Journal of Pharmacokinetics and Pharmacodynamics (2026) 53:56
https://doi.org/10.1007/s10928-026-10057-4

ORIGINAL PAPER

Formalizing method choice: a data-driven selection of mechanistic models for predicting human heart drug partitioning

Seweryn Ulaszek[1,2] · Bartek Lisowski[1] · Monika Jesionek[1] · Barbara Wiśniowska[3] · Sebastian Polak[1,4]

Received: 8 June 2026 / Accepted: 26 July 2026
© The Author(s) 2026

Abstract

We developed a workflow based on a human-derived dataset, aiming to predict the human heart-to-plasma partition coefficient, K_p. The workflow combines established mechanistic tissue-composition models with an AI/ML-based selector that supports compound-specific selection among mechanistic heart K_p prediction methods. Three published mechanistic approaches were implemented and parameterized in a harmonized manner consistent with the described methodologies to generate baseline K_p predictions. The ML-based selector was trained using physicochemical, ADME, and pharmacokinetic descriptors to recognize compound-specific differences in model performance. This hybrid strategy preserves mechanistic interpretability while using empirical pattern recognition to formalize a priori model choice across compounds. In internal cross-validation, the selector reduced extreme outliers and provided a reproducible, data-driven rule for compound-specific method selection, although it did not materially outperform the strongest standalone mechanistic baseline in numerical error. Although no individual mechanistic method performed consistently across the full chemical space, the Poulin–Theil approach showed the best overall performance on human whole-heart data. The workflow was also applied prospectively to transthyretin stabilizers, for which reliable heart K_p estimates may support further mechanistic modelling of cardiac drug exposure and downstream pharmacological effects. Together, these findings suggest that combining mechanistic calculators with an AI/ML-based selector may provide a practical basis for a priori, compound-specific method selection, reducing reliance on subjective choice of a single mechanistic approach.

Keywords Partition coefficient · Heart · Drug distribution · Pharmacokinetics

Introduction

Heart tissue penetration is an important determinant of cardiac drug exposure and is therefore relevant to physiologically based pharmacokinetic (PBPK) modelling. Within PBPK frameworks, tissue distribution is commonly parameterized using tissue-to-plasma partition coefficients (K_p), which quantify the steady-state concentration ratio between tissue and plasma [1]. Over the past two decades, multiple mechanistic tissue-composition-based approaches have been proposed to estimate $K_p$values from compound properties such as lipophilicity, pKa, and fraction unbound in plasma together with tissue composition parameters, including the methods of Poulin–Theil, Rodgers–Rowland, and Schmitt [1–7]. In practice, however, predictive performance varies across compounds and tissues, and comparative evaluations indicate that no single published method for estimating K_p is uniformly optimal when applied broadly [8]. This suggests that K_p estimation should be considered not only a calculation problem, but also a model-selection problem.

Yun and colleagues proposed a decision-tree–based framework for selecting, for a given compound and tissue, the K_p prediction algorithm most likely to perform best

✉ Sebastian Polak
sebastian.polak@uj.edu.pl

1 Chair of Pharmaceutical Technology and Biopharmaceutics, Faculty of Pharmacy, Jagiellonian University Medical College, Medyczna 9, Kraków 30-688, Poland

2 Doctoral School of Medical and Health Sciences, Jagiellonian University Medical College, Kraków, Poland

3 Department of Social Pharmacy, Faculty of Pharmacy, Jagiellonian University Medical College, Medyczna 9 Street, Krakow 30-688, Poland

4 Certara Predictive Technologies (a Certara Company), Level 2-Acero, 1 Concourse Way, Sheffield S1 2BJ, UK

Published online: 25 September 2026

among multiple published methods [9]. Their work demonstrated the feasibility of structured, data-driven algorithm selection for tissue distribution modelling across multiple organs. Building on that concept, the present study focuses on human whole-heart K_p (denoted throughout this publication as $K_{p,\mathrm{heart}}$) through a selector framework trained on human heart distribution data and evaluated across multiple empirical routines for compound-specific method selection [10]. The resulting workflow is intended to support cardiac exposure modelling, including prospective application to transthyretin stabilizers.

Accordingly, the aim of the current work was to develop and test a reproducible framework for predicting $K_{p,\mathrm{heart}}$ using only the data on human whole-heart drug concentration (Fig. 1). Three mechanistic models (Poulin–Theil, Rodgers–Rowland, and Schmitt) were implemented using equations and assumptions derived from original publications [1, 3, 4]. No recalibration of these equations was performed. The objective was to assess whether a structured selection layer improves heart-specific predictive performance when evaluated within a human heart dataset.

As a prospective translational application, the workflow was further applied to transthyretin stabilizers. For this drug class, heart K_p is relevant because it can influence estimated cardiac exposure and can therefore support subsequent mechanistic modeling of drug effects in the heart. Tafamidis was selected as an illustrative example because a previously developed PBPK model was available [11], allowing the predicted $K_{p,\mathrm{heart}}$ value to be propagated into an explicit heart compartment and used to examine its impact on organ-level exposure predictions.

Theoretical

The theoretical framework of this study is based on three published tissue-composition models implemented to calculate $K_{p,\mathrm{heart}}$: Poulin–Theil (PT), Rodgers–Rowland (R&R), and Schmitt. These models estimate heart-to-plasma partitioning from compound-specific physicochemical properties, plasma binding, ionization state, and physiological tissue-composition parameters. They were selected because they represent established, physiologically interpretable frameworks for tissue partitioning and could be applied systematically across the curated dataset using publicly available or inferable compound-specific inputs. Although all three models predict the same endpoint, $K_{p,\mathrm{heart}}$, they differ in how they formalize lipid partitioning, tissue water distribution, plasma and tissue protein binding, ionization-dependent distribution, and phospholipid interactions. Therefore, the equations below are presented to define the mechanistic assumptions underlying each candidate $K_{p,\mathrm{heart}}$ calculator before their empirical comparison and AI/ML-based selection. All three models were implemented using equations and assumptions derived from their published approaches, without recalibration of the original mechanistic structure.

