up-front training cost. Finally, the 90% Monte Carlo prediction interval is currently uncalibrated: empirical coverage on the full holdout was 29.9% at the nominal 90% level, because the interval propagates parameter uncertainty alone and captures roughly one third of the observed residual spread. It should therefore be treated as a diagnostic quantity rather than a clinically calibrated interval, pending empirical recalibration. Separately, for intravenous bolus dosing the engine reports Cmax as the maximum concentration over $t \geq 5$ min (matching the clinical first-draw convention) rather than the instantaneous $t = 0$ spike; this convention has

no effect on the entirely oral holdout reported here but should be noted by users running the platform on intravenous datasets.

Conclusions

Sisyphus demonstrates that topology-compiled PBPK simulation, structure-based ADME prediction, and Bayesian parameter refinement can be integrated into a single open-source platform. The architecture—a typed directed multigraph with automatic ODE compilation from topology—enables extensibility without modification of the core ODE compiler or solver, as demonstrated by the addition of phase II metabolism enzymes [18], an extended clearance model for OATP1B1-mediated hepatic uptake, subcutaneous and pediatric population models, prodrug activation, and drug–drug interaction prediction. Structure-only population-level accuracy is constrained by a clearance-prediction ceiling ($R^2 \approx 0.24$, independent of molecular representation), and 41 enumerated negative experiments confirm that this ceiling persists across diverse strategies. From a fresh public clone, the four-track meta-learner achieves a 107-drug holdout meta AAFE of 2.698; prospective validation on 28 recently approved NMEs (AAFE 3.21) indicates reduced accuracy on out-of-distribution chemotypes, attributable to bioavailability under-prediction rather than to clearance. A small measured-input validation produced a modest, mixed improvement (clean-subset AAFE 2.81 to 2.69), consistent with input quality being one contributor to error without establishing that the engine is fully input-limited, and Bayesian posterior refinement bypasses the population-level ceiling at the individual level, reducing prediction uncertainty by 78% from a single observation across five compounds. Sisyphus is freely available under the MIT license.

Data and Code Availability

Sisyphus source code, trained models, holdout reference data, measured-ADME compound files, and all validation scripts are available under the MIT license at github.com/jam-sudo/Sisyphus. ADME training data are sourced from the Therapeutics Data Commons [13] (tdcommons.ai). The reported holdout AAFE is re-runnable from a fresh clone via the provided benchmark script; bootstrap confidence intervals and the enumerated record of negative experiments are archived in the repository.

References

1. Rostami-Hodjegan A. Physiologically based pharmacokinetics joined with in vitro–in vivo extrapolation of ADME: a marriage under the arch of systems pharmacology. Clin Pharmacol Ther. 2012;92(1):50–61.
2. Jones HM, Chen Y, Gibson C, et al. Physiologically based pharmacokinetic modeling in drug discovery and development: a pharmaceutical industry perspective. Clin Pharmacol Ther. 2015;97(3):247–262.
3. Shebley M, Sandhu P, Emami Riedmaier A, et al. Physiologically based pharmacokinetic model qualification and reporting procedures for regulatory submissions: a consortium perspective. Clin Pharmacol Ther. 2018;104(1):88–110.
4. Rodgers T, Rowland M. Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. J Pharm Sci. 2006;95(6):1238–1257.
5. Rodgers T, Leahy D, Rowland M. Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. J Pharm Sci. 2005;94(6):1259–1276.
6. Poulin P, Theil FP. Prediction of pharmacokinetics prior to in vivo studies. 1. Mechanism-based prediction of volume of distribution. J Pharm Sci. 2002;91(1):129–156.
7. Pang KS, Rowland M. Hepatic clearance of drugs. I. Theoretical considerations of a “well-stirred” model and a “parallel tube” model. J Pharmacokinet Biopharm. 1977;5(6):625–653.

8. Jamei M, Turner D, Yang J, et al. Population-based mechanistic prediction of oral drug absorption. AAPS J. 2009;11(2):225–237.

