Received: 10 February 2024 | Revised: 9 April 2024 | Accepted: 30 April 2024

DOI: 10.1111/cts.13824

ARTICLE

Machine learning framework to predict pharmacokinetic profile of small molecule drugs based on chemical structure

Nikhil Pillai[1] | **Alexandra Abos**[2] | **Donato Teutonico**[3] | **Panteleimon D. Mavroudis**[1]

[1]Global DMPK Modeling & Simulation, Sanofi, Cambridge, Massachusetts, USA

[2]Commercial Data and Analytics, Sanofi, Barcelona, Spain

[3]Translational Medicine & Early Development, Sanofi, Vitry-sur-Seine, France

Correspondence
Nikhil Pillai and Panteleimon D. Mavroudis, Global DMPK Modeling & Simulation, Sanofi, 350 Water St, Cambridge, MA 02141, USA.
Email: nikhil.pillai@sanofi.com and panteleimon.mavroudis@sanofi.com

Abstract

Accurate prediction of a new compound's pharmacokinetic (PK) profile is pivotal for the success of drug discovery programs. An initial assessment of PK in preclinical species and humans is typically performed through allometric scaling and mathematical modeling. These methods use parameters estimated from in vitro or in vivo experiments, which although helpful for an initial estimation, require extensive animal experiments. Furthermore, mathematical models are limited by the mechanistic underpinning of the drugs' absorption, distribution, metabolism, and elimination (ADME) which are largely unknown in the early stages of drug discovery. In this work, we propose a novel methodology in which concentration versus time profile of small molecules in rats is directly predicted by machine learning (ML) using structure-driven molecular properties as input and thus mitigating the need for animal experimentation. The proposed framework initially predicts ADME properties based on molecular structure and then uses them as input to a ML model to predict the PK profile. For the compounds tested, our results demonstrate that PK profiles can be adequately predicted using the proposed algorithm, especially for compounds with Tanimoto score greater than 0.5, the average mean absolute percentage error between predicted PK profile and observed PK profile data was found to be less than 150%. The suggested framework aims to facilitate PK predictions and thus support molecular screening and design earlier in the drug discovery process.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

PK evaluation plays a critical role in understanding a compound's safety and efficacy and determining its starting and efficacious dose for future clinical studies. There are major efforts in utilizing machine learning (ML) to facilitate compound screening and lead discovery by building robust QSAR models to predict ADME properties. ML-based prediction of in vivo PK dynamics is much less pronounced, probably due to limited availability of large quantity of in vivo PK profile data.

Nikhil Pillai and Alexandra Abos have contributed equally to the manuscript.

WHAT QUESTION DID THIS STUDY ADDRESS?

Role of machine learning to predict plasma PK profiles of small molecules based on chemical structure.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

Novel application of machine learning to predict plasma PK profiles based on chemical structure of small molecules in comparison to traditional approaches.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

The framework presented in this article provides a capability to perform virtual PK analysis of preclinical species (Rats). These predictions can be utilized to get insights into compound exposure at the very early stages of drug discovery process and thus facilitate compound screening. Thus, may help in reducing the cost and timeline of drug discovery process.

INTRODUCTION

Drug discovery is a complex process that encompasses several distinct stages. Initially, in *target identification* and *validation*, potential biological targets for a specific disease are identified. Next, in *lead discovery and optimization*, many compounds directed to the specified target are synthesized, screened, and optimized with respect to their absorption, distribution, metabolism, elimination (ADME) characteristics and physiochemical (PC) properties. Finally, the selected compounds progress through preclinical development to assess their pharmacokinetics (PK), efficacy and toxicity in animal models before potential regulatory approval and clinical testing.[1] Oftentimes, due to lack of efficacy or lack of exposure in preclinical species, researchers need to repeat the steps of lead identification, a process that increases the expense of the drug discovery process and may lead to more animal testing.

PK evaluation plays a critical role in understanding a compound's safety and efficacy and determining its starting and efficacious dose for future clinical studies.[2–4] Traditionally, PK in humans is predicted by scaling PK from preclinical species (e.g. mice, rats, dogs, mini-pigs and non-human primates (NHP)), a method that requires extensive use of animal experiments. Toward the goal of mitigating animal experiments, the US Food and Drug Administration (FDA) modernization act 2.0 authorizes the use of certain alternatives to animal testing including computer-based models to obtain an exemption from FDA to investigate safety and effectiveness of a drug.[5,6] Machine learning (ML) is one such alternative approach which has recently shown promise in making the drug discovery and development process more efficient.[7–9]