Poulin–Theil (PT)

The heart-to-plasma partition coefficient was calculated using a Berezhkovskiy-corrected partition expression [1, 12, 13]:

$$K_{p,\mathrm{heart}} = \frac{(f_{NL,t}+0.3f_{NP,t})P+0.7f_{NP,t}+V_{W,t}/f_{ut}}{(f_{NL,p}+0.3f_{NP,p})P+0.7f_{NP,p}+V_{W,p}/f_{up}} \quad (1)$$

where $f_{NL,t}$ and $f_{NP,t}$ are the tissue neutral lipid and neutral phospholipid fractions, $f_{NL,p}$ and $f_{NP,p}$ are the corresponding plasma fractions, $V_{W,t}$ and $V_{W,p}$ are the tissue and plasma water terms, f_{ut} is the tissue unbound fraction, f_{up} is the plasma unbound fraction, and P is the octanol to water partition coefficient of the unionized compound. The latter was calculated as:

$$P = 10^{\log P} \quad (2)$$

In the present workflow, the plasma and tissue binding terms were defined as:

$$B_p = \frac{1-f_{up}}{f_{up}} \ and \quad (3)$$

$$B_t = 0.5B_p \quad (4)$$

and the tissue unbound fraction was then estimated as:

$$f_{ut} = \frac{1}{1+B_t} \quad (5)$$

Rodgers–Rowland (R&R)

For acids and weak bases, the unbound tissue-to-plasma water partition coefficient was represented as the sum of intracellular water, extracellular water, lipid, and extracellular protein-binding contributions [3]:

$$K_{pu} = f_{EW} + \frac{X}{Y}f_{IW} + \frac{Pf_{NL}+(0.3P+0.7)f_{NP}}{Y} + K_{aPR}[PR]_T \quad (6)$$

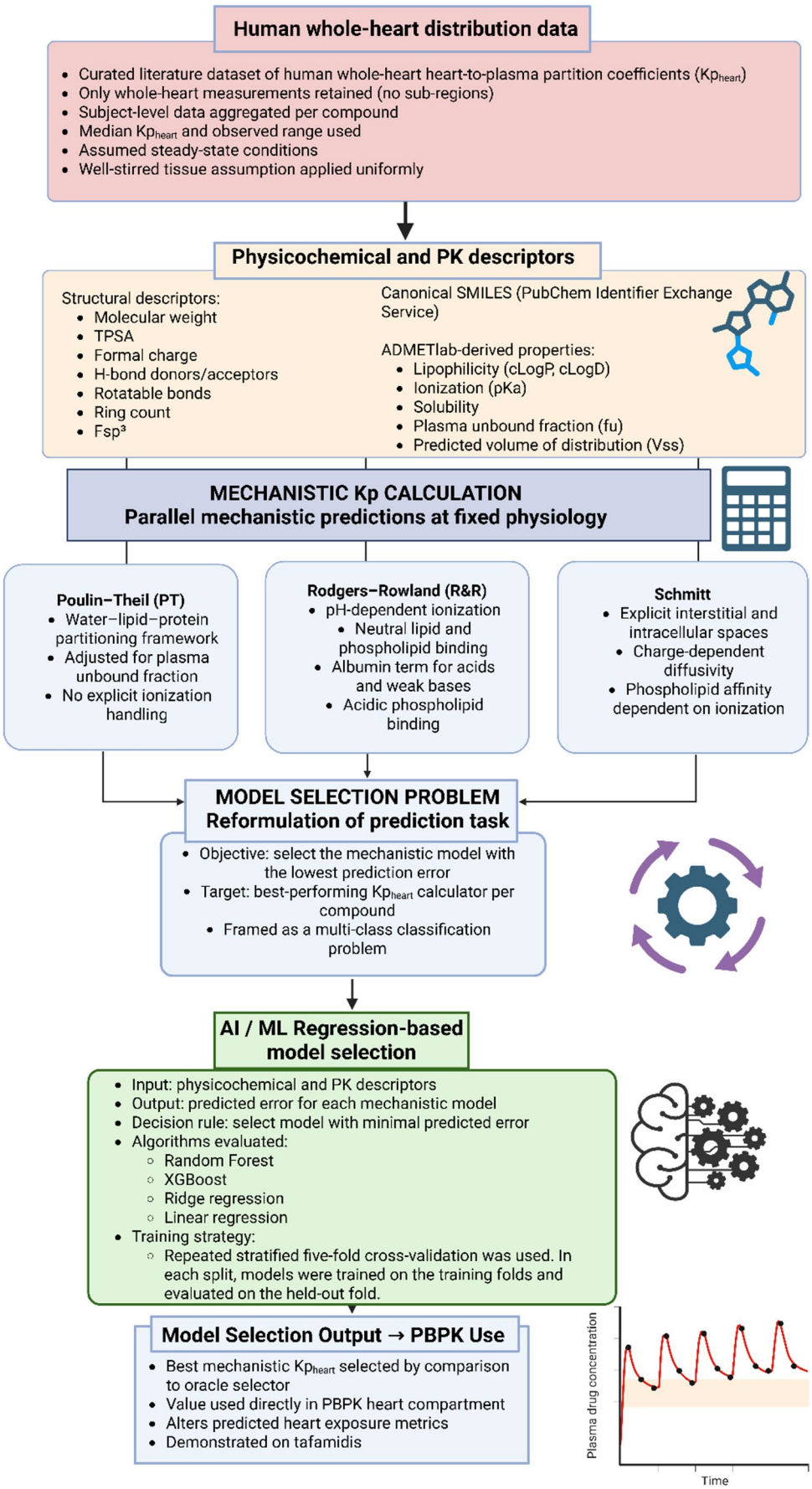


Fig. 1 Diagram of the proposed workflow

and the corresponding tissue-to-plasma partition coefficient was calculated as:

$$K_p = K_{pu} f_{up} \tag{7}$$

where f_{EW} and f_{IW} are the extracellular and intracellular water fractions, f_{NL} and f_{NP} are the neutral lipid and neutral phospholipid fractions, K_{aPR} is the extracellular protein association term, $[PR]_T$ is the tissue extracellular protein term, and X and Y are ionization terms.

For acids, the ionization terms were defined as:

$$X = 1 + 10^{(pH_t - pK_a)} \tag{8}$$

$$Y = 1 + 10^{(pH_p - pK_a)} \tag{9}$$

For weak bases, the ionization terms were defined as:

$$X = 1 + 10^{(pK_a - pH_t)} \tag{10}$$

$$Y = 1 + 10^{(pK_a - pH_p)} \tag{11}$$

For strong bases, the Rodgers formulation additionally included acidic phospholipid binding [2]:

$$K_{pu} = f_{EW} + \frac{X}{Y} f_{IW} + \frac{P f_{NL} + (0.3P + 0.7) f_{NP}}{Y} + \frac{K_{aAP} [AP^-]_T \, 10^{(pK_a - pH_t)}}{Y} \tag{12}$$

where K_{aAP} is the acidic phospholipid association term and $[AP^-]_T$ is the tissue acidic phospholipid term. The corresponding tissue-to-plasma partition coefficient was then calculated as:

$$K_p = K_{pu} f_{up} \tag{13}$$

For strong bases, the ionization terms were defined as:

$$X = 1 + 10^{(pK_a - pH_t)} \tag{14}$$

$$Y = 1 + 10^{(pK_a - pH_p)} \tag{15}$$

The acidic phospholipid affinity term was estimated from blood-cell partitioning logic according to:

$$K_{pu,BC} = \frac{B{:}P + H - 1}{H f_{up}} \tag{16}$$

where $B : P$ is the blood-to-plasma ratio, H is the hematocrit, and $K_{pu,BC}$ is the unbound blood-cell-to-plasma partition coefficient. In the present workflow, compounds with $pK_a < 7$ were treated as weak bases and compounds with $pK_a \geq 7$ as strong bases. This implementation follows the Rodgers framework for moderate-to-strong bases together with its extension to acids and weak bases [2, 3].

