molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

Prediction of the Contribution Ratio of a Target Metabolic Enzyme to Clearance from Chemical Structure Information

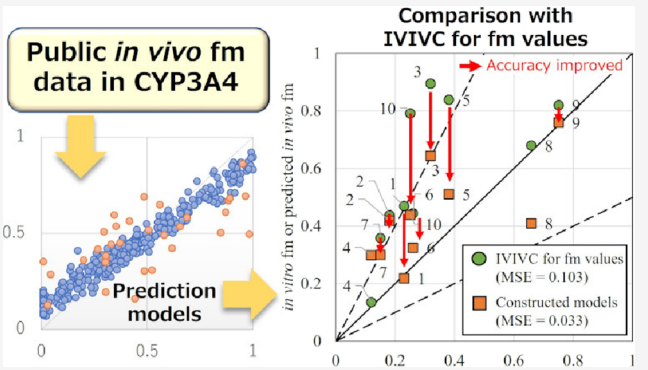
Reiko Watanabe,* Toshio Kawata, Shinya Ueda, Takumi Shinbo, Mitsuo Higashimori, Yayoi Natsume-Kitatani, and Kenji Mizuguchi

Cite This: *Mol. Pharmaceutics* 2023, 20, 419−426

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The contribution ratio of metabolic enzymes such as cytochrome P450 to *in vivo* clearance (fraction metabolized: fm) is a pharmacokinetic index that is particularly important for the quantitative evaluation of drug−drug interactions. Since obtaining experimental *in vivo* fm values is challenging, those derived from *in vitro* experiments have often been used alternatively. This study aimed to explore the possibility of constructing machine learning models for predicting *in vivo* fm using chemical structure information alone. We collected *in vivo* fm values and chemical structures of 319 compounds from a public database with careful manual curation and constructed predictive models using several machine learning methods. The results showed that *in vivo* fm values can be obtained from structural information alone with a performance comparable to that based on *in vitro* experimental values and that the prediction accuracy for the compounds involved in CYP induction or inhibition is significantly higher than that by using *in vitro* values. Our new approach to predicting *in vivo* fm values in the early stages of drug discovery should help improve the efficiency of the drug optimization process.



KEYWORDS: *pharmacokinetics, fm, CYP3A4, drug−drug interaction, in silico prediction*

INTRODUCTION

Co-administration of two or more drugs has a potential risk of drug−drug interaction (DDI); that is, a drug might be involved in altering the behaviors of the other drug. DDI often causes unexpected adverse events and increases the incidence or the severity of a known adverse event. Cytochrome P450 (CYP) is important for the metabolism of many drugs, and drugs metabolized by CYP families are exposed to a potential risk of DDIs. An unexpected severe adverse event of new drugs not only threatens a patient's life but also affects corporates' performance and reputation due to withdrawal.[1] Therefore, it is recommended to weed out drug candidates with potential DDI risk at the early stage of drug development. The development of guidelines or providing guidance toward DDI has been discussed by three bureaus (FDA, EMA, and PMDA) since the end of the '90s. They recommended that the DDI risks could be predicted by mechanistic static pharmacokinetics (MSPK) and/or physiologically based pharmacokinetic (PBPK) modeling approach using the physical and chemical properties and *in vitro* data, and clinical DDI studies could be waived if no or minimal clinical impact from DDI prediction is expected.

The contribution ratio of a target metabolic enzyme to clearance (fraction metabolized: fm) is a particularly important index for quantitative DDI predictions. However, the only *in vivo* study to obtain fm, i.e., a clinical pharmacokinetic study in co-administration with a potent inhibitor for a target metabolic enzyme, is a physically demanding study for the subject and is rarely conducted; therefore, it is difficult to obtain *in vivo* fm directly. One of the methods used instead of finding *in vivo* fm is *in vitro/in vivo* correlation (IVIVC),[2] which allows the prediction of the *in vivo* performance of a pharmaceutical product based on its *in vitro* pharmacokinetic profiles. Although IVIVC is widely used for the prediction of *in vivo* sensitivity to CYP enzymes, IVIVC does not always meet in any CYP substrate. Ohno et al.[3] reported that the contribution of the CYP3A4 enzyme determined by *in vitro* study was not highly correlated with that by *in vivo* study, the former being defined as the fm by CYP3A4 for inhibition based on intrinsic clearance (CLint), which was determined by examining the effect of CYP3A4-selective inhibitors/antibodies on drug

Received: August 17, 2022
Revised: November 22, 2022
Accepted: November 22, 2022
Published: December 20, 2022

metabolism by human liver microsome (HLM) in vitro. The group suggested that CYP3A substrates on the IVIVC line have the following characteristics: very low urinary excretion and achievement of rapid equilibrium in the liver. Efforts have been made to improve the IVIVC, such as by correction using a relative activity factor (RAF)[4] using recombinant human CYP and HLM and the percentage contribution of individual CYPs using CLint values obtained for substrates in both pooled HLM or recombinant CYPs,[5,6] which is almost consistent with RAF. Although the importance of IVIVC is widely recognized and various improvements have been made, considering that many drug candidates are terminated before the clinical development phase, *in vitro* studies using human biological materials (e.g., liver microsomes and hepatocytes) are time-consuming and expensive. Moreover, laboratory experiments and synthesis of candidate compounds require many chemicals, including organic solvents, which have a high environmental impact. Therefore, if *in vivo* fm could be predicted without any experimental data, it would bring revolutionary changes to screening, lead generation, and lead optimization.