ML was shown to efficiently predict ADME properties (e.g., clearance (CL), intrinsic clearance (CLint), volume of distribution (Vdss)) as well as PC properties (e.g., lipophilicity (logP), solubility, fraction unbound (fu), pKa) of small molecules in multiple studies.[10–14] ML-driven predictions were shown to significantly reduce timelines of drug discovery process by expediting molecular screening and rank ordering of novel compounds.[15] Furthermore, we have recently demonstrated that combining these ML approaches with traditional pharmacometric models such as 1-compartment, or physiologically-based pharmacokinetic (PBPK) modeling, can lead to reasonable predictions of PK profiles for small molecules.[13]

In this work, we expand our previous efforts and demonstrate a proof of concept by using a novel ML framework where small molecule PK dynamics are predicted solely using ML. The rationale for replacing the pharmacokinetic modeling step applied previously (1-compartment, PBPK model) with ML is to eliminate the assumptions accompanying mechanistic models which are hard to be validated in the early development stages. The presented framework can help mitigate the need for animal experiments and provide insights into compound exposure that can be further utilized in molecule screening and reduce the costs and timelines associated with the drug discovery process.[13]

METHODS

Data

Data available from two internal databases were used for model development and testing, and these compounds were part of different series. First dataset consisted of molecular structures of 530 drug compounds which were represented as Simplified molecular-input line-entry system (SMILES) strings, along with their in vivo clearance and in vivo volume of distribution which were generated based on non-compartmental analysis (NCA). These data were utilized to identify which features will contribute to accurate prediction of PK profiles. The second dataset consisted of PK profiles (concentration vs. time data) of 397 compounds in rats administered 1 mg/kg intravenous (IV) dose. By combining these two datasets we created a group of 391 compounds for which SMILES string, PK parameters, and corresponding PK profiles were available.

Prior to ML model development, SMILES strings were preprocessed for salt stripping, were converted to their canonical forms and standardization was performed.[16] These SMILES strings were then represented as fingerprints (Morgan fingerprint 2048 bit) or molecular descriptors using RDKit. And 200 structure-based descriptors were determined using RDKit such as QED, TPSA (topological polar surface area), number of hydrogen bond donors/acceptors etc.[17]

Model framework

The proposed framework is shown in Figure 1. Small molecules' structure was available in the form of SMILES string as shown in panel (1). RDKit package (version 2018. 09.1) in Python (version 3.6.10) was used to transform the SMILES string into either descriptors or molecular fingerprints as shown in panel 2.[17] An exhaustive feature selection approach was performed where different combinations of features (PK parameters, descriptors, and fingerprints) were evaluated to identify the best-performing input parameters. Different scenarios were defined, based on the inclusion and exclusion of features. This included combination of CL and Vdss as input features, CL, Vdss, and RDKit descriptors as input features, CL, Vdss, RDKit fingerprint as input features and CL, Vdss, RDKit descriptor and RDKit fingerprint as input features. PK profiles were predicted using these models and the predicted PK profiles were compared with the observed PK profiles and the performance metrics (MAPE, R^2 etc.) were calculated. The final model was selected based on the performance metrics. These features were evaluated since they may implicitly affect the PK profiles, for example, RDKit descriptors include properties such as molecular weight, logP (lipophilicity), etc., which have been shown to impact PK profiles and are typically used as an input in well-established PBPK software's such as Simcyp and PK-Sim. After performing the feature selection, ML models were developed to predict the most important features ((Clearance (CL) and volume of distribution (Vdss))) based on the molecular structure.[13] Finally, predicted CL and Vdss were used as an input to an additional machine learning (ML) model which predicts the PK profile of the compound.

Machine Learning methodology and feature selection

As a first step to build the ML model, the dataset was split into training and test set. The training set consisted of 330 compounds (85% of total compounds) whereas the test dataset consisted of 61 compounds (15% of total compounds).[13] These 61 compounds were part of the test set for all ML models developed for this study. Next, a feature selection approach was performed where different combinations of features (PK parameters, descriptors and fingerprints) were evaluated. Parameters identified from feature selection were in vivo CL, and Vdss. Two ML models were then built to predict those features based on molecular representation. A detailed description of the ML models developed to predict in vivo CL and Vdss are shown in our previous work.[13]

Thereafter, a ML model was developed to predict the PK profile based on in vivo CL and Vdss. Different algorithms, such as generalized linear models (simple and with Poisson distribution), random forest,[18] support vector regressor,[19] light gradient boosting machine (GBM)[20] and XGBoost[21] were tested to evaluate their predictive capability based on the available dataset (Table S1). All of the above models use predicted CL, predicted Vdss, and time points (1–24 h with sample every hour) as input and predict the concentration values at each time point as output. Validation of the model was performed by using five fold cross-validation on the training set. Hyperparameter optimization was also performed during the cross-validation stage (Table S2). Performance metrics such as root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and coefficient of determination (R^2) were utilized for model comparisons.