Schmitt

The Schmitt model was implemented as an interstitial/cellular tissue-composition framework. Tissue-to-plasma partitioning was expressed as [4]:

$$K_{t:p} = \left(\frac{F_{int}}{f_{u,int}} + \frac{F_{cell}}{f_{u,cell}} \right) f_{u,p} \tag{17}$$

where F_{int} and F_{cell} are the interstitial and cellular volume fractions, and $f_{u,int}$, $f_{u,cell}$, and $f_{u,p}$ are the unbound fractions in interstitial space, cellular space, and plasma, respectively.

The interstitial unbound fraction was derived from plasma binding as:

$$\frac{1}{f_{u,int}} = F_{W,int} + \frac{F_{P,int}}{F_{P,pl}} \left(\frac{1}{f_{u,p}} - F_{W,pl} \right) \tag{18}$$

where $F_{W,int}$ and $F_{W,pl}$ are the water fractions in interstitium and plasma, and $F_{P,int}$ and $F_{P,pl}$ are the corresponding protein fractions.

The cellular unbound fraction was calculated from the summed contributions of water, neutral lipid, neutral phospholipid, acidic phospholipid, and protein:

$$\frac{1}{f_{u,cell}} = F_W + K_{nL} F_{nL} + K_{nPL} F_{nPL} + K_{aPL} F_{aPL} + K_P F_P \tag{19}$$

where F_W, F_{nL}, F_{nPL}, F_{aPL}, and F_P are the respective cellular fractions, and K_{nL}, K_{nPL}, K_{aPL}, and K_P are the corresponding partition coefficients.

When direct neutral phospholipid partitioning data were unavailable, K_{nPL} was estimated from lipophilicity as:

$$K_{nPL} = 10^{\log P} \tag{20}$$

following tissue-composition framework proposed by Schmitt in [4].

Methods

Human heart physiological and tissue-composition parameters required for implementation of the mechanistic models were taken from the original model publications and incorporated as fixed values within the unified workflow [1–4]. Parameters representing physiological conditions are listed in Online Resource 2, and the overall workflow is shown in Fig. 1.

Data curation

Training data were drawn from a curated database of publicly available $K_{p,heart}$ values [14]. Only whole-heart measurements were retained, whereas records referring to specific anatomical heart regions were excluded. Subject-level entries were then aggregated so that each drug contributed a single record defined by the median $K_{p,\mathrm{heart}}$ together with its observed range (minimum–maximum). As steady-state conditions were rarely explicitly reported in the original studies, all observations were treated under a steady-state assumption for the purpose of consistent comparative model development. Each compound was assigned an isomeric SMILES string using the PubChem Identifier Exchange Service and subsequently queried in ADMETlab 3.0 [15]. From this source, routinely obtainable parameters relevant to drug distribution were assembled together with the core inputs required by the mechanistic calculators [16, 17]. Accordingly, the compound-level physicochemical, ADME, and pharmacokinetic descriptors used in the workflow were computationally predicted or database-derived descriptors obtained through ADMETlab 3.0, rather than systematically measured experimental input values. These included predicted descriptors describing lipophilicity and solubility (e.g., clogP, clogD at physiological pH, and clogS), ionization (acidic and basic pKa), plasma unbound fraction, and empirical distribution metrics such as predicted steady-state volume of distribution ($V_{d,ss}$). In addition, size- and polarity-related descriptors associated with tissue partitioning were included, such as molecular weight, topological polar surface area (TPSA), formal charge, numbers of hydrogen-bond acceptors and donors, rotatable bond count, ring count, F_{sp^3}, and related structural descriptors [18, 19]. Collectively, these features captured compound properties relevant to both the empirical selector and the mechanistic models. Records lacking the information required for mechanistic prediction or selector-model training were excluded from the dataset. The complete list of ADMETlab 3.0 descriptors used for model training, cross-validation dataset construction, and prospective tafamidis $K_{p,\mathrm{heart}}$ prediction is provided in the supplementary materials (Online Resource 1). Blood-to-plasma partition ratios for each drug were predicted with built-in Simcyp v25 predictor, using logP, pKa, fu with assumed plasma pH of 7.4 and haematocrit of 45% [20].

Model selector

For each compound, $K_{p,\mathrm{heart}}$ was calculated using the three mechanistic models described above. These predicted values were then compared with the observed median $K_{p,\mathrm{heart}}$.

The model-selection problem was formulated as a regression-based selection task. Instead of directly assigning compounds to mechanistic-method classes, separate regression models were trained to estimate the expected log-scale prediction error associated with each mechanistic approach. For each compound and mechanistic model m, the absolute log-error was defined as:

$$e_m = |log\,(Kp, m + \epsilon\,) - log\,(Kp, obs + \epsilon\,)| \tag{21}$$

where $K_{p,m}$ denotes the partition coefficient predicted by model m, $K_{p,\mathrm{obs}}$ denotes the observed median $K_{p,\mathrm{heart}}$, and $\epsilon = 10^{-12}$ is a small constant introduced for numerical stability.

The selector was trained to predict $\widehat{e}_{PT}$, $\widehat{e}_{RR}$, and $\widehat{e}_{Sch}$ from compound descriptors, and the final mechanistic method was chosen as the one associated with the lowest predicted error:

$$\widehat{m} = \underset{m}{\mathrm{argmin}}\widehat{e_m} \tag{22}$$

This formulation preserves interpretability, because the mechanistic equations remain unchanged, while allowing data-driven adjustment for systematic method-specific differences in predictive performance.

Cross-validation and preprocessing

Repeated stratified 5-fold cross-validation with 10 repetitions was used for internal training and evaluation of the selector. Stratification was based on the oracle class, defined for each compound as the mechanistic method with the lowest observed absolute log-error. Within each split, preprocessing was performed using the training fold only and then transferred unchanged to the corresponding test fold. Numeric predictors were imputed using training-fold medians, whereas categorical predictors were encoded using training-fold factor levels, with unseen levels assigned to an "Unknown" category.

Selector backends

Four regression backends were evaluated for predicting method-specific log-errors: random forest (RF), extreme gradient boosting (XGBoost), ridge regression (RIDGE), and ordinary linear regression (LM).

Random forest

Random forest is an ensemble learning method that aggregates predictions from multiple decision trees trained on

bootstrap-resampled datasets. At each split, a random subset of predictors is considered, reducing correlation between trees and improving generalization. This approach captures nonlinear relationships and high-order interactions without requiring explicit feature engineering and is generally robust to mixed descriptor scales and heterogeneous predictor sets. In the present study, random forest was implemented in regression mode to predict the expected log-error associated with each mechanistic K_p method [21].