Recently, computer-aided drug design using *in silico* models to predict the absorption, distribution, metabolism, excretion, and toxicity (ADMET) parameters has been widely implemented in the field of drug discovery and development.[7,8] This approach is effective for evaluating the physicochemical properties and *in vivo* pharmacokinetics during the early stages of drug discovery.[9,10] In addition, the use of *in silico* prediction techniques is expected to minimize the expenses and risks of subsequent withdrawals during clinical trials. To the best of our knowledge, there have been no previous reports on predicting *in vivo* fm from structural information alone.

In this study, the potential for the prediction of *in vivo* fm using machine learning was explored. Available information on drug candidates at an earlier stage is very limited, and we selected CYP3A4 as a model metabolic enzyme because of the availability of *in vivo* data on CYP3A4-mediated DDI from public sources. The objective of our study was to construct a prediction model based on the chemical structure for obtaining *in vivo* fm and demonstrate its potential as an alternative method to IVIVC in the early stages of drug development.

MATERIALS AND METHODS

Data Source and Construction of the Dataset. *In vivo* data of CYP3A4-mediated DDI were obtained from the Drug Interaction Database "DIDB" provided by the University of Washington Drug Interaction Solutions, Seattle, Washington, USA.[11] DIDB collected qualitative and quantitative human *in vitro* and clinical (*in vivo*) information on interacting co-medications and excipients based on the most relevant and up-to-date information from the large body of publications and regulatory documentation. First, the *in vivo* data reporting DDI with five typical CYP3A4 inhibitors (ketoconazole, itraconazole, ritonavir, telaprevir, and voriconazole) were extracted from the database. For each tested substrate, the ratio change of the area under the concentration−time curve (AUC) of the substrate (AUC for combination with typical inhibitors/AUC for the administration of the substrate alone) was calculated. If the dose level and regimen of CYP3A4 inhibitors were considered insufficient to inhibit CYP activity or were substantially different from those of clinical practice, data obtained under such conditions were excluded from the dataset. In addition, if the data for both the prodrug and its active metabolite were available in DIDB, the data on the prodrug were removed. When only the ratios of clearance (CL) or oral clearance (CL/F) were available from the DIDB, the ratios of CL or CL/F were utilized on behalf of the AUC ratios. Based on the AUC ratios, *in vivo* fm for each substrate was calculated using the following formula: *in vivo* fm = (1−1/AUC ratio)/IR, where IR is the inhibition ratio identified for each inhibitor. The IR values for each inhibitor are shown in Table 1.

Table 1. IR of Each Inhibitor

CYP3A4 inhibitors	IR
ketoconazole (200−400 mg)	0.98
itraconazole (100−200 mg)	0.95
ritonavir (800 mg)	0.99
telaprevir (2250 mg)	0.97
voriconazole (400 mg)	0.98

The quantities of *in vivo* data (i.e., number of reports of AUC ratio, CL ratio, or CL/F ratio archived in DIDB) vary widely depending on the tested substrates. Typical histograms of *in vivo* fm data are shown in Figure S1 (Supporting Information 1). For substrates with substantial deviations between the median and most frequent values, we thoroughly reviewed the original publications and excluded data obtained from studies where the condition was understood to be inappropriate. The data were included in the dataset if a sole datum of *in vivo* fm was available for a substrate. If multiple data points were available for a substrate, the data were curated by the following process. The median value of *in vivo* fm was calculated, and the upper and lower boundaries were defined as ±0.1 to 0.45 of the median value for each substrate by visual inspection of a histogram. Data outside the upper/lower boundaries were excluded from the calculation of the median value. If the calculated median value was outside the upper/lower boundary, *in vivo* fm was regarded as "missing". The obtained median value was used as the typical *in vivo* fm for each substrate.

The chemical structure of each compound was identified by obtaining the ChEMBL ID from the ChEMBL database.[12] Information on chemical conformations such as R/S that is not accurately described in the structural information in ChEMBL were obtained from PubChem.[13] As a result of the curation process, the dataset includes *in vivo* fm values and information on the chemical structure of 319 compounds. There were 17 compounds with *in vivo* fm values outside the range of 0−1.0 in the calculation. We assigned 0.01 to the compounds with *in vivo* fm values below 0 and 1.0 to those with *in vivo* fm values above 1.0. The created dataset of *in vivo* fm was defined as Dataset_fm. The details of Dataset_fm are shown in Supporting Information 2.

Descriptor Calculation. Structural information with SMILES was converted to a 2D Structure Data File (SDF) using Open Babel[14] with "make dative bonds" and "Generate 2D coordinates" options, and the structures were standardized using the MolVS tool written in Python using the RDKit[15] chemistry framework. We used Mordred software (v.1.2.0)[16] to calculate two-dimensional (2D) physicochemical descriptors and jCompoundMapper[17] to describe the fingerprints such as AP2D, ECFP, ECFPvariant, and MACCS. The fingerprints were hashed and folded into a fixed size of 4096 bit.

