



OPEN

# Adapting physiologically-based pharmacokinetic models for machine learning applications

Sohaib Habiballah<sup>1</sup> & Brad Reisfeld<sup>1,2</sup>✉

Both machine learning and physiologically-based pharmacokinetic models are becoming essential components of the drug development process. Integrating the predictive capabilities of physiologically-based pharmacokinetic (PBPK) models within machine learning (ML) pipelines could offer significant benefits in improving the accuracy and scope of drug screening and evaluation procedures. Here, we describe the development and testing of a self-contained machine learning module capable of faithfully recapitulating summary pharmacokinetic (PK) parameters produced by a full PBPK model, given a set of input drug-specific and regimen-specific information. Because of its widespread use in characterizing the disposition of orally administered drugs, the PBPK model chosen to demonstrate the methodology was an open-source implementation of a state-of-the-art compartmental and transit model called *OpenCAT*. The model was tested for drug formulations spanning a large range of solubility and absorption characteristics, and was evaluated for concordance against predictions of *OpenCAT* and relevant experimental data. In general, the values predicted by the ML models were within 20% of those of the PBPK model across the range of drug and formulation properties. However, summary PK parameter predictions from both the ML model and full PBPK model were occasionally poor with respect to those derived from experiments, suggesting deficiencies in the underlying PBPK model.

Machine learning is increasingly used in many aspects of drug discovery and development<sup>1–3</sup>, with a common use being the virtual screening of chemical libraries. Many of the studies published in the literature have focused on advances in screening for bioactivity<sup>4–7</sup>, with many fewer centered on pharmacokinetics<sup>8–11</sup>. These latter studies have proven to be relatively effective across large chemical libraries, but do not generally have the predictive power of special-purpose tools like physiologically based pharmacokinetic (PBPK) models.

PBPK models, which incorporates anatomical, physiological, and biochemical relationships, are widely used in the drug development process and have proven useful in predicting drug ADME for a range of therapeutics<sup>12,13</sup>, including pediatric drugs<sup>14</sup>, biopharmaceutics<sup>15</sup>, generic drugs<sup>16</sup>, monoclonal antibodies<sup>17</sup>, and nanoparticles<sup>18</sup>. In the case of orally administered drugs, compartmental absorption and transit (CAT) PBPK models are often used. These models include many details of drug absorption, metabolism, and transport in the gastrointestinal tract, which is often divided into discrete segments, each of which may account for drug in various states, e.g., undissolved, dissolved, and absorbed into the enterocytes. Implementations of such models include the ACAT (advanced compartmental and transient) model<sup>19</sup>, part of the proprietary software GastroPlus<sup>®</sup>, and the *OpenCAT* model<sup>20,21</sup>, developed de novo using the open source software GNU MCSim<sup>22</sup>.

Usage of PBPK models, like ACAT or *OpenCAT*, requires the solution of (often complex) systems of differential equations that may make implementation difficult in certain workflows, like those common in machine learning pipelines<sup>23</sup>. One approach to address this limitations is to capture the essential features of PBPK simulation predictions within a trained ML model. Such a model could take as input drug-specific information and would output essential PK parameters as a stage in a multi-element analysis pipeline. The aim of this work was to explore the feasibility of this approach using *OpenCAT* as the underlying PBPK model.

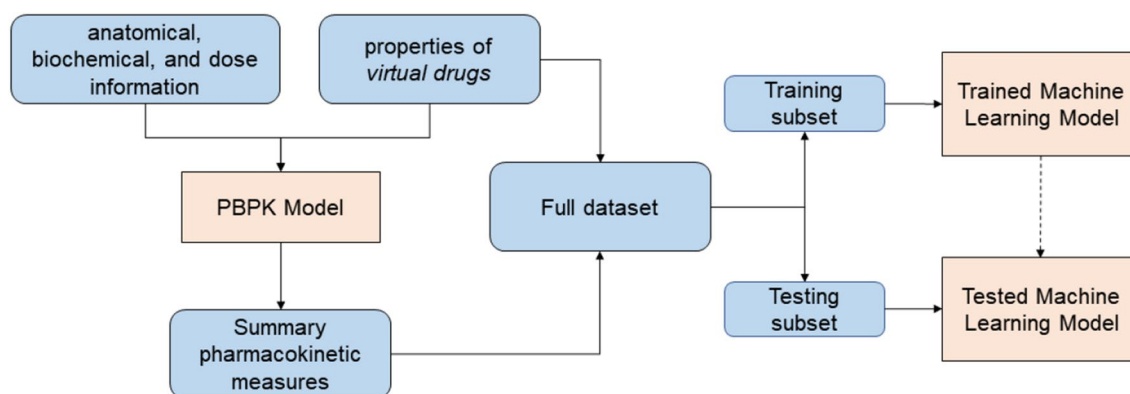
<sup>1</sup>Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO 80523-1301, USA. <sup>2</sup>School of Public Health, Colorado State University, Fort Collins, CO 80523-1612, USA. ✉email: brad.reisfeld@colostate.edu

## Methods

To achieve the project aim, a machine-learning model was developed to take as input drug-specific properties and produce as output various summary pharmacokinetic measures comparable to those generated by the full PBPK model. The development workflow comprised several steps: (i) Generating a dataset comprising properties of a large set of *virtual drugs*, (ii) Simulating the pharmacokinetics of each member of the dataset using the PBPK model and calculating summary pharmacokinetic measures, (iii) Developing and training machine learning models using the drug properties and summary PK measures, (iv) Critically evaluating the models in terms of their predictive capabilities. A schematic of the workflow is shown in Fig. 1.