Extreme gradient boosting

XGBoost is a gradient boosting algorithm that sequentially constructs decision trees to minimize a specified loss function, here the squared error loss. Each new tree is fitted to the residual structure not captured by the preceding ensemble, allowing efficient representation of complex nonlinear patterns. Regularization, shrinkage, and row- and column-subsampling contribute to improved generalization and reduced overfitting risk. In the present workflow, XGBoost was implemented in regression mode to predict method-specific log-errors from the encoded descriptor matrix [22].

Ridge regression

Ridge regression is a linear model with L_2-regularization, obtained by minimizing squared error together with a penalty on coefficient magnitude. This penalization stabilizes estimation in the presence of correlated predictors and limits variance inflation in descriptor-rich settings. In the current study, predictors were standardized before model fitting, and ridge regression served as a regularized linear baseline for method-specific error prediction [23].

Ordinary linear regression (LM)

Ordinary linear regression was included as a transparent unregularized baseline for method-specific error prediction. The model estimates a linear relationship between the encoded compound descriptors and the observed log-scale prediction error for each mechanistic K_p method by minimizing the residual sum of squares. This use is consistent with previous applications of linear-model frameworks to characterize predictive errors of complex mechanistic models as a function of system descriptors, although in the present workflow the predicted errors were further used to support selection among competing mechanistic K_p methods. In contrast to ridge regression, LM does not apply coefficient shrinkage and is therefore more sensitive to multicollinearity, high-dimensional descriptor sets, and influential observations. However, its inclusion provides a simple and interpretable reference model for assessing whether more flexible machine-learning approaches, such as random forest or XGBoost, provide added value beyond a purely linear descriptor–error relationship. In the present workflow, LM was implemented in regression mode to predict method-specific log-errors from the encoded descriptor matrix [24].

Evaluation strategy

Selector development and backend comparison were conducted using repeated stratified five-fold cross-validation with 10 repetitions. Stratification was based on the oracle class, defined retrospectively for each compound as the mechanistic method associated with the lowest observed absolute log-error. The oracle was used exclusively as a reference for evaluating selector agreement and was not available prospectively.

In each split, models were trained on four folds and evaluated on the remaining fold. The term “evaluation fold” is used here to distinguish these cross-validation evaluations from an independent final test set, which was not used in the present analysis. Within each split, preprocessing was performed using the training fold only and then transferred unchanged to the corresponding evaluation fold in order to prevent information leakage. Numeric predictors were imputed using training-fold medians, whereas categorical predictors were encoded using factor levels defined in the training fold, with previously unseen levels mapped to an “Unknown” category.

Out-of-fold (OOF) predictions were then pooled across all repeated cross-validation runs. An OOF prediction was defined as a prediction generated for a compound while that compound was held out from model training and included only in the evaluation fold. Thus, pooled OOF performance summarizes held-out predictions across all folds and repetitions and provides an aggregate cross-validation-based estimate of selector performance.

Selector performance was evaluated by comparing the predicted best mechanistic method with the oracle class across all cross-validation evaluation folds. The primary decision metric was balanced accuracy, defined as the mean recall across the three oracle classes: Poulin–Theil, Rodgers–Rowland, and Schmitt. For a given oracle class, recall was calculated as the proportion of compounds assigned to that oracle class for which the selector correctly identified the same mechanistic method. This metric was selected to reduce sensitivity to oracle-class imbalance.

Backend ranking was based primarily on median cross-validation balanced accuracy against the oracle, with median Cohen’s κ and median logMAE used as secondary ranking criteria. After backend selection, the chosen random forest selector was refit on the full labelled dataset and applied to tafamidis and acoramidis as an illustrative prospective use case.

As an illustrative prospective use case, tafamidis $K_{p,\mathrm{heart}}$ estimates obtained with Poulin–Theil, Rodgers–Rowland, and Schmitt were separately implemented in a previously developed whole-body tafamidis PBPK model [25]. The simulations were performed under the same dosing scenario, and systemic plasma and heart exposures were summarized over the day-7 dosing interval from 144 to 168 h after treatment initiation. The heart compartment was treated as distributive only, without intrinsic clearance.

Additional selector-performance metrics included overall selection accuracy, Cohen's κ, confusion matrices, and a binomial test against a random-choice baseline of 1/3. Overall accuracy quantified the proportion of compounds for which the selector chose the same mechanistic model as the oracle [26].

Cohen's κ quantified chance-corrected agreement between the selector and the oracle [27].

Confusion matrices were used to summarize class-specific agreement and misclassification structure [28]. The binomial test assessed whether the observed selection accuracy exceeded that expected under random choice among three mechanistic models.

To quantify numerical agreement between selected and observed $K_{p,\mathrm{heart}}$ values, regression-based performance measures were also calculated, including log-scale mean absolute error (logMAE), mean absolute error (MAE), root mean squared error (RMSE), and geometric mean fold error (GMFE). These metrics were calculated for the selector output and, for comparison, for each individual mechanistic model and for the oracle assignment.

Predicted–observed agreement was further assessed using the concordance correlation coefficient (CCC) calculated on natural-log-transformed $K_{p,\mathrm{heart}}$ values and reported as logCCC. This metric evaluates concordance with the line of identity, rather than correlation alone, and therefore captures both the strength of the predicted–observed association and systematic deviation from perfect agreement.

For new compounds, the workflow remains unchanged: three $K_{p,\mathrm{heart}}$ values are first calculated using the mechanistic models under fixed physiological assumptions, after which the trained selector identifies the method predicted to yield the lowest error and reports the corresponding $K_{p,\mathrm{heart}}$ together with all three mechanistic estimates. In this way, the framework is intended to support and standardize mechanistic K_p-method selection for further mechanistic modelling of drugs' pharmacokinetics, rather than relying solely on modeler intuition. All calculations and analyses were performed in R version 4.4.2 (R Core Team, 2024) using readr, dplyr, tidyr, randomForest, xgboost, ggplot2, and a lightweight in-house ridge implementation [29]

Results

After filtering to compounds with available observed $K_{p,\mathrm{heart}}$ values, valid predictions from all three mechanistic models, and calculable corresponding error terms, 55 compounds were retained for selector development and internal evaluation. Oracle-class distribution in the filtered dataset was strongly imbalanced, with 44 compounds assigned to Poulin–Theil, 9 to Rodgers–Rowland, and 2 to Schmitt.

Selector performance was evaluated using repeated stratified 5-fold cross-validation with 10 repetitions, and fold-level performance is summarized in Table 1. Backend differences were modest, and no algorithm showed strong oracle-class recovery. Random forest achieved the highest median balanced accuracy and was therefore selected according to the predefined backend-ranking rule, although Cohen's κ indicated limited chance-corrected agreement. The remaining backends showed either lower balanced accuracy or greater variability across evaluation folds.