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426

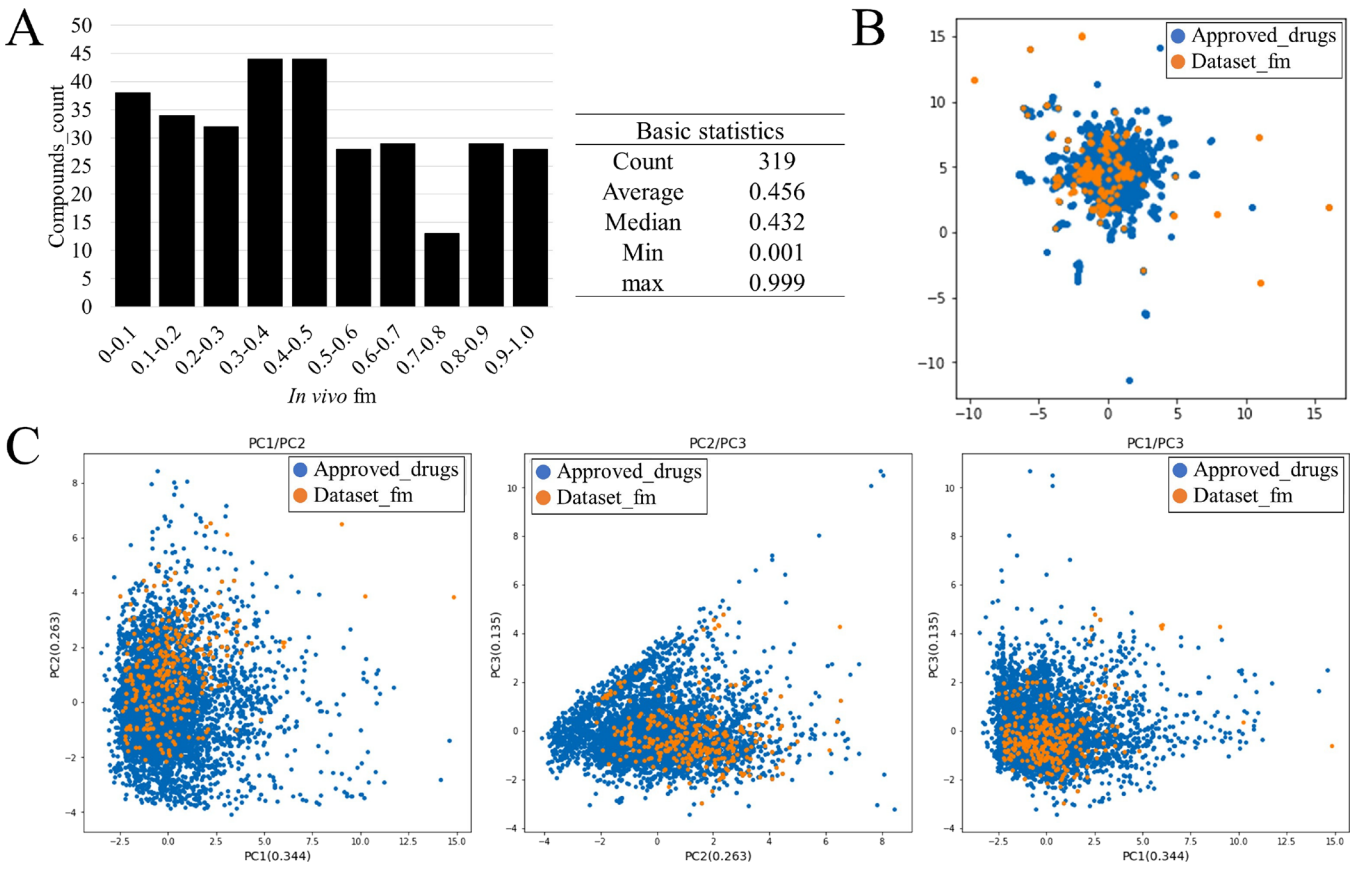


Figure 1. Distribution and visualization of chemical space in Dataset_fm. (A) Distribution of in vivo fm values in Dataset_fm and its basic statistics. (B) Chemical space visualized with fingerprint, ECFP4, in UMAP. (C) Chemical space visualized with 11 physicochemical features. Blue dots, 5858 small-molecule approved drugs; Orange dots, the Dataset_fm with 319 compounds.

Distribution and Chemical Space Analysis. The distributed 11 physicochemical descriptors are generally considered to be important parameters for synthetic expansion, and they were visualized using principal component analysis (PCA). These descriptors included SLogP, LogDpH7.4, molecular weight (MW), acidic pKa (apKa), basic pKa (bpKa), number of hydrogen bond acceptors (HBAcc), number of hydrogen bond donners (HBDon), topological polar surface area (TopoPSA), number of aromatic atoms (nAromAtom), number of aromatic bonds (nAromBond), and number of rotatable bonds (nRot). The distribution of extended connectivity circular fingerprints 4 (ECFP4) was also visualized with uniform manifold approximation and projection (UMAP).[18] Structural information on the 5858 approved drugs extracted from the KEGG DRUG Database[19] was used as a control.