**Dataset generation.** *Drug properties.* The aim of this part of the workflow was to generate a large set of *virtual drugs* whose properties spanned the four classes of the Biopharmaceutics Classification System (BCS)<sup>24</sup>, which organizes compounds into four categories: Class I—high permeability, high solubility; Class II—high permeability, low solubility; Class III—low permeability, high solubility; and Class IV—low permeability, low solubility. To generate these property sets, drug properties (molecular mass, molar volume, acidic dissociation constant, effective permeability, precipitation rate constant, drug solubility, particle radius, and drug density) and additional parameters influencing drug pharmacokinetics (the ratio of the drug unbound fraction over its partition coefficient ( $FuPC_i$ ), metabolic parameters, dose magnitude, and subject body weight) were taken from the literature<sup>25–31</sup> to establish realistic ranges for all measures (see Table 1). Assuming a uniform distribution for each of the ranges, a Monte Carlo sampling procedure was employed in which a given sample was constructed by drawing from each parameter distribution. In total, 15,000 *virtual drugs* were generated for use in subsequent steps in the workflow.

To assess whether the generated *virtual drugs* were representative of actual drugs, 40 drugs across all BCS classes were selected from ChEMBL<sup>32</sup> and PubChem<sup>33</sup> and their properties determined. Tests confirmed that 90% of the real drugs had at least one ‘close match’ in the *virtual drugs* dataset and all of them had at least one ‘moderate match’. In this analysis, ‘close match’ and ‘moderate match’ meant that all of the following properties



**Figure 1.** Elements and steps in the workflow.

Parameter name	Range	Units
Molecular mass	20–1700	$\frac{\text{g}}{\text{gmol}}$
Molar volume	30–1200	$\frac{\text{g}}{\text{gmol}}$
Acidic dissociation constant	0–14	–
Particle radius	5–50	$\mu\text{m}$
Drug density	0.1–0.7	$\frac{\text{g}}{\text{ml}}$
Drug solubility	0.001–1000	$\frac{\text{mg}}{\text{ml}}$
Precipitation rate constant	0.01–1	$\text{h}^{-1}$
Effective permeability of GI tract epithelia	0.015–8	$\frac{\text{cm}}{\text{s}}$
Unbound fraction over partition coefficient in compartment $i$ , $FuPC_i$	0.01–100	–
Subject's body mass	45–125	kg
Dose magnitude	$1-50 \times 10^3$	$\mu\text{mol}$
Metabolism rate constant, $K_m$	0.04–10	mM
Metabolism rate constant, $V_{max}$	$10^{-6}-10^{-2}$	$\frac{\mu\text{mol}}{(\text{min})(\text{mg})(\text{microsomal proteins})}$

**Table 1.** Ranges for property of interest determined by examining physical properties of actual drugs spanning all BCS classes.

of the *virtual drug* were within 15% and 50%, respectively, of those of an actual drug: molecular weight, density, molar volume, pKa, solubility, and effective permeability.

Though assigning drugs to a specific BCS classes is not always straightforward<sup>34,35</sup>, for the purpose of this work, classes were assigned based simply on specific thresholds of solubility and permeability<sup>24,36</sup>. Using this procedure, the library of virtual drugs comprised 6392 Class I, 2097 Class II, 4920 Class III, and 1591 Class IV drugs.

**Pharmacokinetic information.** For each member of the set of *virtual drugs*, a simulation was performed using the OpenCAT model, whose structure is shown in Fig. 2. These simulations produced predicted time-course concentrations of the chemical in each state (unreleased, undissolved, dissolved) and compartment. For the purpose of comparison to experimental values from the literature, the detailed simulation results were condensed into the summary pharmacokinetic metrics (SPKMs)  $C_{\max}/\text{dose}$ ,  $t_{\max}$ , and  $AUC/\text{dose}$  using the predicted PK information from the blood. SPKMs values were normalized using min-max scaling.

Finally, to create the full dataset, the properties of each virtual drug were combined with its corresponding normalized SPKMs.

**Model development.** Prior to model development, values were randomly selected from the full dataset (inputs + normalized SPKMs) to create training and testing subsets comprising 80% and 20% of the values, respectively. For each machine learning model, the drug properties and additional parameters noted earlier were used as *features*, while the normalized SPKMs were used as *labels*. Though a multivariate model could be constructed to predict all three normalized SPKMs, it was found that distinct models for each label (i.e., each normalized SPKM) consistently performed better in recapitulating target values.

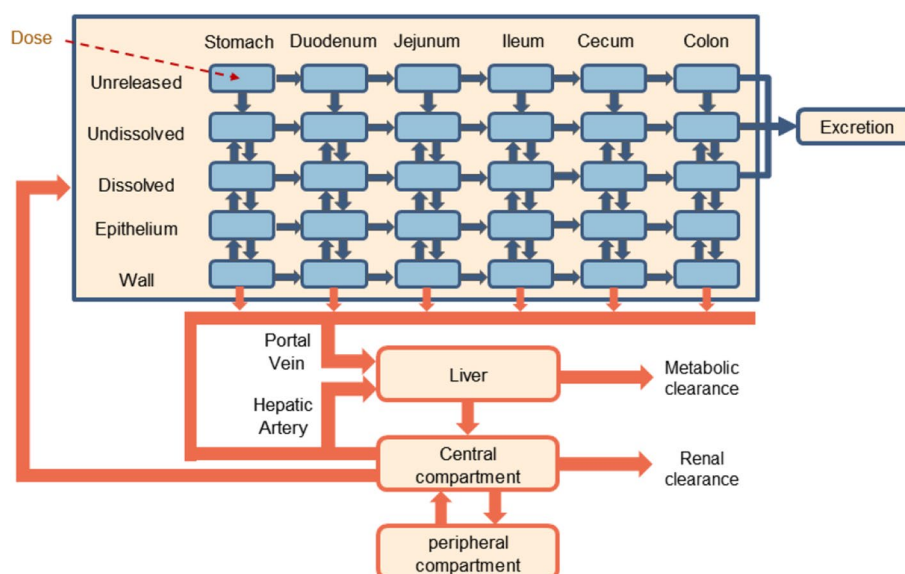
**Algorithm selection:** To inform the process of algorithm selection, two machine learning algorithms appropriate for regression analyses, were evaluated: random forest<sup>37</sup> and gradient boosting<sup>38,39</sup>. Preliminary evaluations indicated that both gradient boosting and random forest algorithm were equally suitable for this application, but because it proved to be more computationally efficient in cases of interest, random forest (RF) regression was selected.