In terms of numerical agreement between predicted and observed $K_{p,\mathrm{heart}}$, median test logMAE was lowest for random forest (0.898), followed by XGB (0.962), RIDGE (1.09), and LM (1.25). Based on the predefined ranking rule, which prioritized median test balanced accuracy against the oracle and used median test Cohen's κ as a secondary criterion, random forest was selected as the final backend for prospective application.

A pooled out-of-fold comparison between the selected RF-based workflow and the three mechanistic baselines is shown in Table 2. In this pooled OOF analysis, the RF selector achieved logMAE 0.900, GMFE 2.46, and logCCC 0.377 across 550 held-out predictions. The

Table 1 Cross-validation performance of regression-based selector backends. Performance values are presented as median [minimum–maximum] of fold-level metrics across 50 held-out evaluation folds generated from 10 repetitions of 5-fold cross-validation

Backend	Balanced accuracy*	Cohen's κ	Accuracy	logMAE	GMFE	Median *N*
Random forest	0.500 [0.333–0.500]	0.000 [−0.222–0.163]	0.818 [0.636–0.889]	0.898 [0.432–1.456]	2.456 [1.541–4.288]	11
Linear regression	0.454 [0.125–0.833]	0.018 [−0.351–0.593]	0.545 [0.222–0.889]	1.252 [0.378–2.937]	3.497 [1.460–18.865]	11
Ridge regression	0.435 [0.148–0.944]	0.060 [−0.305–0.765]	0.556 [0.250–0.917]	1.089 [0.387–2.693]	2.971 [1.473–14.780]	10
XGBoost	0.389 [0.148–0.750]	−0.116 [−0.352–0.621]	0.727 [0.333–0.909]	0.962 [0.466–2.082]	2.617 [1.594–8.024]	11

N denotes the median number of compounds in the held-out evaluation fold

* Balanced accuracy was calculated as the mean recall across oracle classes and was predefined as the primary backend-selection criterion

Table 2 Pooled out-of-fold numerical performance of the selected random forest-based selector, the three mechanistic baseline models, and the retrospective oracle benchmark

Model	logMAE	MAE	RMSE	GMFE	logCCC	*N*
Random forest selector	0.900	4.41	9.51	2.46	0.377	550
Poulin–Theil	0.875	4.37	9.50	2.40	0.402	550
Rodgers–Rowland	1.578	39816.22	294567.00	4.85	0.264	550
Schmitt	5.089	5.96	11.06	162.17	−0.064	550
Oracle	0.703	3.91	8.95	2.02	0.588	550

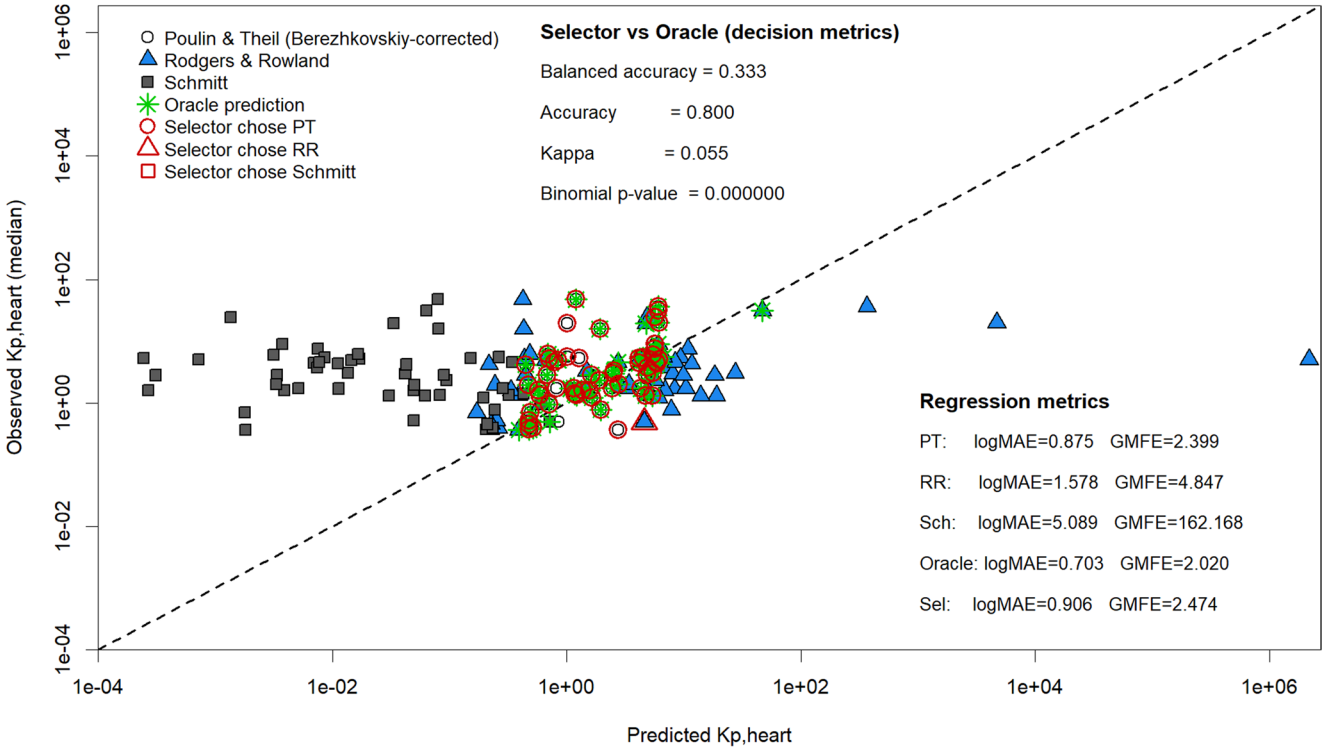


Fig. 2 Predicted versus observed $K_{p,heart}$ values using one representative selector assignment per compound, derived from repeated out-of-fold predictions across stratified 5-fold cross-validation (10 repetitions). Selector-assigned predictions are highlighted in red, and oracle predictions are shown in green

strongest standalone mechanistic baseline was Poulin–Theil, with slightly better numerical agreement and concordance (logMAE 0.875, GMFE 2.40, logCCC 0.402). Rodgers–Rowland showed weaker agreement (logMAE 1.58, GMFE 4.85, logCCC 0.264), whereas Schmitt underperformed markedly (logMAE 5.09, GMFE 162.17, logCCC −0.064). As expected, the unattainable oracle benchmark remained best overall (logMAE 0.703, GMFE 2.02, logCCC 0.588).

After internal backend selection, the random forest selector was refit on the full internal dataset and applied prospectively to the transthyretin stabilisers. For both tafamidis and acoramidis, the selector chose Poulin–Theil as the preferred mechanistic method. The resulting selected $K_{p,\mathrm{heart}}$ estimates were 5.86 for tafamidis and 0.593 for acoramidis.