Model Construction. To eliminate the influence of the split bias on the result, 20 datasets were created by randomly splitting the data into two subsets (training and test sets in the ratio of 9:1). Five machine learning methods were compared: random forest (RF), support vector machine (SVM), artificial neural network (ANN), k-nearest neighbors (kNN), and ridge and lasso linear regression (LRs). RF, SVM, and ANN were selected as representative nonlinear algorithms, LRs as the linear algorithm, and kNN as the nonparametric algorithm. Initially, descriptors with a variance lower than 0.01 in a training set were removed, and then descriptors were narrowed down to 1000 using recursive feature elimination (RFE). Finally, the descriptors selected by Boruta[20] were used for model construction. Hyperparameters were optimized using Optuna,[21] which is a Bayesian optimization method, during training with nested fivefold cross-validation. The average and standard deviation of the mean square error (MSE) in the test sets were compared as an evaluation metric. For each split, the model with the lowest MSE out of 10 iterations of feature selection by Boruta was selected as the best model. Scikit-learn (v. 0.23.2) was used to construct prediction models. The details of the machine learning settings are provided in Supporting Information 3.

Collection of the In Vivo and In Vitro fm Values in IVIVC. The *in vivo* and *in vitro* fm values in a previously reported IVIVC were collected from Figure 3 of Hisaka et al.[22] The values were extracted by reading the plot with WebPlotDigitizer.[23] The mean value was calculated when a single compound had multiple in vitro fm values. The *in vivo* and *in vitro* fm values for 29 compounds were collected and are plotted in Figure S2.

RESULTS

Distribution and Chemical Space Analysis of Dataset_fm. The distribution and basic statistics of the constructed dataset, Dataset_fm, are shown in Figure 1A. No major bias was observed, and *in vivo* fm was evenly distributed in the range of 0 to 1.0. To visualize the chemical space of the dataset, we performed two types of dimensionality reduction techniques: UMAP of fingerprints in ECFP4 and PCA of 11 descriptors, which are generally considered to be important parameters for synthetic expansion, SLogP, LogDpH7.4, MW,

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426

apKa, bpKa, HBAcc, HBDon, TopoPSA, nAromAtom, nAromBond, and nRot. As a control, a representative set of 5858 small-molecule approved drugs retrieved from KEGG DRUG was analyzed. Figure 1B,C shows a plot of UMAP and PCA with the first three principal components, respectively. The dataset covered a chemical space similar to that of approved drugs, with a much smaller number of compounds.

Models to Predict In Vivo fm Value. Regression models were created using five types of algorithms: RF, SVM, ANN, kNN, and LRs in 20 patterns of random splits of the training and test sets to eliminate the influence of the split bias on the result at a ratio of 9:1. Hyperparameter tuning was performed using nested fivefold cross-validation during model training. The finalized model using the training set was validated against the test set, not used for the training process, and the average of MSEs in the test set of each split was used as an evaluation metric. The process of model construction is illustrated in Supplemental Information 3. In addition to the original linear scale, a logit-transformed scale was used as the variable object, imposing weights on the higher and lower ranges of in vivo fm. The statistical results are shown in Table 2. The accuracy was relatively higher in nonlinear models such as RF, SVM, and ANN than that in kNN and LRs. The best MSE in the test set was 0.054 ± 0.013, with the RF in the original linear scale. In RF models, 84−125 physicochemical important descriptors and fingerprints were finally selected. A list of the top 30 descriptors determined to be high contributors across our constructed models was created as shown in Table S1. Descriptors related to the polarity and electronegativity of the compounds, like ATS, GATS, and E-state, were considered highly important. Lipophilicity also highly contributed to the predicted fm values. Physicochemical parameters occupied the higher ranks, and no fingerprints were included among the top 30 descriptors. The best models in the original and logit scales were obtained with RF, and their results are plotted in Figure 2 as examples. The MSE in the best models were 0.041 and 0.032, and the percentage of samples predicted within 2- and 1/2-fold were 87.5 and 81.3% in the original and logit scales, respectively. The average and standard deviation of the percentage of samples predicted within 2- and 1/2-fold in the top 10 models were 78.8 ± 3.4% in the original scale. Although the model on the logit scale improved underestimation of the higher range and overestimation of the lower range in the training set over the model on the original scale, there was no significant improvement in the MSE and the percentage of samples within 2- and 1/2-fold of the test set. The individual results in each split on the original and logit scales are shown in Tables S2 and S3, respectively, in Supporting Information 1.