PK profiles predicted using the proposed ML framework were compared with observed data. Metrics such as R^2, MAPE, MAE, ratio of $C_{\max}$ predicted versus $C_{\max}$ observed, and ratio of $\mathrm{AUC}_{24\mathrm{h}}$ (Area Under the Curve until 24 h) predicted versus $\mathrm{AUC}_{24\mathrm{h}}$ observed were utilized for performance evaluation.[13,22]

RESULTS

To visualize the chemical space and evaluate the distribution of the molecules in the test set in comparison with the molecules in the training dataset, we utilized t-distribution stochastic neighbor embedding (t-SNE) approach using chemplot package in python.[23] It was observed that the molecules in the test dataset are evenly distributed across the chemical space of the train dataset, which suggests that the test dataset is a good representation of the training data (Figure 2).[13]

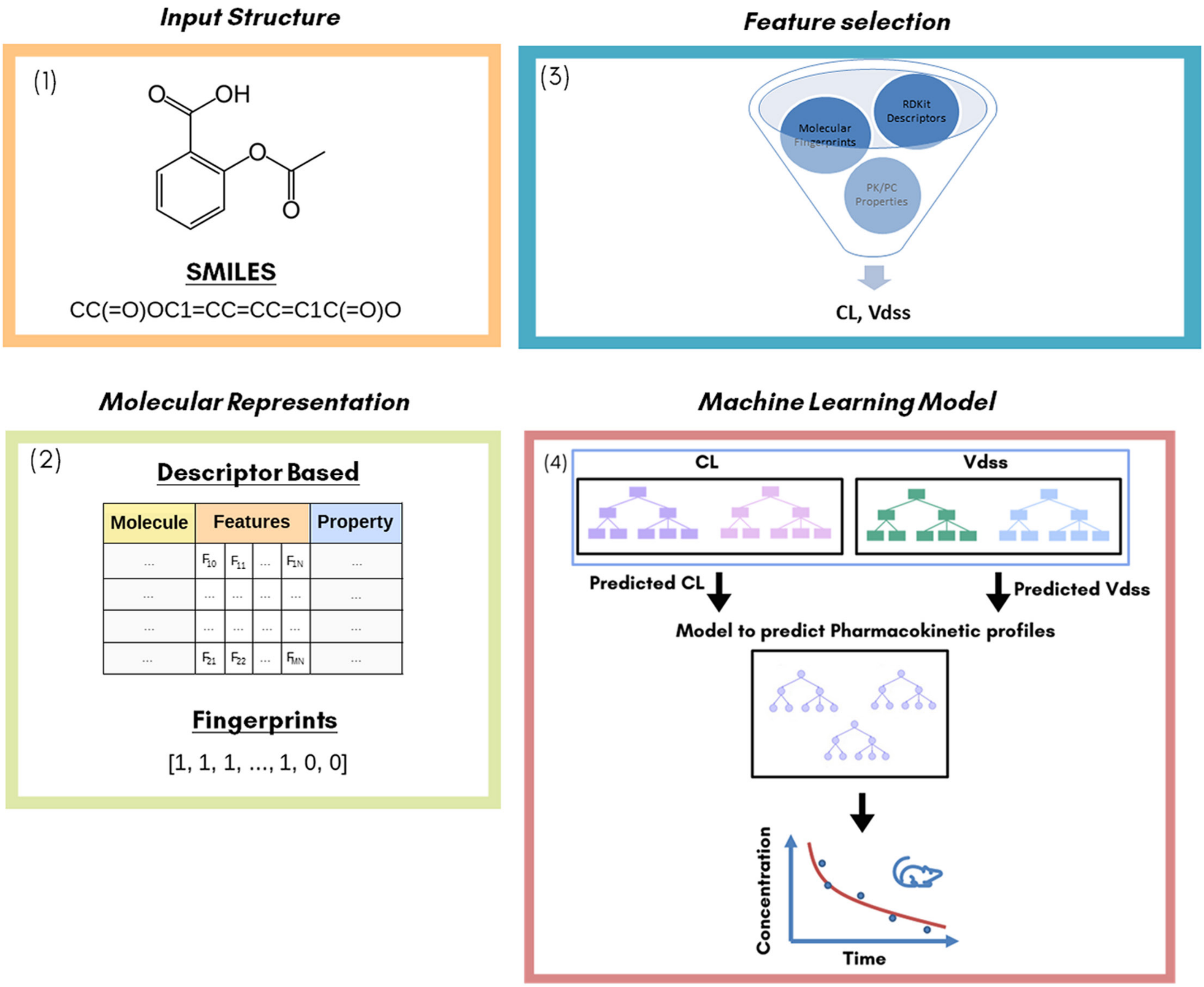

FIGURE 1 Modeling framework utilized in this study. (1) Chemical structures represented as SMILES string were used as input. (2) RDKit package was utilized to extract Fingerprints and Descriptors based on the SMILES string. (3) Feature selection performed to identify key features. (4) Chemical structures were then used as an input to a machine learning model to predict PK parameters identified in step 3, which were then used as an input to a second machine learning model to predict the PK profile.