**Model training:** Utilizing the training data set, the algorithm's hyperparameters were tuned by performing a grid search. In addition to hyperparameters, the number of trees were optimized to maximize the accuracy of the algorithm while minimizing overfitting. From this process, the number of trees was set to 150 for all models.

**Model evaluation.** Assessment metrics: To assess the performance of each ML-based model relative to those from the OpenCAT model, two metrics were used: (i) the relative error in predicted normalized SPKMs between the OpenCAT and ML models and (ii) the adjusted coefficient of determination across the entire set of *virtual drugs*, which is defined as

$$R^2(\text{adj}) = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1},$$

where  $n$  is the number of values in the data set,  $k$  is the number of independent features included, and  $R^2$  is the *unadjusted* coefficient of determination given by the conventional definition:  $R^2 = 1 - SSR/SST$ , where  $SSR$  is the sum of squares of the residuals and  $SST$  is the total sum of squares.



**Figure 2.** Structural overview of the OpenCAT PBPK model.

**Feature importance assessment:** A study was conducted to evaluate the influence of feature set (number of features and features selected) on the accuracy of the model. Subsets having fewer features than the full feature set (FFS) are known as reduced feature sets (RFS). The RFS were generated iteratively and utilized the feature importance score<sup>40</sup> for aggregation. The performance of each model was evaluated when trained using the FFS and all RFS. The *optimal* RFS was selected as the set with the highest  $R^2$ (adj) value relative to the OpenCAT predictions.

**Testing against experimental data:** To further test the models, the normalized SPKMs were compared to those from OpenCAT and from experimentally-measured values for ten specific drugs (see Table 2).

**Software.** Python<sup>60</sup> (v3.8) was used for general data processing. All machine learning simulations were conducted using scikit-learn (v1.1.3)<sup>61</sup>, and pharmacokinetic simulations were implemented and run using GNU MCSim<sup>22</sup> (v6.1.0).

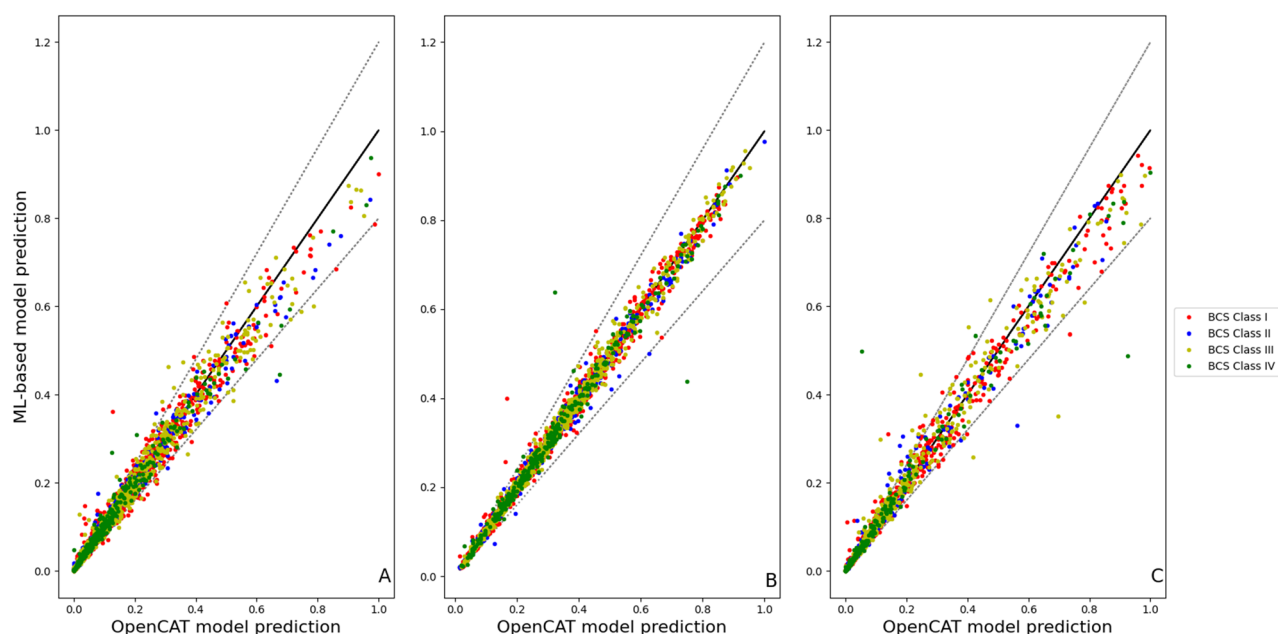
## Results

**Model verification. Full feature set model.** Figure 3 shows a comparison of the predictions of the full feature set (FFS) models with those from OpenCAT across the testing subset of *virtual drugs*. The  $R^2$ (adj) values for the individual models were 0.85, 0.93, 0.86 for  $C_{\max}/dose$ ,  $t_{\max}$ , and  $AUC/dose$ , respectively.