Table 2. The Statistical Results in the Prediction Models in the Original and Logit Scales

algorithm	training/test	MSE_original	MSE_logit
RF	train	0.007 ± 0.000	0.005 ± 0.004
	test	0.054 ± 0.013	0.058 ± 0.010
SVM	train	0.014 ± 0.008	0.025 ± 0.010
	test	0.059 ± 0.014	0.079 ± 0.015
ANN	train	0.024 ± 0.006	0.010 ± 0.009
	test	0.069 ± 0.021	0.090 ± 0.016
kNN	train	0.041 ± 0.005	0.035 ± 0.006
	test	0.066 ± 0.017	0.072 ± 0.014
LRs	train	0.032 ± 0.007	0.035 ± 0.008
	test	1.395 ± 5.788	0.077 ± 0.015

Comparison with an IVIVC for fm Values. As no in silico prediction model for in vivo fm values using machine learning has been reported to date, we assessed the performance of our method by using a dataset obtained from a previously reported IVIVC study (see "Materials and Methods"). The *in vivo* and corresponding *in vitro* fm values were plotted for the 29 compounds in this dataset (Figure S2). By regarding this plot as the prediction of *in vivo* fm values from *in vitro* data, we calculated the MSE to be 0.055 and observed that 75.9% of the compounds were within the range of twofold of the experimental *in vivo* values. Those figures are comparable to the results of our RF models reported in the

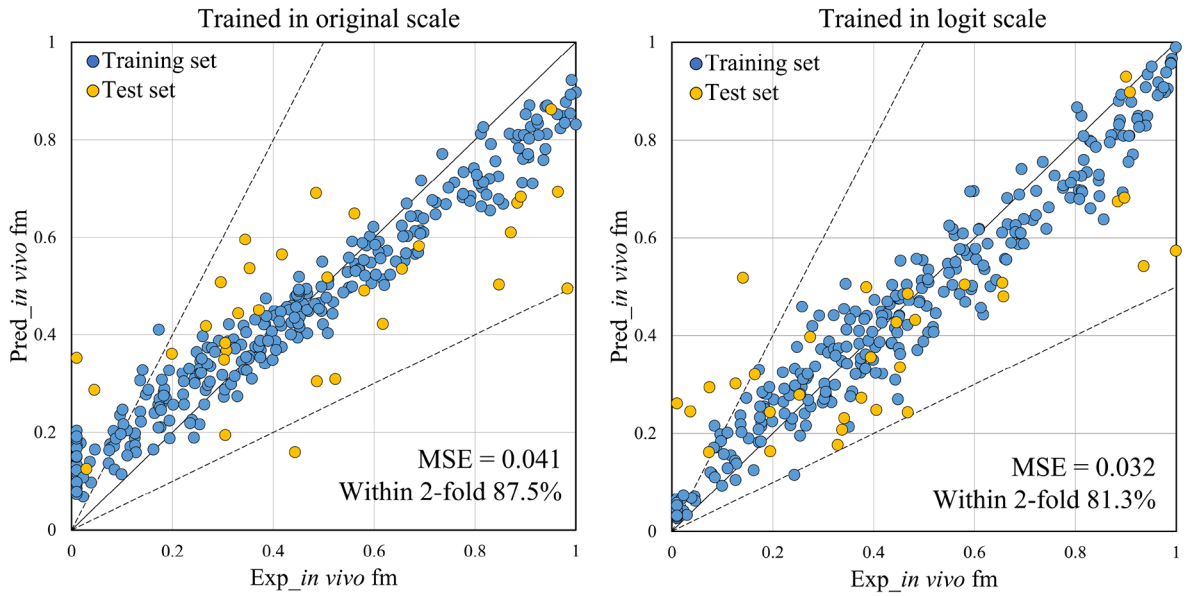


Figure 2. Plots of the prediction results with RF modes. Plot of the best models with RF among constructed models and MSE and the percentage of samples within twofold in original and logit scales. White circles, training set; black circles, test set; dashed lines, 2- and 1/2-fold.