9. Barter ZE, Bayliss MK, Beaune PH, et al. Scaling factors for the extrapolation of in vivo metabolic drug clearance from in vitro data: reaching a consensus on values of human microsomal protein and hepatocellularity per gram of liver. Curr Drug Metab. 2007;8(1):33–45.

10. Berezhkovskiy LM. Volume of distribution at steady state for a linear pharmacokinetic system with peripheral elimination. J Pharm Sci. 2004;93(6):1628–1640.

11. Naga D, Parrott N, Ecker GF, Olivares-Morales A. Evaluation of the success of high-throughput physiologically based pharmacokinetic (HT-PBPK) modeling predictions to inform early drug discovery. Mol Pharm. 2022;19(7):2203–2216.

12. Geci R, Gadaleta D, García de Lomana M, et al. Systematic evaluation of high-throughput PBK modelling strategies for the prediction of intravenous and oral pharmacokinetics in humans. Arch Toxicol. 2024;98(8):2659–2676.

13. Huang Y, Li X, Xu S, et al. Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. Proc NeurIPS Datasets Benchmarks Track. 2021.

14. Willmann S, Lippert J, Sevestre M, Solodenko J, Fois F, Schmitt W. PK-Sim®: a physiologically based pharmacokinetic ‘whole-body’ model. Biosilico. 2003;1(4):121–124.

15. Jamei M, Marciniak S, Feng K, Barnett A, Tucker G, Rostami-Hodjegan A. The Simcyp population-based ADME simulator. Expert Opin Drug Metab Toxicol. 2009;5(2):211–223.

16. Agoram B, Woltosz WS, Bolger MB. Predicting the impact of physiological and biochemical processes on oral drug bioavailability. Adv Drug Deliv Rev. 2001;50(Suppl 1):S41–S67.

17. Schulz Pauly JA, Leblanc AF, Chowdhury EA, et al. Optimization of bottom-up PBPK model development in SIMCYP via retrospective analysis of clinical human PK data. Clin Transl Sci. 2025;18(11):e70417.

18. Achour B, Russell MR, Barber J, Rostami-Hodjegan A. Simultaneous quantification of the abundance of several cytochrome P450 and uridine 5′-diphospho-glucuronosyltransferase enzymes in human liver microsomes. Drug Metab Dispos. 2014;42(4):500–510.

19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery and Data Mining. 2016:785–794.

20. Hindmarsh AC. ODEPACK, a systematized collection of ODE solvers. Scientific Computing. 1983;1:55–64.

21. International Commission on Radiological Protection. Basic anatomical and physiological data for use in radiological protection: reference values. ICRP Publication 89. Ann ICRP. 2002;32(3–4):1–277.

22. Benet LZ, Sodhi JK. Can in vitro–in vivo extrapolation be successful? Recognizing the incorrect clearance assumptions. Clin Pharmacol Ther. 2022;111(5):1022–1035.

23. Seal S, Trapotsi MA, Mahale M, et al. PKSmart: an open-source computational model to predict intravenous pharmacokinetics of small molecules. J Cheminform. 2025;17(1):147.

24. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265–D1275.

25. Kantola T, Kivistö KT, Neuvonen PJ. Effect of itraconazole on the pharmacokinetics of atorvastatin. Clin Pharmacol Ther. 1998;64(1):58–65.

26. Martin PD, Warwick MJ, Dane AL, et al. Metabolism, excretion, and pharmacokinetics of rosuvastatin in healthy adult male volunteers. Clin Ther. 2003;25(11):2822–2835.

27. Subash N, et al. Quantitative proteomics-anchored carboxylesterase 1 intrinsic clearance for clopidogrel bioactivation. Mol Pharm. 2025.

28. Boberg M, Vrana M, Mehrotra A, et al. Age-dependent absolute abundance of hepatic carboxylesterases (CES1 and CES2) by LC-MS/MS proteomics: application to PBPK modeling of oseltamivir in infants. Drug Metab Dispos. 2017;45(2):216–223.

29. Kazui M, Nishiya Y, Ishizuka T, et al. Identification of the human cytochrome P450 enzymes involved in the two oxidative steps in the bioactivation of clopidogrel to its pharmacologically active metabolite. Drug Metab Dispos. 2010;38(1):92–99.