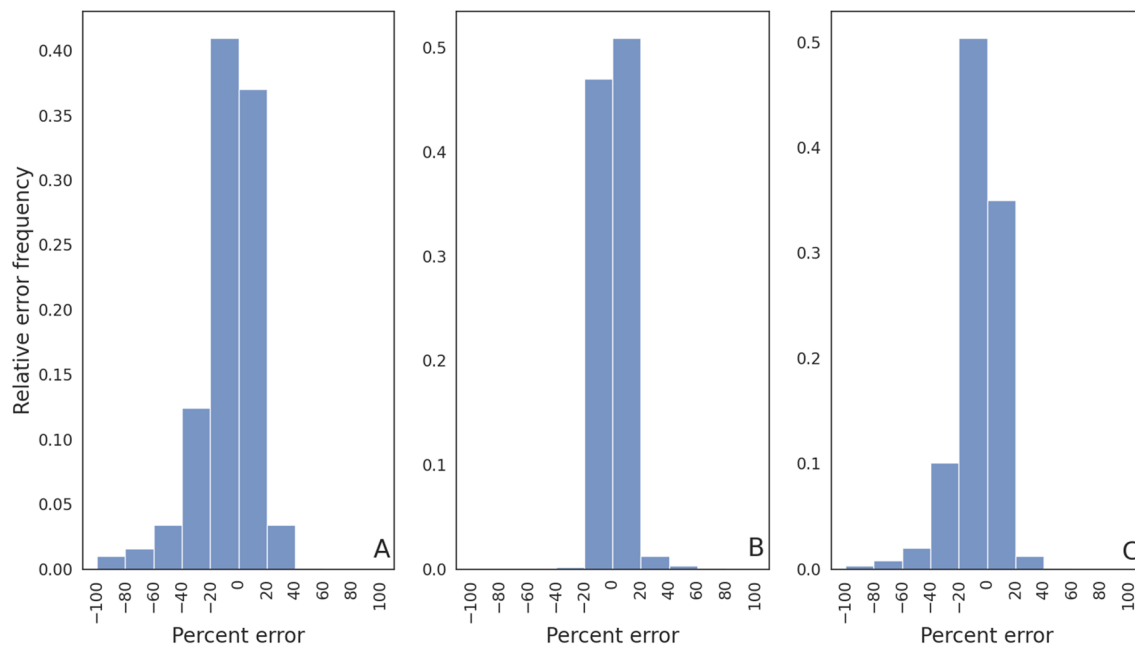
To assess the degree of relative error between the machine learning and OpenCAT model predictions across the testing subset of *virtual drugs*, results were binned to create error frequency distributions (see Fig. 4). In the panels of this figure, the abscissa represents the relative error percentage, while the ordinate of the plots represents

Drug name	CAS number	BCS class	Literature references
Acetaminophen	103-90-2	Class III/IV	41–43
Codeine	76-57-3	Class I	44,45
Diazepam	439-14-5	Class I	42,46
Enalapril	75847-73-3	Class III	47,48
Fluvastatin	93957-54-1	Class II	49,50
Metoprolol	51384-51-1	Class I	51,52
Midazolam	59467-70-8	Class I	53,54
Ranitidine	66357-35-5	Class III	42,55
Trazodone	19794-93-5	Class II	56,57
Valacyclovir hydrochloride	124832-27-5	Class III	58,59

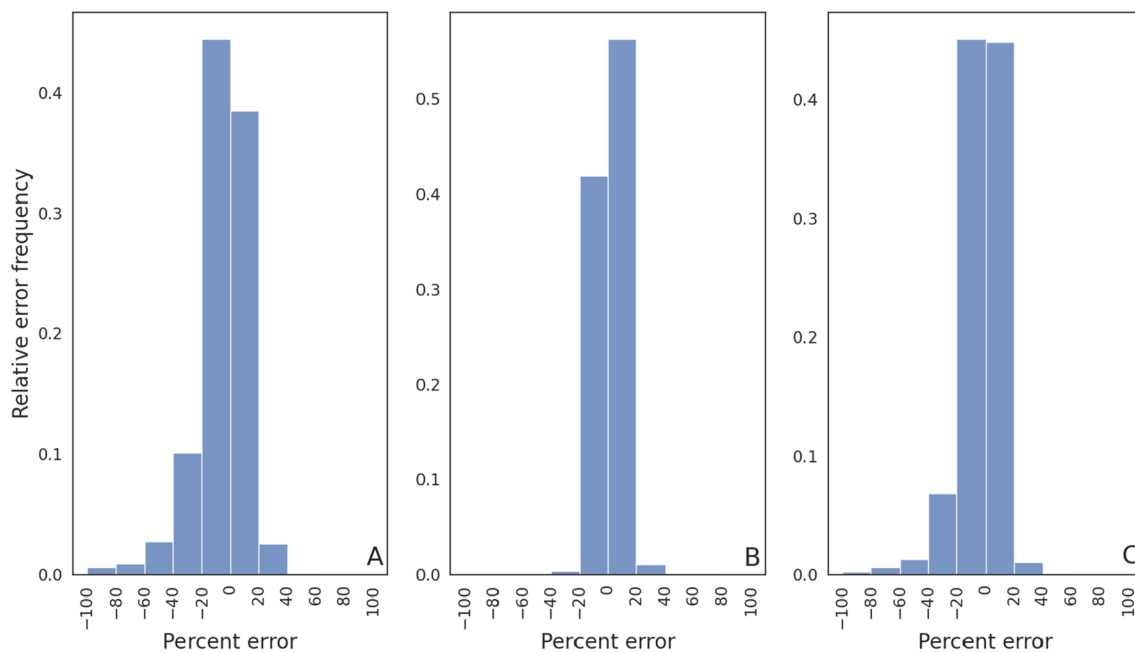
**Table 2.** Drugs used for model evaluation.



**Figure 3.** Comparison of normalized machine learning-based model predictions and their OpenCAT counterparts using the testing subset for (A)  $C_{\max}/dose$ , (B)  $t_{\max}$ , and (C)  $AUC/dose$ . In all panels, the solid line represents perfect agreement and the dashed lines indicate the  $\pm 20\%$  error bounds.