To evaluate the performance of the ML model used to predict PK profiles, we compared the observed data with the ML predictions for the compounds in the test dataset. Figure 3 shows the scatter plot of measured concentration data and the predicted concentration data. The points are evenly distributed across the identity line (dotted black) suggesting that the model is able to capture the measured data well. XGBoost model with gamma distribution was found to have the best performance metrics in comparison to other algorithms tested to predict PK profiles using the proposed framework (Table S1).

Further evaluation was performed by visually inspecting the PK profiles generated by the proposed method versus the observed data (Figure 4). The predictions generated by the proposed ML framework were able to capture both mono-exponential and bi-exponential PK for the most compounds present in the test set. For benchmarking, predictions generated by PK-Sim (open systems pharmacology) PBPK model informed using QSAR predictions were overlayed.[13] The shaded region accounts for the PK variability observed in predicted PK profiles generated by assuming different distribution models available in PK sim, namely, Berezhkovskiy,[24] PK Sim standard,[25] Poulin and Theil,[22] Rogers and Rowland,[26] and Schmidt.[27] It can be observed that the predictions generated by the proposed framework were comparable to the predictions generated by the PBPK framework (Figure 4, Figure S1).

The distribution of ratio of AUC_{24h} observed to the ratio of AUC_{24h} predicted and maximum concentration ($C_{\max}$) observed to $C_{\max}$ predicted generated based on predictions from the proposed framework are shown in Figure 5. The median value of predicted $C_{\max}$ and AUC_{24h}

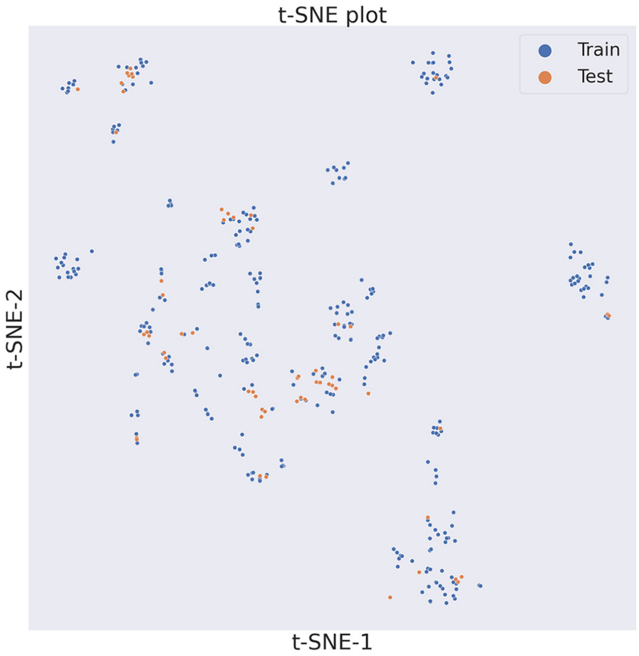

FIGURE 2 Visualization of the chemical space of compounds used to train the model and the compounds which were used in the test set. T distributed stochastic neighbor embedding (t-SNE) approach is used to perform dimensionality reduction to help facilitate visualization of chemical space.

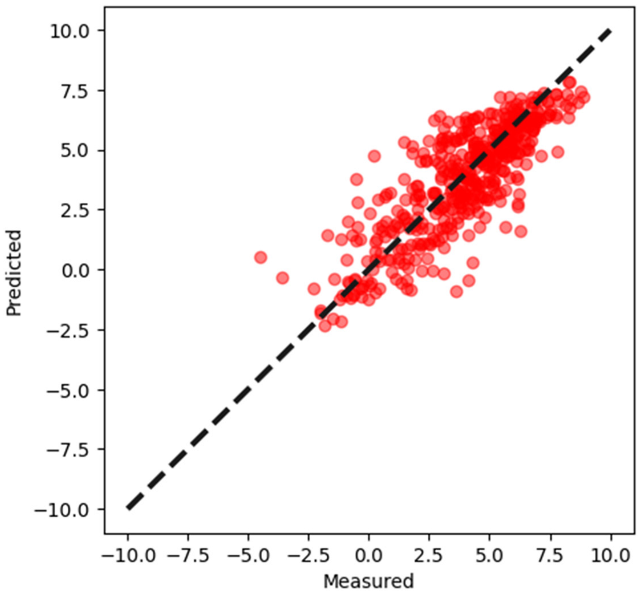

FIGURE 3 Scatter plot of concentrations predicted using machine learning framework (y-axis) versus observed concentrations (x-axis) for the test dataset.