Selector choices were dominated by Poulin–Theil, with occasional Rodgers–Rowland and rare Schmitt selections, reflecting the skewed distribution of oracle-best labels Fig. 2 presents all three mechanistic predictions together with one representative selector-assigned value per compound, derived from repeated cross-validation evaluations.

In the tafamidis PBPK use case, the three alternative $K_{p,\mathrm{heart}}$ estimates produced markedly different predicted cardiac exposures during the day-7 dosing interval (144–168 h), while systemic plasma AUC_{τ} remained similar across simulations. This pattern is consistent with the heart being modelled as a distribution-only compartment, where $K_{p,\mathrm{heart}}$ directly determines tissue partitioning but does not introduce local clearance. A comparison of AUC values obtained using different $K_{p,\mathrm{heart}}$ estimates is shown in Fig. 3.

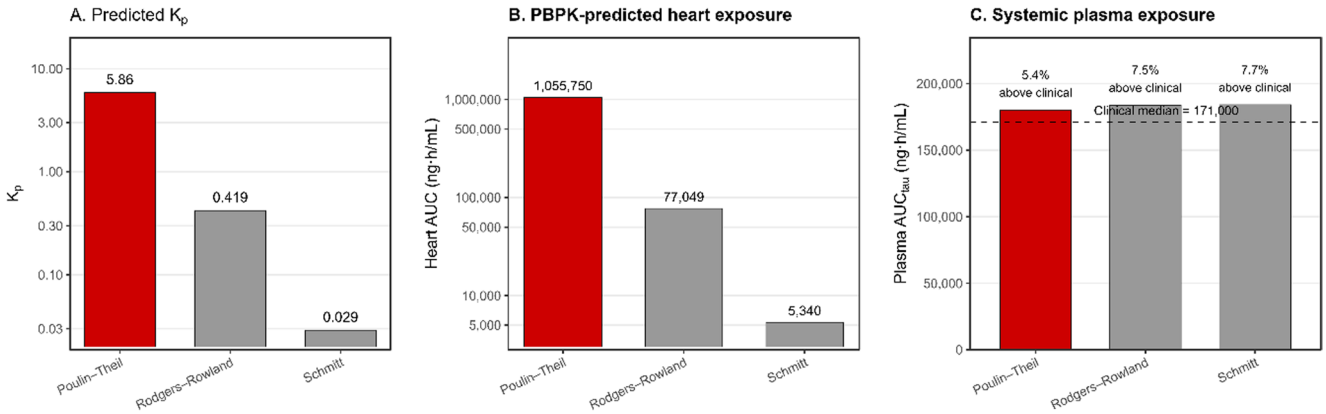


Fig. 3 Prospective PBPK use-case analysis of alternative $K_{p,heart}$ estimates. Poulin–Theil, Rodgers–Rowland, and Schmitt $K_{p,heart}$ values were propagated into the tafamidis PBPK model, and AUC values were calculated over the day-7 dosing interval from 144 to 168 h after treatment initiation. Panels show: (A) $K_{p,heart}$ values, (B) PBPK-predicted heart AUC_τ, and (C) systemic plasma AUC_τ. The dashed line in panel C denotes the observed clinical trial median plasma AUC_τ of 171,000 ng·h/mL

Discussion

All three models used for K_p prediction and evaluated in this study are based on physiologically and biochemically motivated principles and have previously been used for tissue-distribution estimation. However, determining which model should be preferred for predicting xenobiotic partitioning from plasma to a specific human organ remains unresolved. This difficulty reflects the scarcity of curated human tissue-partitioning data and the lack of a universally dominant mechanistic approach across chemical space. By relying on experimentally determined human $K_{p,\mathrm{heart}}$ values and restricting evaluation to the human heart setting, the present work provides a human-centered benchmark for method comparison.

A major consequence of the current workflow is that selector performance is interpreted through repeated internal cross-validation, rather than through a single train/test split. This is important because the earlier single-split framework tested by us proved unstable with respect to backend ranking. Under the repeated stratified cross-validation design, random forest emerged as the best-performing selector backend according to the predefined criterion, whereas the earlier split-based workflow had favoured different model. This indicates that backend choice was sensitive to resampling design and confirms that repeated internal evaluation is more appropriate for a dataset of this size.

The selector results should be interpreted primarily through balanced accuracy and Cohen's κ, rather than raw accuracy alone. The median balanced accuracy of 0.500 for random forest means that, after averaging recall equally across the three mechanistic classes, the selector correctly recovered about half of the oracle-assigned compounds. In the present imbalanced setting, this is more informative than overall accuracy because it reduces the apparent advantage of strategies dominated by the majority PT class. By contrast, the relatively high median random forest accuracy of 0.818 could easily be overinterpreted if considered alone, because a selector that frequently defaults to PT will appear accurate when PT dominates the oracle distribution.

This interpretation is reinforced by the κ values. Cohen's κ measures agreement beyond that expected by chance, taking class prevalence into account. The fact that median κ values remained low for all backends, including 0 for random forest, indicates that much of the apparent agreement was still driven by oracle-class imbalance. In practical terms, random forest performed best under the predefined class-balanced ranking criterion, but its decisions remained strongly constrained by the predominance of the PT class. The slightly higher median κ values for RIDGE and LM suggest some additional chance-corrected structure, but not enough to overcome their weaker balanced accuracy.

The oracle benchmark should itself be interpreted cautiously. It is not a prospective predictor, but a retrospective upper bound that identifies which mechanistic method happened to be least wrong after the observed $K_{p,\mathrm{heart}}$ value is already known. Accordingly, the gap between the selector and the oracle reflects remaining room for improvement in method selection, not a realistic target that can be achieved directly in practice.

The pooled OOF comparison clarifies the practical role of the selector. The random forest selector achieved logMAE 0.900 and GMFE 2.46, meaning that selected $K_{p,\mathrm{heart}}$ values differed from observations by about 2.5-fold on average. However, the strongest standalone mechanistic baseline, Poulin–Theil, performed slightly better numerically, with logMAE 0.875 and GMFE 2.40. Thus, under the present dataset and implementation assumptions, the selector did not materially outperform the best baseline in terms of absolute numerical error.

This should not be interpreted as a failure of the approach. Rather, it defines its actual value in the current setting. The selector's main contribution is not a large reduction in $K_{p,\mathrm{heart}}$ error relative to the strongest mechanistic model, but the provision of a transparent, reproducible, and data-driven rule for method choice. When several mechanistic methods are plausible, the selector formalizes a decision that would otherwise often rely on modeler intuition. In that sense, its practical role is to support and standardize mechanistic method selection rather than to replace the mechanistic models themselves.