**Figure 4.** Relative error frequency for the full feature training set FFS for (A)  $C_{\max}/dose$ , (B)  $t_{\max}$ , and (C)  $AUC/dose$ .



**Figure 5.** Relative error frequency for the optimal reduced feature training set (RFS) for (A)  $C_{\max}/dose$ , (B)  $t_{\max}$ , and (C)  $AUC/dose$ .

Model	± 20%			± 40%		
	$C_{\max}/dose$	$t_{\max}$	$AUC/dose$	$C_{\max}/dose$	$t_{\max}$	$AUC/dose$
FFS	0.77	0.97	0.82	0.93	0.99	0.95
Optimal RFS	0.83	0.98	0.90	0.96	0.99	0.98

**Table 3.** Fraction of predictions within relative error ranges for the full and reduced training set models.

the relative error magnitude frequency. As summarized in Table 3, the total fraction of samples having relative errors in the range  $\pm 20\%$  was 0.77, 0.97, and 0.82 for  $C_{\max}/dose$ ,  $t_{\max}$ , and  $AUC/dose$ , respectively. At a threshold of  $\pm 40\%$ , the fractions were increased to 0.93, 0.99, and 0.95 for these same metrics.

**Optimal reduced feature set model.** Following the identification of the optimal reduce feature set, an evaluation identical to that for the FFS models was conducted. In this case, the  $R^2(\text{adj})$  values were 0.93, 0.98, and 0.95 for the  $C_{\max}/dose$ ,  $t_{\max}$ , and  $AUC/dose$  models, respectively.

Similar to the previous case, histograms were generated to quantify the frequency of obtaining a prediction within a certain error percentage (see Fig. 5). As listed in Table 3, for this model the total fraction of samples having relative errors in the range  $\pm 20\%$  was 0.83 for  $C_{\max}/dose$ , 0.98 for  $t_{\max}$ , and 0.90 for  $AUC/dose$ . For errors in the range of  $\pm 40\%$ , these fractions were increased to 0.96, 0.99, and 0.98.

**Feature importance.** As noted earlier, to assess the influence of each of the features on the model predictions, a feature importance study was conducted. It was found that the eight most influential features overall—and those used to create the optimal RFS model—were (i) the fraction unbound to partition coefficient ratio for the liver ( $FuPC_{\text{liver}}$ ), (ii, iii) the two metabolism rate constants ( $V_{\max}$  and  $K_M$ ), (iv) the subject's body mass, (v) the drug solubility, and (vi, vii, viii) the fraction unbound to partition coefficient ratio for the colon, stomach, and duodenum ( $FuPC_{\text{colon}}$ ,  $FuPC_{\text{stomach}}$ ,  $FuPC_{\text{duod}}$ ). Other results of this study are shown in Fig. 6, which depicts the feature importance scores for both the FFS and optimal RFS models.

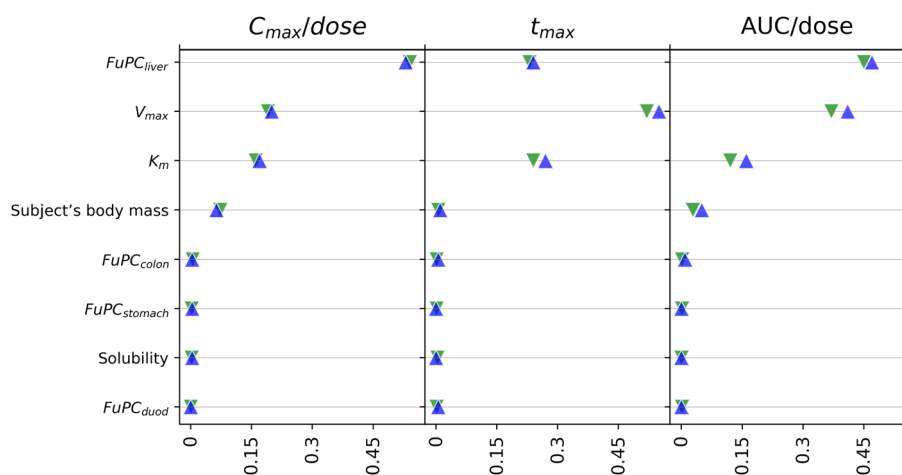
**Model verification against experimental data.** As an additional test, the optimized RFS models underwent comparison to the experimental data described earlier. The resulting values for the min-max normalized PK parameters are depicted in Fig. 7. There were often several values of the same SPKM from different experiments and/or cited uncertainty in these values. These values are represented in the figure as boxes (first to the third quartiles) and whiskers (minimum and maximum values). The predicted SPKMs from the ML and OpenCAT models are indicated by symbols. As expected, the agreement between the ML models and experimentally-obtained data was similar to that of the full PBPK model. While good agreement with experimental values was seen for many drugs and BCS classes ( $R^2(\text{adj})$  of 0.61, 0.79, and 0.77 for  $C_{\max}/dose$ ,  $t_{\max}$ , and  $AUC/dose$ , respectively), models showed relatively poor predictive capabilities for others. These deficiencies may correspond to shortcomings previously described in the literature for the ACAT model for certain kinds of drugs<sup>62</sup>.