is within two-fold of that of observed C_{max} and AUC_{24h}. Figure S2 shows the ratio of AUC_{24h} observed to AUC_{24h} predicted and C_{max} observed to C_{max} predicted for individual PBPK model predictions and average of PBPK model predictions. It can be observed that although the median value for predictions generated by PBPK models is similar to that of the predictions generated by ML framework, the interquartile range for predictions generated by PBPK models is slightly larger than the predictions generated by machine learning framework when looking at the ratio of C_{max} observed to C_{max} predicted.

Finally, performance metrics of predictions generated by ML framework were evaluated by comparing them with the observed data (Table 1). It can be observed from the MAPE, AUC_{24h} ratio, and C_{max} ratio that the error between the predictions and observations is around 3-fold, which suggests reasonable accuracy.

DISCUSSION

There are continuous efforts in the pharmaceutical industry to optimize drug development process by reducing the timelines for the development of new products and minimizing the associated costs while limiting as much as possible the need to sacrifice animals.[28] ML is an excellent tool that can transform historical data to additional knowledge, based on which decisions can be made to facilitate the drug discovery process. Although there are major efforts in utilizing ML to facilitate compound screening and lead discovery by building robust QSAR models to predict ADME properties,[15,29] ML-based prediction of in vivo PK dynamics is much less pronounced, probably due to limited availability of large quantity of in vivo PK profile data in comparison to in vitro assays.

Traditionally, PK dynamics of a compound are predicted by using PBPK and compartmental modeling, which require the researchers to generate large amount of in vitro/in vivo data. This motivates the need to explore the application of ML in this field, which may mitigate some of assumptions required while selecting certain mechanistic models and need to generate large amount of new data to gain insights into PK Characteristics.

The proposed framework uses as input ML-driven predictions of in vivo CL and in vivo Vdss to predict PK profiles. These parameters were identified to play a key role in the prediction of PK profile after performing a feature importance on a bigger set of input features which included descriptors, fingerprints, and PK properties. For the proof of concept developed in this work and to ensure any unbiased prediction of PK profiles (predicting PK profiles for compounds which were in the training set in any of the ML models developed), we utilized the available in-house dataset to develop the ML model to predict in vivo CL and Vdss. We further evaluated the error propagated of these models when utilizing the QSAR predictions of CL and Vdss to inform the ML model to predict PK profile