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426

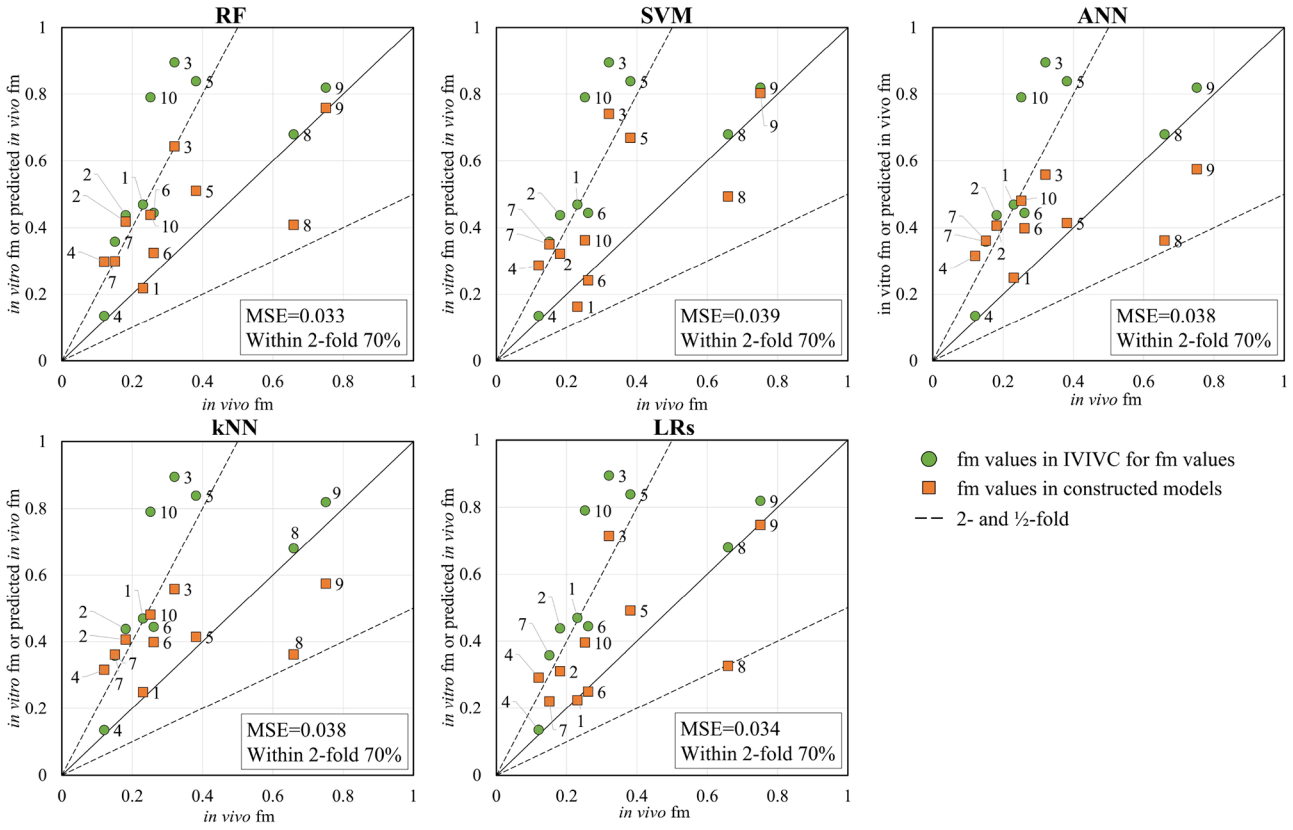


Figure 3. Plots of experimental *in vivo* fm and predicted fm or *in vitro* fm. Extracted prediction results of 10 compounds showing induction and inhibition potency toward CYPs in five machine learning algorithms (1; carbamazepine, 2; cerivastatin, 3; clarithromycin, 4; efavirenz, 5; indinavir, 6; nevirapine, 7; ritonavir, 8; saquinavir, 9; sildenafil, 10; ziprasidone). White circles: samples in the IVIVC for fm values, gray squares: samples in the constructed models. *x*-axis; experimental *in vivo* fm, *y*-axis; experimental *in vitro* fm values in the IVIVC or predicted fm values in the constructed models.

previous section (75.9% versus 78.8 ± 3.4% in the percentage within 2- and 1/2-fold, and 0.055 versus 0.054 ± 0.013 in MSE).

Since 23 out of the 29 compounds in IVIVC overlapped with those in Dataset_fm, the remaining six compounds (amitriptyline, carbamazepine, cisapride, clarithromycin, indinavir, and nifedipine) in IVIVC that were not used in training comprised the external test set. *In vivo* and *in vitro* fm values for these six compounds are shown in Table S4. The MSE of these six compounds was 0.101 in IVIVC. The fm values of the external test set were predicted using the constructed RF models, and the mean and standard deviation of MSE was 0.097 ± 0.002, indicating that our models, using structural information alone, achieved accuracy comparable to that of IVIVC for fm values. The individual results are shown in Table S5.

Prediction Accuracy for Compounds Associated with the Potency for CYP3A4 Inhibition or Induction. Ten of the 29 compounds, carbamazepine, clarithromycin, cerivastatin, efavirenz, indinavir, nevirapine, ritonavir, saquinavir, sildenafil, and ziprasidone, were previously reported to act as CYP inducers and/or inhibitors (see Table S6 for details). It has been noted that the correlation between *in vitro* and *in vivo* fm values tends to be low for those compounds (white circles in Figure S2) and the MSE for the 10 compounds was indeed larger than those for the other compounds (0.030 vs 0.103). These observations suggest that the IVIVC approach will be less effective for compounds that have the potential to induce or inhibit CYPs, and therefore, we decided to evaluate our machine-learning approach using those 10 compounds.

Since seven of those 10 compounds were also found in Dataset_fm, we removed them and re-trained our models using the remaining 312 compounds in Dataset_fm, and *in vivo* fm values of 10 compounds that act as CYP inducers and/or inhibitors were predicted by re-trained models. MSE and the percentage of samples within 2- and 1/2-fold in 10 compounds were calculated and compared with that of these compounds using the IVIVC for fm values. The MSE of the constructed model was approximately one-third that of the IVIVC (0.033−0.039 vs 0.103) in five machine learning algorithms, and the percentage of samples within 2- and 1/2-fold increased from 40 to 70%, indicating that the model constructed by machine learning can predict *in vivo* fm values for compounds involving CYP inhibition and induction more accurately than that by using *in vitro* experimental values. The details of the data used for the plots are shown in Table S7 (Supporting Information 2).