## Discussion

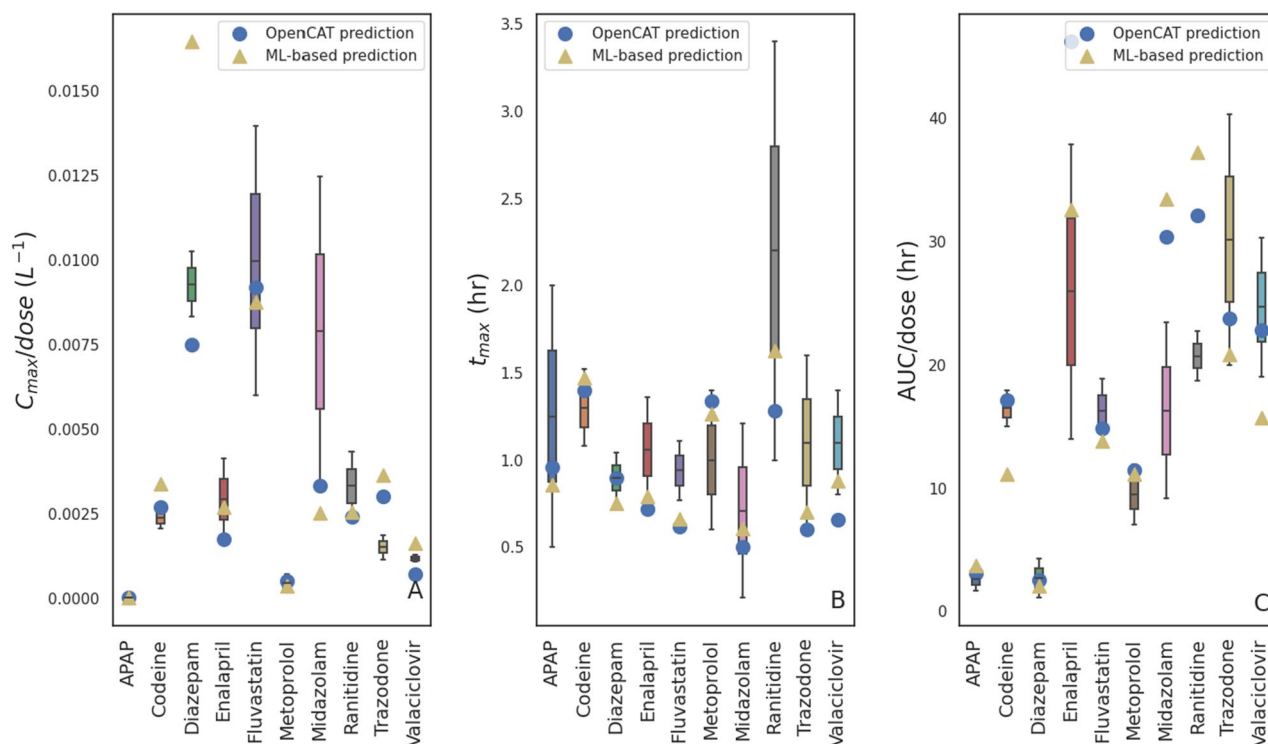
Agreement between predicted SPKMs for the machine learning models and full OpenCAT PBPK model were generally very good, suggesting that the methodology can be a viable means to introduce PBPK-level accuracy in pharmacokinetic predictions to a machine learning workflow. Models based on the optimized reduced feature set outperformed those based on the full feature set for all SPKMs evaluated, indicating that a feature optimization step is warranted to produce a model with the best fidelity with respect to the original PBPK model.

Though the primary focus here was on translating the OpenCAT model to a self-contained module appropriate as a component in a machine learning pipeline, it is expected that the methodology will be amenable to almost any PBPK model. Moreover, while this study focused on a specific set of model inputs and outputs (SPKMs), these can easily be customized for the application of interest.

Despite its promise, there are two potential deficiencies that must be considered. First, the generation of the set of *virtual drugs* used to underpin the methodology relied on randomly sampling values across parameter



**Figure 6.** Feature importance assessment for the ML-based models. Green inverted triangles represent the feature importance scores associated with the FFS model while blue triangles represent the scores for the optimal RFS model.



**Figure 7.** Comparison of experimental results vs those of the presents ML-based models and OpenCAT for (A)  $C_{max}/dose$ , (B)  $t_{max}$ , and (C)  $AUC/dose$ .

ranges. Although realistic values were used to establish these ranges, some combinations of properties likely resulted in unrealistic drug candidates. Second, the derived ML model will suffer the same predictive deficiencies and anomalies as the underlying PBPK model, so a prudent choice of the underlying model must be made.

### Data availability

The datasets generated and/or analysed during the current study are available in a Zenodo repository, <https://doi.org/10.5281/zenodo.7837360>.

Received: 8 April 2023; Accepted: 6 September 2023

Published online: 11 September 2023

### References

- Issa, N. T., Stathias, V., Schürer, S. & Dakshanamurthy, S. Machine and deep learning approaches for cancer drug repurposing. In *Seminars in Cancer Biology*, vol. 68, 132–142 (Elsevier, 2021).
- Martinelli, D. Generative machine learning for de novo drug discovery: A systematic review. *Comput. Biol. Med.* 105403 (2022).
- Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nat. Mater.* **18**, 435–441. <https://doi.org/10.1038/s41563-019-0338-z> (2019).
- An, T. *et al.* A machine learning-based approach to ER $\alpha$  bioactivity and drug ADMET prediction. *Front. Genet.* **13**, 1087273. <https://doi.org/10.3389/fgene.2022.1087273> (2022).
- Periwal, V. *et al.* Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. *PLoS Comput. Biol.* **18**, e1010029. <https://doi.org/10.1371/journal.pcbi.1010029> (2022).
- Lane, T. R. *et al.* Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. *Mol. Pharm.* **18**, 403–415. <https://doi.org/10.1021/acs.molpharmaceut.0c01013> (2021).
- Robinson, M. C., Glen, R. C. & Lee, A. A. Validating the validation: Reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. *J. Comput. Aided Mol. Des.* **34**, 717–730. <https://doi.org/10.1007/s10822-019-00274-0> (2020).
- Miljković, F. *et al.* Machine learning models for human in vivo pharmacokinetic parameters with in-house validation. *Mol. Pharm.* **18**, 4520–4530. <https://doi.org/10.1021/acs.molpharmaceut.1c00718> (2021).
- Destere, A. *et al.* A hybrid model associating population pharmacokinetics with machine learning: A case study with iohexol clearance estimation. *Clin. Pharmacokinet.* **61**, 1157–1165. <https://doi.org/10.1007/s40262-022-01138-x> (2022).
- Danishuddin, N., Kumar, V., Faheem, M. & Woo Lee, K. A decade of machine learning-based predictive models for human pharmacokinetics: Advances and challenges. *Drug Discov. Today* **27**, 529–537. <https://doi.org/10.1016/j.drudis.2021.09.013> (2022).
- Ota, R. & Yamashita, F. Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics. *J. Control. Release* **352**, 961–969. <https://doi.org/10.1016/j.jconrel.2022.11.014> (2022).
- Shaik, A. N., Khan, A. A. & ADMET & DMPK. Physiologically based pharmacokinetic (PBPK) modeling and simulation in drug discovery and development. *Admet & Dmpk.* **7**, 1–3. <https://doi.org/10.5599/admet.667> (2019).
- Zhuang, X. & Lu, C. PBPK modeling and simulation in drug research and development. *Acta Pharm. Sin.* **6**, 430–440. <https://doi.org/10.1016/j.apsb.2016.04.004> (2016).
- Wang, K. *et al.* Physiologically based pharmacokinetic models are effective support for pediatric drug development. *AAPS Pharm-SciTech* **22**, 208. <https://doi.org/10.1208/s12249-021-02076-w> (2021).