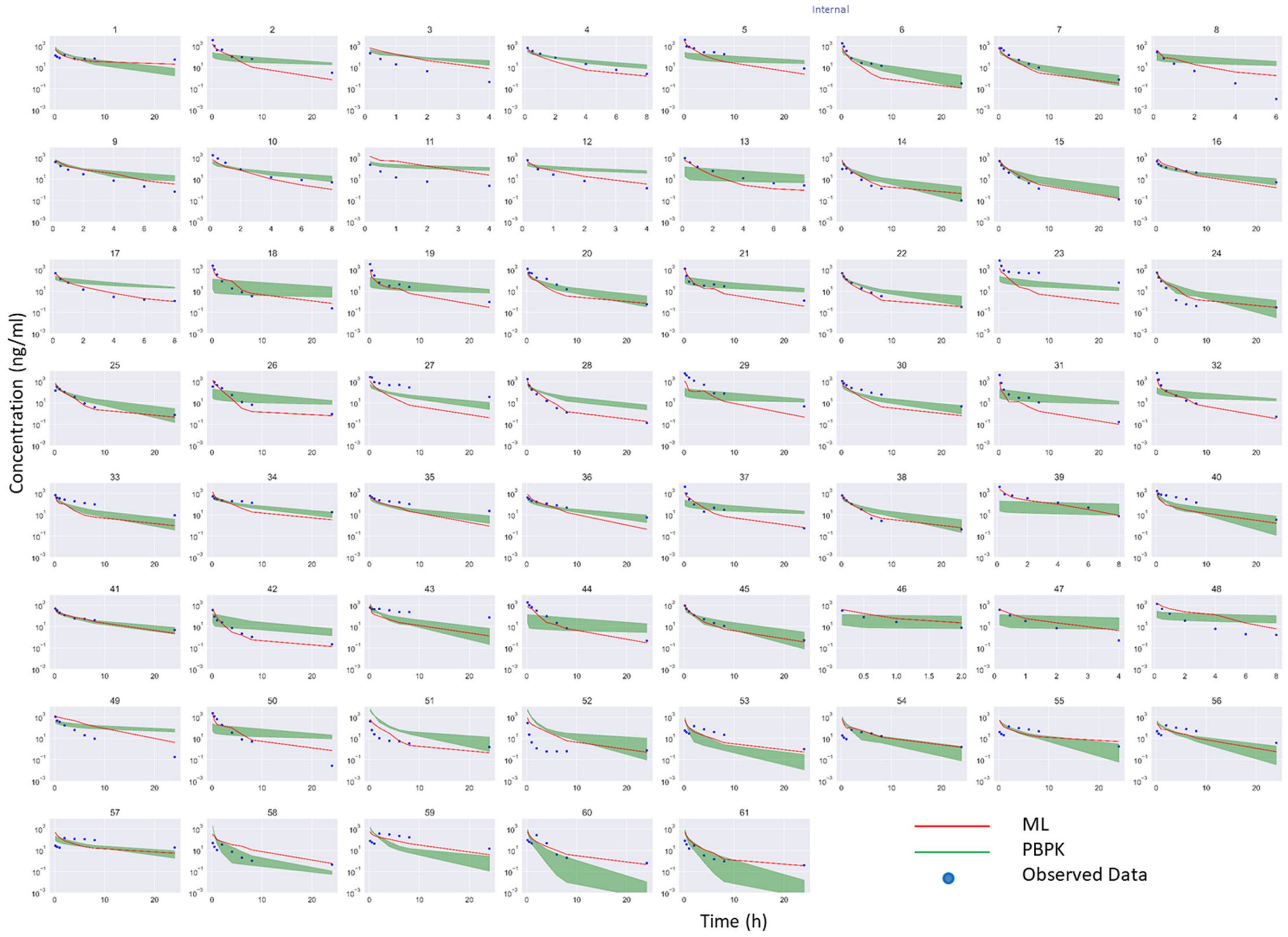

FIGURE 4 PK profiles (concentration vs. time) predicted using ML framework (red line), PBPK modeling (Green region) overlayed over observed data (Blue dots).

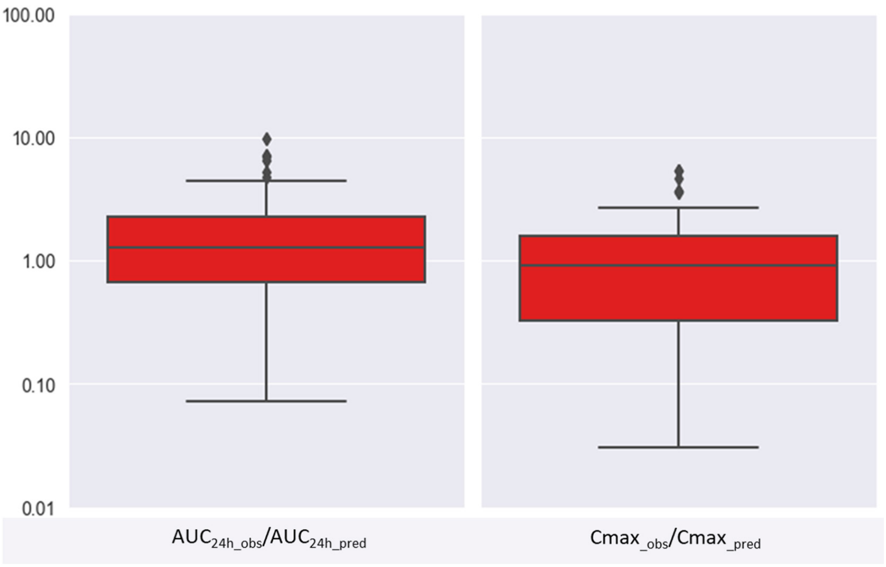

FIGURE 5 Distribution of ratio of AUC_{24h} observed versus AUC_{24h} predicted, and ratio of $C_{\max}$ observed versus $C_{\max}$ predicted for predictions generated using ML framework.

TABLE 1 Metrics for predictions generated using ML framework.

Metrics	ML
MAPE (%)	267
R^2	0.65
Median $AUC_{24h\_obs}/AUC_{24h\_pred}$	1.27
Median $Cmax_{obs}/Cmax_{pred}$	0.92

and comparing them to the scenario where ML model to predict PK profile is informed using experimentally measured in vivo CL and Vdss. Our analysis showed that there was around twofold difference between the error metrics, which further suggested that the QSAR models developed for this work have reasonable accuracy (Table S3).