DISCUSSION

We developed the first *in silico* prediction model for *in vivo* fm by machine learning from compound structural information without requiring experimental data. The overall accuracy of the created models was comparable to that of the IVIVC for fm values, and the errors in the prediction results of compounds involved in CYP inhibition and induction were smaller than those of the IVIVC.

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426

Compared with the IVIVC for fm values, which is based on *in vitro* experimental values, the constructed model better predicted in vivo fm in compounds involved in CYP inhibition and induction, as shown in Figure 3. Despite the relative ease of obtaining HLM and recombinant human CYP enzymes and advances such as the use of RAF that accounts for CYP expression levels in the human liver, extrapolation from *in vitro* data to *in vivo* data requires more detailed consideration of the overall pharmacokinetics, such as CYP induction via nuclear receptors and CYP inhibition represented by time-dependent inhibition. We believe that the prediction model for *in vivo* fm using machine learning can consequently reflect the possibility of CYP inhibition and induction that cannot be thoroughly evaluated in a simplified *in vitro* experimental system by learning *in vivo* the fm value itself, which is the end result of the event. This is the strength of machine learning models, which can capture a variety of mechanisms in an integrated manner.

However, there are limitations in building models using public data. As shown in Figure 1, the current dataset is in a low-density state relative to the target chemical space. The accuracy easily affected how the test and training sets were split and covered only a tiny portion of the vast chemical space. This remains a concern when building a model using limited data. The amount and quality of data substantially impact the prediction accuracy and its applicable domain. The prediction accuracy will be low if there is no similar data in the training data.[24] A sensitivity analysis performed using Simcyp simulator version 19 (Certara UK Ltd., Sheffield, UK) demonstrated a more precipitous increase of AUC ratios at a higher range (>c.a. 0.8) of tested fm values, indicating that higher accuracy is required to predict *in vivo* fm values to achieve more accurate DDI prediction, especially in the case that the *in vivo* fm values are anticipated to be higher (>c.a. 0.8) (Figure S3, in Supporting Information 1). This suggests that it is essential to collect more data to cover a broader chemical space and have more diversity, as well as a particularly high *in vivo* fm range of data to increase the prediction accuracy in a substantial fm range. In addition, racial differences in metabolic enzymes constitute a significant factor in racial disparities in pharmacokinetics, and the identification of *in vivo* fm is expected to assist in predicting racial differences. However, it is challenging to collect *in vivo* fm values for CYP2C9, 2C19, and 2D6 from public databases, which are of particular note for genetic polymorphisms.

GlaxoSmithKline released raw clinical data on new drugs in 2012,[25] which was welcomed as the first step toward information disclosure. The movement to follow this lead is spreading worldwide,[26] and there has been a large range of open biomedical datasets available for training new machine learning algorithms developed by governments, medical societies, and international research collaborations.[27−29] The possibility of learning from a large amount of data containing similar structures is increasing. By using this method to build a prediction model using in-company data, we expect to develop a prediction model with a prediction accuracy higher than that of *in vitro* tests.

CONCLUSION

We collected *in vivo* fm values for 319 CYP3A4 substrates from public data and demonstrated the direct prediction of human *in vivo* fm from structural information using machine learning methods with a prediction accuracy comparable to that obtained with *in vitro* data for the first time. The prediction errors for compounds involved in CYP inhibition and induction were smaller than those of the IVIVC for fm values. The ability to predict human in vivo fm values from structural information, which until now could only be estimated through cumbersome experiments, may prove this method helpful that can be used as an alternative to IVIVC in the early stages of drug development. It is anticipated that *in vivo* fm values can be predicted with higher accuracy by this method as the number of public data increases and pharmaceutical companies begin analyzing their internal data. Because *in vivo* fm is one of the most important parameters to be considered in DDI, the constructed model can be incorporated into the system to infer DDI by predicting in vivo fm values directly from compound structures as an alternative to animal testing and IVIVC; this prediction based on structural information enables evaluation in the early stages of drug development.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.2c00698.

Supporting Information 1. (Figure S1) Typical histogram for each substrate; (Figure S2) collection of the in vivo and in vitro fm values in IVIVC; (Figure S3) sensitivity analysis of AUC ratios (with ketoconazole/without ketoconazole) of each substrate against tested in vivo fm; (Table S1) list of the top 30 descriptors; (Table S2) individual MSE of training and test sets in each split in the original scale; (Table S3) individual MSE of training and test sets in each split in the logit scale; (Table S4) six compounds in the external test set and their corresponding fm values; (Table S5) individual MSE of the external test set containing six compounds; (Table S6) cerivastatin, carbamazepine, clarithromycin, efavirenz, indinavir, nevirapine, ritonavir, saquinavir, sildenafil, and ziprasidone, as CYP substrates, inhibitors, and inducers; (Table S7) prediction accuracy in compounds associated with the potency for enzyme inhibition or induction; (Table S8) the summary of study conditions (tested substrates, inhibitor, and their dose and regimen) used for the simulation of DDI; (Supporting Information 3) details of the model construction (PDF)

Supporting Information 2. Dataset_fm (XLSX)