It is also instructive to contextualize the present framework within broader efforts to optimize tissue distribution predictions. For instance, when preclinical pharmacokinetic data are available, several groups have demonstrated that optimizing simplified PBPK models or deriving empirical tissue scaling factors from animal models (e.g., rats, dogs, and monkeys) can significantly improve the accuracy of human volume of distribution and drug disposition estimates [30, 31]. Other strategies bypass in vivo animal testing altogether by utilizing protein binding assays to correct theoretically calculated partition coefficients, facilitating earlier human pharmacokinetic predictions [32]. Furthermore, pure machine learning and quantitative structure-activity relationship (QSAR) methods, such as repeated random forest algorithms, have been effectively deployed to directly predict tissue K_p values and handle missing data matrices across multiple organs [33]. While these approaches emphasize cross-species translation, in vitro-to-in vivo extrapolation, or direct empirical parameter prediction, the workflow proposed here occupies a distinct space. By restricting the learning environment entirely to human data and keeping the underlying mechanistic equations intact, our approach uses pattern recognition not to predict the partitioning endpoint directly, but exclusively to formalize the decision-making process among competing mechanistic models in a target-specific context.

The strong dominance of Poulin–Theil in the oracle distribution is itself informative. It suggests that, under the fixed physiological assumptions used here, PT is a strong default model for human whole-heart K_p prediction. This conclusion is also consistent with the prospective stabiliser application, in which the selector chose PT for both tafamidis and acoramidis. However, this same dominance also constrains learning: when 44 of 55 compounds are oracle-assigned to PT, there is limited information available to define the conditions under which Rodgers–Rowland or Schmitt become preferable.

The PBPK example should therefore be interpreted as a translational illustration of the relevance of $K_{p,\mathrm{heart}}$ selection, rather than as an independent validation of the selector. In a whole-body PBPK model, tissue-to-plasma partition coefficients determine how systemic exposure is distributed across organs; consequently, uncertainty in $K_{p,\mathrm{heart}}$ propagates directly into predicted cardiac exposure. This is particularly relevant for tafamidis, where cardiac exposure is pharmacologically meaningful in the context of transthyretin amyloid cardiomyopathy, even though the present heart compartment does not include local clearance or a cardiac pharmacodynamic feedback mechanism. Thus, the PBPK application demonstrates how method choice at the K_p prediction stage can affect downstream organ-level exposure estimates used in mechanistic pharmacokinetic modelling.

Several limitations affect interpretation. First, the benchmark dataset remains small, so repeated cross-validation improves internal use of the available data but does not eliminate uncertainty arising from limited sample size. This data scarcity is a common challenge for models relying strictly on public domain information. While highly predictive QSAR models for tissue distribution have been successfully developed for other organs, such as the model for unbound brain partitioning by Dolgikh et al., which leveraged a massive, in-house proprietary dataset of approximately 1000 compounds, researchers modeling human cardiac exposure without access to closed pharmaceutical databases must operate under severe data constraints [34]. In such data-poor environments, our hybrid approach of using machine learning merely to select among established mechanistic models, rather than fitting a purely empirical model from scratch, provides a necessary and pragmatic solution.

Second, class imbalance is substantial, which makes minority-class recall estimates unstable and depresses chance-corrected agreement.

Third, physiological and tissue-composition values were fixed across the mechanistic calculators according to their original publications. As elegantly demonstrated by Utsey et al., the use of differing tissue composition data can confound the comparative evaluation of mechanistic K_p models and significantly impact predictive performance [35]. While Utsey et al. proposed a standardized tissue composition to control for this variability across multiple organs, we elected to evaluate each model using the specific composition parameters reported by its original authors. Consequently, models that are particularly sensitive to specific inputs, such as phospholipid terms, pH partitioning, extracellular protein assumptions, or membrane-partition surrogates, may be inherently advantaged or disadvantaged by their native physiological assumptions.

Fourth, the prospective application to stabilisers demonstrates workflow use, but it does not constitute external validation.

The especially weak performance of Schmitt warrants cautious interpretation. In principle, Schmitt is the most explicitly compositional of the three frameworks. In the

present implementation, however, it depends on several practically necessary assumptions, including treatment of the acidic phospholipid contribution and the use of lipophilicity-derived surrogate input when direct neutral phospholipid partition data are unavailable. Poor performance under these conditions therefore does not necessarily invalidate the Schmitt framework itself but rather suggests that it may be particularly sensitive to parameter quality and implementation choices in the human heart setting.

The conceptual strength of the proposed workflow lies in its hybrid structure. The mechanistic equations remain intact and physiologically interpretable, while a data-driven layer provides empirical guidance on method selection. This preserves mechanistic transparency while reducing reliance on subjective method choice. In the present dataset, that strategy produced meaningful internal discrimination beyond uninformed guessing, but it did not overcome the dominance of the strongest baseline to a large numerical extent.

Overall, when multiple mechanistic tissue-partitioning methods are available, a regression-based selector offers a transparent and reproducible rule for choosing among them. Under the current human whole-heart benchmark, Poulin–Theil remained the strongest standalone baseline and random forest the best empirical selector backend. The principal limitation is therefore not algorithmic complexity, but benchmark structure: the limited size and strong class imbalance of the available human $K_{p,\mathrm{heart}}$ dataset. Expansion of the human benchmark dataset, together with future incorporation of physiological variability and improved descriptor representation of phospholipid- and pH-dependent partitioning, represents the most direct path toward improving selector robustness and broader utility.

Conclusion

In conclusion, the proposed hybrid mechanistic–AI/ML workflow provides a reproducible framework for selecting heart K_p prediction methods on a compound-specific basis. Although Poulin–Theil remained the strongest standalone model in the current analysis, the selector formalized method choice and demonstrated how uncertainty in $K_{p,heart}$ estimation can propagate into PBPK-predicted cardiac exposure. Further improvement of this approach will require broader, higher-quality, and more diverse human heart distribution data to better characterize the conditions under which different mechanistic models are preferable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10928-026-10057-4.

Author contributions S.U. and S.P. conceptualized the study and developed the methodological framework. S.U. performed data curation, implemented the mechanistic Kp calculations, developed and evaluated the AI/ML-based selector, performed the formal analysis, and prepared the figures and tables. S.U. wrote the main manuscript text. S.P. supervised the study and acquired funding. B.L., M.J., B.W., and S.P. critically reviewed and edited the manuscript. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Funding This research was funded by the European Union within the VITAL Horizon Europe project (Grant number 101136728).