Capability of ML model to predict PK profile data in comparison to other methods can be observed in Figures 3 and 4, Table 1 and Table S4. The proposed ML method can capture both mono-exponential and bi-exponential PK profiles unlike one-compartment model which assumes single volume of distribution and thus is limited to predicting mono-exponential PK profile. XGBoost algorithm utilized in this work is an ensemble model based on gradient boosting, where multiple weak models (decision trees) are combined to create a stronger predictive model. It supports different objective functions, and in this work, the loss function is based on log-likelihood of the gamma distribution. Gamma distribution is a two-parameter probability distribution that describes the time until an event occurs. The two parameters, shape (α) and rate (β), control the shape of the distribution curve and the distribution decay, respectively,[30] which could explain the positive results using in vivo CL, and Vdss as model features.

AUC_{24h} derived from predicted PK profiles were close to the AUC_{24h} derived from observed PK profile data (Figure 5). Furthermore, the average value of ratio of C_{max} observed versus C_{max} predicted, and ratio of AUC_{24h} observed versus AUC_{24h} predicted was within twofold which suggests reasonable accuracy.[31] Similarly, from Figure S2, it can be observed that even though the median value of ratio of C_{max}'s and AUC's for the average and individual PBPK profile is closer to 1, the inter-quartile range is wider in comparison to the prediction generated based on proposed ML framework. This may be linked to the uncertainty associated with the properties or additional unknown mechanisms which were not utilized to inform the PBPK model.

It can be observed that for certain compounds, the proposed framework has high error in predicting the PK profiles, this may be attributed to the difference in chemical structure of those compounds in comparison to the compounds utilized for training the model. To evaluate this, we measured Tanimoto score between the compounds in the test set and the compounds in the training dataset. Tanimoto score was utilized to measure the similarity between the two datasets. It was observed that 41 compounds in the test set had less than 0.7 Tanimoto score, with 24 of them having less than 0.5 Tanimoto score, which suggests significant difference between the chemical structure[32] (Figure S3). Average MAPE of these 24 compounds was found to be 501% whereas for rest of the compounds, the average MAPE was 145%. ML models developed in this work can be further improved by adding more data to the training set in the future, which may make implementation of deep learning approaches feasible as well.

With advancement of ML in drug discovery, multiple research groups have been trying to predict PK properties and PK profiles by utilizing PBPK models informed using predictions of QSAR models,[13,33] and by utilizing chemical structures along with certain experimentally derived in vitro properties to inform ML models which is used for predicting PK profiles.[34] However, the framework presented in this article is the first example of predicting PK profiles solely based on chemical structures as input. This proof of concept demonstrates the feasibility of ML models to accurately predict PK profiles and thus provides encouragement to further explore this area. As most ML methodologies, the black-box nature of the presented work can limit its application on evaluating plasma PK for different species and different doses than those present in the training set. Methodology requires dynamic involvement of extended PK datasets to build intelligence on PK differences between species, and nonlinearities that may arise due to saturation of involved clearance mechanisms.

In conclusion, the ML framework presented in this article can predict PK profiles with reasonable accuracy for most of the scenarios tested. These efforts aim to enable PK profile predictions earlier in the drug discovery process thus helping scientists gain insights into exposure of the compounds in vivo and helping them with prioritization and screening of compounds.

AUTHOR CONTRIBUTIONS

NP, AA, DT and PM wrote the manuscript. NP and PM designed the research. NP and AA performed the research. NP, AA, DT, and PM analyzed the data. NP and AA contributed to new analytical tools.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Yongyi Luo for helping understand the PK assays and data. Dr. Sirimas Sudsakorn and Dr. Henrik Cordes for facilitating data management and access. Dr. Hans Matter and Dr. Gerhard Hessler for their discussion on machine learning methodologies.

FUNDING INFORMATION

This article is funded by Sanofi.

CONFLICT OF INTEREST STATEMENT

All authors were employed by Sanofi while the manuscript was written. The authors declared no competing interests for this work.

ORCID

Nikhil Pillai https://orcid.org/0000-0003-3272-0603
Panteleimon D. Mavroudis https://orcid.org/0000-0002-0512-4147