15. Anand, O., Pepin, X. J. H., Kolhatkar, V. & Seo, P. The Use of Physiologically Based Pharmacokinetic Analyses-in Biopharmaceutics Applications -Regulatory and Industry Perspectives. *Pharm. Res.* **39**, 1681–1700. <https://doi.org/10.1007/s11095-022-03280-4> (2022).
16. Manolis, E. *et al.* Using mechanistic models to support development of complex generic drug products: European Medicines Agency perspective. *CPT Pharm. Syst. Pharmacol.* <https://doi.org/10.1002/psp4.12906> (2023).
17. Glassman, P. M. & Balthasar, J. P. Physiologically-based modeling of monoclonal antibody pharmacokinetics in drug discovery and development. *Drug Metab. Pharmacokinet.* **34**, 3–13 (2019).
18. Li, M., Zou, P., Tyner, K. & Lee, S. Physiologically based pharmacokinetic (PBPK) modeling of pharmaceutical nanoparticles. *AAPS J.* **19**, 26–42 (2017).
19. Agoram, B., Woltoz, W. S. & Bolger, M. B. Predicting the impact of physiological and biochemical processes on oral drug bioavailability. *Adv. Drug Deliv. Rev.* **50**, S41–S67 (2001).
20. Bois, F. Y. PopKAT Default Pharmacokinetic Models Structures and Parameters. Tech. Rep. FDA: 1U01FD005838-01, INERIS, Verneuil en Halatte, France (2020).
21. Hsieh, N.-H. *et al.* A Bayesian population physiologically based pharmacokinetic absorption modeling approach to support generic drug development: Application to Bupropion hydrochloride oral dosage forms. *J. Pharmacokinet. Pharmacodyn.* **48**, 893–908. <https://doi.org/10.1007/s10928-021-09778-5> (2021).
22. Bois, F. Y. & Maszle, D. R. Gnumcsim: A monte carlo simulation program. *J. Stat. Softw.* (1997).
23. Ping, D. *The Machine Learning Solutions Architect Handbook: Create Machine Learning Platforms to Run Solutions in an Enterprise Setting* (Packt Publishing, 2022).
24. Amidon, G. L., Lennernäs, H., Shah, V. P. & Crison, J. R. A theoretical basis for a biopharmaceutic drug classification: The correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm. Res.* **12**, 413–420. <https://doi.org/10.1023/A:1016212804288> (1995).
25. Manallack, D. T. The p k a distribution of drugs: Application to drug discovery. *Perspect. Med. Chem.* **1**, 1177391X0700100003 (2007).
26. Williams, H. D. *et al.* Strategies to address low drug solubility in discovery and development. *Pharmacol. Rev.* **65**, 315–499 (2013).
27. Dahlgren, D. *et al.* Regional intestinal permeability of three model drugs in human. *Mol. Pharm.* **13**, 3013–3021 (2016).
28. Reith, D. *et al.* Simultaneous modelling of the Michaelis–Menten kinetics of paracetamol sulphation and glucuronidation. *Clin. Exp. Pharmacol. Physiol.* **36**, 35–42 (2009).
29. Claassen, K., Willmann, S., Eissing, T., Preusser, T. & Block, M. A detailed physiologically based model to simulate the pharmacokinetics and hormonal pharmacodynamics of enalapril on the circulating endocrine renin-angiotensin-aldosterone system. *Front. Physiol.* **4**, 4 (2013).
30. Isbister, G. K. *et al.* Zero-order metoprolol pharmacokinetics after therapeutic doses: Severe toxicity and cardiogenic shock. *Clin. Toxicol.* **54**, 881–885 (2016).
31. Abebe, B. T. *et al.* Pharmacokinetic drug-drug interactions between tropium chloride and ranitidine substrates of organic cation transporters in healthy human subjects. *J. Clin. Pharmacol.* **60**, 312–323 (2020).
32. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–1107. <https://doi.org/10.1093/nar/gkr777> (2012).
33. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380. <https://doi.org/10.1093/nar/gkac956> (2023).
34. Zur, M., Hanson, A. S. & Dahan, A. The complexity of intestinal permeability: Assigning the correct BCS classification through careful data interpretation. *Eur. J. Pharm. Sci.* **61**, 11–17. <https://doi.org/10.1016/j.ejps.2013.11.007> (2014).
35. Bergström, C. A. S., Andersson, S. B. E., Fagerberg, J. H., Ragnarsson, G. & Lindahl, A. Is the full potential of the biopharmaceutics classification system reached?. *Eur. J. Pharm. Sci.* **57**, 224–231. <https://doi.org/10.1016/j.ejps.2013.09.010> (2014).
36. Bransford, P. *et al.* Ich m9 guideline in development on biopharmaceutics classification system-based biowaivers: An industrial perspective from the iq consortium. *Mol. Pharm.* **17**, 361–372 (2019).
37. Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 278–282 (IEEE, 1995).
38. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 1189–1232 (2001).
39. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** (2017).
40. Nembrini, S., König, I. R. & Wright, M. N. The revival of the gini importance?. *Bioinformatics* **34**, 3711–3718 (2018).
41. Kalantzi, L. *et al.* Biowaiver monographs for immediate release solid oral dosage forms: Acetaminophen (paracetamol). *J. Pharm. Sci.* **95**, 4–14 (2006).
42. Papich, M. G. & Martinez, M. N. Applying biopharmaceutical classification system (BCS) criteria to predict oral absorption of drugs in dogs: Challenges and pitfalls. *AAPS J.* **17**, 948–964 (2015).
43. Critchley, J., Critchley, L., Anderson, P. & Tomlinson, B. Differences in the single-oral-dose pharmacokinetics and urinary excretion of paracetamol and its conjugates between Hong Kong Chinese and Caucasian subjects. *J. Clin. Pharm. Ther.* **30**, 179–184 (2005).
44. Zur, M., Hanson, A. S. & Dahan, A. The complexity of intestinal permeability: Assigning the correct BCS classification through careful data interpretation. *Eur. J. Pharm. Sci.* **61**, 11–17 (2014).
45. Kim, I. *et al.* Plasma and oral fluid pharmacokinetics and pharmacodynamics after oral codeine administration. *Clin. Chem.* **48**, 1486–1496 (2002).
46. Friedman, H. *et al.* Pharmacokinetics and pharmacodynamics of oral diazepam: Effect of dose, plasma concentration, and time. *Clin. Pharmacol. Ther.* **52**, 139–150 (1992).
47. Verbeeck, R. K. *et al.* Biowaiver monographs for immediate-release solid oral dosage forms: Enalapril. *J. Pharm. Sci.* **106**, 1933–1943 (2017).
48. Arafat, T. *et al.* Pharmacokinetics and pharmacodynamics profiles of enalapril maleate in healthy volunteers following determination of enalapril and enalaprilat by two specific enzyme immunoassays. *J. Clin. Pharm. Ther.* **30**, 319–328 (2005).
49. Sandhala, D. & Lankalapalli, S. New method for the in vivo estimation of fluvastatin and its application for pharmacokinetic studies in rabbit. *Indian J. Pharm. Educ. Res.* **54**, 1153–1158 (2020).
50. Smith, H., Jokubaitis, L., Troendle, A., Hwang, D. & Robinson, W. Pharmacokinetics of fluvastatin and specific drug interactions. *Am. J. Hypertens.* **6**, 375S–382S (1993).
51. Yang, Y. *et al.* Biopharmaceutics classification of selected  $\beta$ -blockers: Solubility and permeability class membership. *Mol. Pharm.* **4**, 608–614 (2007).
52. Stout, S. M. *et al.* Influence of metoprolol dosage release formulation on the pharmacokinetic drug interaction with paroxetine. *J. Clin. Pharmacol.* **51**, 389–396 (2011).
53. Guittet, C., Manso, M., Burton, I., Granier, L.-A. & Marçon, F. A two-way randomized cross-over pharmacokinetic and pharmacodynamic study of an innovative oral solution of midazolam (adv6209). *Pharm. Res.* **34**, 1840–1848 (2017).
54. Link, B. *et al.* Pharmacokinetics of intravenous and oral midazolam in plasma and saliva in humans: Usefulness of saliva as matrix for cyp3a phenotyping. *Br. J. Clin. Pharmacol.* **66**, 473–484 (2008).
55. Van Hecken, A., Tjandramaga, T., Mullie, A., Verbesselt, R. & De Schepper, P. Ranitidine: Single dose pharmacokinetics and absolute bioavailability in man. *Br. J. Clin. Pharmacol.* **14**, 195–200 (1982).
56. Danielak, D. *et al.* Physiologically based dissolution testing in a drug development process—A case study of a successful application in a bioequivalence study of trazodone er formulations under fed conditions. *AAPS PharmSciTech* **21**, 1–11 (2020).