Data availability The dataset and descriptor files used in this study are provided as supplementary files. These include the curated human heart-to-plasma partition coefficient dataset, compound descriptors used for selector development, and files required to reproduce the mechanistic Kp calculations and AI/ML-based model-selection analysis.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Poulin P, Theil F (Jan. 2000) A Priori Prediction of Tissue:Plasma Partition Coefficients of Drugs to Facilitate the Use of Physiologically-Based Pharmacokinetic Models in Drug Discovery. J Pharm Sci 89(1):16–35. https://doi.org/10.1002/(SICI)1520-6017(200001)89:1%3C16::AID-JPS3%3E3.0.CO;2-E
2. Rodgers T, Leahy D, Rowland M (2005) Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases. J Pharm Sci 94(6):1259–1276. https://doi.org/10.1002/jps.20322
3. Rodgers T, Rowland M (2006) Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions. J Pharm Sci 95(6):1238–1257. https://doi.org/10.1002/jps.20502
4. Schmitt W (Mar. 2008) General approach for the calculation of tissue to plasma partition coefficients. Toxicol Vitro 22(2):457–467. https://doi.org/10.1016/j.tiv.2007.09.010
5. Korzekwa K, Nagar S (2017) Drug distribution part 2. Predicting volume of distribution from plasma protein binding and membrane partitioning. Pharm Res 34(3):544–551. https://doi.org/10.1007/s11095-016-2086-y

6. Jansson R, Bredberg U, Ashton M (2008) Prediction of drug tissue to plasma concentration ratios using a measured volume of distribution in combination with lipophilicity. J Pharm Sci 97(6):2324–2339. https://doi.org/10.1002/jps.21130
7. Arundel PA (1997) A Multi-Compartmental Model Generally Applicable to Physiologically-Based Pharmacokinetics, *IFAC Proc. Vol.*, vol. 30, no. 2, pp. 129–133, Mar. https://doi.org/10.1016/S1474-6670(17)44557-5
8. Jones RD et al (2011) PhRMA CPCDC initiative on predictive models of human pharmacokinetics, part 2: comparative assessment of prediction methods of human volume of distribution. J Pharm Sci 100(10):4074–4089. https://doi.org/10.1002/jps.22553
9. Yun YE, Cotton CA, Edginton AN (2014) Development of a decision tree to classify the most accurate tissue-specific tissue to plasma partition coefficient algorithm for a given compound. J Pharmacokinet Pharmacodyn 41(1):1–14. https://doi.org/10.1007/s10928-013-9342-0
10. Green D (2018) Physiologically Based Pharmacokinetic Analyses — Format and Content Guidance for Industry, Aug. [Online]. Available: https://www.fda.gov/files/drugs/published/Physiologically-Based-Pharmacokinetic-Analyses-%E2%80%94-Format-and-Content-Guidance-for-Industry.pdf?utm_source=chatgpt.com
11. Ulaszek S, Lisowski B, Wiśniowska B, Polak S (2026) Exploring Tafamidis Effects Through PBPK-QSP Modelling, Feb. 09, Medicine and Pharmacology. https://doi.org/10.20944/preprints202602.0585.v1
12. Berezhkovskiy LM (2004) Volume of distribution at steady state for a linear pharmacokinetic system with peripheral elimination. J Pharm Sci 93(6):1628–1640. https://doi.org/10.1002/jps.20073
13. Poulin P, Theil FP (2002) Prediction of pharmacokinetics prior to in vivo studies. II. Generic physiologically based pharmacokinetic models of drug disposition. J Pharm Sci 91(5):1358–1370. https://doi.org/10.1002/jps.10128
14. Tylutki Z, Polak S (2015) Plasma vs heart tissue concentration in humans – literature data analysis of drugs distribution. Biopharm Drug Dispos 36(6):337–351. https://doi.org/10.1002/bdd.1944
15. Fu L et al (2024) ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res 52(W1):W422–W431. https://doi.org/10.1093/nar/gkae236
16. Xiong G et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 49(W1):W5–W14. https://doi.org/10.1093/nar/gkab255
17. PubChem Identifier Exchange Service. Accessed: Sep. 08, 2025. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi
18. Honorio KM, Moda TL, Andricopulo AD (2013) Pharmacokinetic properties and in silico ADME modeling in drug discovery. Med Chem 9(2):163–176. https://doi.org/10.2174/1573406411309020002
19. Karlgren M, Bergström CAS (2015) How Physicochemical Properties of Drugs Affect Their Metabolism and Clearance. Nov. https://doi.org/10.1039/9781782622376-00001
20. Ezuruike U et al (2022) Guide to development of compound files for PBPK modeling in the Simcyp population-based simulator. CPT Pharmacomet Syst Pharmacol 11(7):805–821. https://doi.org/10.1002/psp4.12791
21. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
22. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. pp. 785–794. https://doi.org/10.1145/2939672.2939785
23. Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634
24. Knightes CD, Cyterski M (2005) Evaluating predictive errors of a complex environmental model using a general linear model and least square means. Ecol Model 186(3):366–374. https://doi.org/10.1016/j.ecolmodel.2005.01.034
25. Ulaszek S, Lisowski B, Wiśniowska B, Polak S (2026) Exploring tafamidis effects through PBPK–QSP modelling. Pharmaceutics. https://doi.org/10.3390/pharmaceutics18030367
26. Bishop CM (2006) Pattern recognition and machine learning. in Information science and statistics. Springer, New York
27. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20(1):37–46. https://doi.org/10.1177/001316446002000104
28. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. in Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7
29. Team RC (2024) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Oct 31. https://doi.org/10.32614/R.manuals
30. Yau E, Gertz M, Ogungbenro K, Aarons L, Olivares-Morales A (2023) A 'middle-out approach' for the prediction of human drug disposition from preclinical data using simplified physiologically based pharmacokinetic (PBPK) models. CPT Pharmacomet Syst Pharmacol 12(3):346–359. https://doi.org/10.1002/psp4.12915
31. Ning M et al (Feb.2025) A cross-species assessment of in silico prediction methods of steady-state volume of distribution using Simcyp simulators. J Pharm Sci 114(2):1410–1422. https://doi.org/10.1016/j.xphs.2024.12.018
32. Mayumi K, Tachibana M, Yoshida M, Ohnishi S, Kanazu T, Hasegawa H (2020) The novel in vitro method to calculate tissue-to-plasma partition coefficient in humans for predicting pharmacokinetic profiles by physiologically-based pharmacokinetic model with high predictability. J Pharm Sci 109(7):2345–2355. https://doi.org/10.1016/j.xphs.2020.04.002
33. Handa K, Yoshimura S, Kageyama M, Iijima T (2024) Development of novel methods for QSAR modeling by machine learning repeatedly: a case study on drug distribution to each tissue. J Chem Inf Model 64(9):3662–3669. https://doi.org/10.1021/acs.jcim.4c00046
34. Dolgikh E et al (Nov.2016) QSAR Model of Unbound Brain-to-Plasma Partition Coefficient, Kp, uu, brain: Incorporating P-glycoprotein Efflux as a Variable. J Chem Inf Model 56(11):2225–2233. https://doi.org/10.1021/acs.jcim.6b00229
35. Utsey K, Gastonguay MS, Russell S, Freling R, Riggs MM, Elmokadem A (2020) Quantification of the impact of partition coefficient prediction methods on physiologically based pharmacokinetic model output using a standardized tissue composition. Drug Metab Dispos 48(10):903–916. https://doi.org/10.1124/dmd.120.090498

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.