REFERENCES

1. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *Br J Pharmacol*. 2011;162(6):1239-1249.
2. Sou T, Hansen J, Liepinsh E, et al. Model-informed drug development for antimicrobials: translational PK and PK/PD modeling to predict an efficacious human dose for Apramycin. *Clin Pharmacol Ther*. 2021;109(4):1063-1073.
3. Cella M, Gorter de Vries F, Burger D, Danhof M, Della Pasqua O. A model-based approach to dose selection in early pediatric development. *Clin Pharmacol Ther*. 2010;87(3):294-302.
4. Mavroudis PD, Pillai N, Wang Q, Pouzin C, Greene B, Fretland J. A multi-model approach to predict efficacious clinical dose for an anti-TGF-β antibody (GC2008) in the treatment of osteogenesis imperfecta. *CPT Pharmacometrics Syst Pharmacol*. 2022;11(11):1485-1496.
5. Stewart A, Denoyer D, Gao X, Toh YC. The FDA modernisation act 2.0: bringing non-animal technologies to the regulatory table. *Drug Discov Today*. 2023;28(4):103496.
6. Wadman M. FDA no longer needs to require animal tests before human drug trials. 2023. https://www.science.org/content/article/fda-no-longer-needs-require-animal-tests-human-drug-trials.
7. Lu J, Bender B, Jin JY, Guan Y. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature Machine Intelligence*. 2021;3(8):696-704.
8. Chen EP, Bondi RW, Michalski PJ. Model-based target pharmacology assessment (mTPA): an approach using PBPK/PD modeling and machine learning to design medicinal chemistry and DMPK strategies in early drug discovery. *J Med Chem*. 2021;64(6):3185-3196.
9. Chen EP, Bondi RW, Zhang C, et al. Applications of model-based target pharmacology assessment in defining drug design and DMPK strategies: GSK experiences. *J Med Chem*. 2022;65(9):6926-6939.
10. Feinberg EN, Joshi E, Pande VS, Cheng AC. Improvement in ADMET prediction with multitask deep featurization. *J Med Chem*. 2020;63(16):8835-8848.
11. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep*. 2017;7(1):42717.
12. Xiong G, Wu Z, Yi J, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res*. 2021;49(W1):W5-w14.
13. Panteleimon D, Mavroudis DT, Abos A, Pillai N. Application of machine learning in combination with mechanistic modeling to predict plasma exposure of small molecules. *Front Syst Biol*. 2023;3:3.
14. Dodd S, Kollipara S, Sanchez-Felix M, et al. Prediction of ARA/PPI drug–drug interactions at the drug discovery and development interface. *J Pharm Sci*. 2019;108(1):87-101.
15. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov*. 2019;18(6):463-477.
16. Swain M. MolVS: Molecule Validation and Standardization. https://molvs.readthedocs.io/en/latest/
17. RDKit.: Open-source cheminformatics. https://www.rdkit.org. https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.
18. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
19. Zhang F, O'Donnell LJ. Chapter 7 – support vector regression. In: Mechelli A, Vieira S, eds. *Machine Learning*. Academic Press; 2020:123-140.
20. Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inform Process Syst*. 2017;30:3146-3154.
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–794.
22. Poulin P, Theil FP. 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*. 2000;89(1):16-35.
23. Cihan Sorkun M, Mullaj D, Koelman J, Er S. ChemPlot, a python library for chemical space visualization. *Chemistry–Methods*. 2022;7(2):1-12.
24. Berezhkovskiy LM. Volume of distribution at steady state for a linear pharmacokinetic system with peripheral elimination. *J Pharm Sci*. 2004;93(6):1628-1640.
25. Willmann S, Lippert J, Schmitt W. From physicochemistry to absorption and distribution: predictive mechanistic modelling and computational tools. *Expert Opin Drug Metab Toxicol*. 2005;1(1):159-168.
26. 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.
27. Schmitt W. General approach for the calculation of tissue to plasma partition coefficients. *Toxicol In Vitro*. 2008;22(2):457-467.
28. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. 2010;9(3):203-214.
29. Pillai N, Dasgupta A, Sudsakorn S, Fretland J, Mavroudis PD. Machine learning guided early drug discovery of small molecules. *Drug Discovery Today*. 2022;27(8):2209-2215.
30. Hogg RV, Craig AT. *Introduction to Mathematical Statistics*. 4th ed. Macmillan; 1978.
31. Maharaj AR, Edginton AN. Physiologically based pharmacokinetic modeling and simulation in pediatric drug development. *CPT Pharmacometrics Syst Pharmacol*. 2014;3(11):e148.
32. Mokaya M, Imrie F, van Hoorn WP, Kalisz A, Bradley AR, Deane CM. Testing the limits of SMILES-based de novo molecular generation with curriculum and deep reinforcement learning. *Nat Mach Intell*. 2023;5(4):386-394.

33. Hosea NA, Jones HM. Predicting pharmacokinetic profiles using in silico derived parameters. *Mol Pharm*. 2013;10(4):1207-1215.
34. Obrezanova O, Martinsson A, Whitehead T, et al. Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. *Mol Pharm*. 2022;19(5):1488-1504.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pillai N, Abos A, Teutonico D, Mavroudis PD. Machine learning framework to predict pharmacokinetic profile of small molecule drugs based on chemical structure. *Clin Transl Sci*. 2024;17:e13824. doi:10.1111/cts.13824