57. Kale, P. & Agrawal, Y. K. Pharmacokinetics of single oral dose trazodone: A randomized, two-period, cross-over trial in healthy, adult, human volunteers under fed condition. *Front. Pharmacol.* **6**, 224 (2015).
58. Metry, M. & Polli, J. E. Evaluation of excipient risk in BCS class i and iii biowaivers. *AAPS J.* **24**, 20 (2022).
59. Lin, H., Tian, Y., Tian, J.-X., Zhang, Z.-J. & Mao, G.-G. Pharmacokinetics and bioequivalence study of valacyclovir hydrochloride capsules after single dose administration in healthy Chinese male volunteers. *Arzneimittelforschung* **60**, 162–167 (2010).
60. van Rossum, G. (Guido). Python reference manual (1995). Issue: R 9525 Publication Title: Department of Computer Science [CS].
61. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. Gobeau, N., Stringer, R., De Buck, S., Tuntland, T. & Faller, B. Evaluation of the gastroplus<sup>TM</sup> advanced compartmental and transit (acat) model in early discovery. *Pharm. Res.* **33**, 2126–2139 (2016).

### Author contributions

B.R. conceived the study, S.H. developed the software and conducted the simulations, S.H. and B.R. analysed the results. Both authors reviewed the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to B.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**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/>.

© The Author(s) 2023