AUTHOR INFORMATION

Corresponding Author

Reiko Watanabe − *Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka 567-0085, Japan; Institute for Protein Research, Osaka University, Osaka 567-0085, Japan;* orcid.org/0000-0001-9359-8731; Email: reiko-watanabe@nibiohn.go.jp

Authors

Toshio Kawata − *Science Enablement Department, Data Science & Innovation Division, Research & Development, AstraZeneca K.K., Osaka 530-0011, Japan*

Shinya Ueda − *Science Enablement Department, Data Science & Innovation Division, Research & Development, AstraZeneca K.K., Osaka 530-0011, Japan*

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426

Takumi Shinbo − *Science Enablement Department, Data Science & Innovation Division, Research & Development, AstraZeneca K.K., Osaka 530-0011, Japan*
Mitsuo Higashimori − *Science Enablement Department, Data Science & Innovation Division, Research & Development, AstraZeneca K.K., Osaka 530-0011, Japan*
Yayoi Natsume-Kitatani − *Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka 567-0085, Japan; Institute of Advanced Medical Sciences, Tokushima University, Tokushima 567-0085, Japan*
Kenji Mizuguchi − *Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka 567-0085, Japan; Institute for Protein Research, Osaka University, Osaka 567-0085, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.molpharmaceut.2c00698

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was conducted in collaboration with AstraZeneca. We thank Dr. Hiroshi Komura of Osaka Metropolitan University and Dr. Rikiya Ohashi of Mitsubishi Tanabe Pharma Corporation for reviewing this manuscript. We also thank Dr. Masataka Kuroda for helping deepen the discussion section. Finally, we would like to thank Editage (www.editage.jp) for the English language editing.

■ ABBREVIATIONS

DDI, drug−drug interaction; CYP, cytochrome P450; OATP2, organic anion transporting polypeptide 2; MSPK, mechanistic static pharmacokinetics; PBPK, physiologically based pharmacokinetic; IVIVC, in vitro/in vivo correlation; RAF, relative activity factor; fm, a contribution ratio of a target metabolic enzyme to clearance; ADMET, the absorption, distribution, metabolism, excretion, and toxicity; AUC, area under the concentration−time curve; CL, clearance; CL/F, oral clearance; MSE, mean square error; LGBM, light gradient boosting machine; RF, random forest; SVM, support vector machine; ANN, artificial neural network; kNN, k-nearest neighbors; LRs, Ridge and Lasso regression; RFE, recursive feature elimination

■ REFERENCES

(1) Onakpoya, I. J.; Heneghan, C. J.; Aronson, J. K. Post-Marketing Withdrawal of 462 Medicinal Products Because of Adverse Drug Reactions: A Systematic Review of the World Literature. *BMC Med.* **2016**, *14*, 10.
(2) Davanco, M. G.; Campos, D. R.; Carvalho, P. O. In Vitro - in Vivo Correlation in the Development of Oral Drug Formulation: A Screenshot of the Last Two Decades. *Int. J. Pharm.* **2020**, *580*, No. 119210.
(3) Ohno, Y.; Hisaka, A.; Suzuki, H. General Framework for the Quantitative Prediction of CYP3A4-Mediated Oral Drug Interactions Based on the Auc Increase by Coadministration of Standard Drugs. *Clin. Pharmacokinet.* **2007**, *46*, 681−696.
(4) Crespi, C.; Gastalver, G.; Palou, A.; Roca, P. Enzymatic Determination of Carbon (14c)-Labeled Glycerol in Biological Samples. *J. Biochem. Biophys. Methods* **1995**, *30*, 179−183.
(5) McGinnity, D. F.; Parker, A. J.; Soars, M.; Riley, R. J. Automated Definition of the Enzymology of Drug Oxidation by the Major Human Drug Metabolizing Cytochrome P450s. *Drug Metab. Dispos.* **2000**, *28*, 1327−1334.
(6) Komura, H.; Iwaki, M. Usefulness of Hepatocytes for Evaluating the Genetic Polymorphism of CYP2D6 Substrates. *Xenobiotica* **2005**, *35*, 575−587.
(7) Racz, A.; Bajusz, D.; Miranda-Quintana, R. A.; Heberger, K. Machine Learning Models for Classification Tasks Related to Drug Safety. *Mol. Diversity* **2021**, *25*, 1409−1424.
(8) Pantaleao, S. Q.; Fernandes, P. O.; Goncalves, J. E.; Maltarollo, V. G.; Honorio, K. M. Recent Advances in the Prediction of Pharmacokinetics Properties in Drug Design Studies: A Review. *ChemMedChem* **2022**, *17*, No. e202100542.
(9) Alqahtani, S. In Silico Adme-Tox Modeling: Progress and Prospects. *Expert Opin. Drug Metab. Toxicol.* **2017**, *13*, 1147−1158.
(10) Pillai, N.; Dasgupta, A.; Sudsakorn, S.; Fretland, J.; Mavroudis, P. D. Machine Learning Guided Early Drug Discovery of Small Molecules. *Drug Discovery Today* **2022**, *27*, 2209−2215.
(11) *The University of Washington Drug Interaction Solutions*. https://www.druginteractionsolutions.org/. (accessed 2019-02-19)
(12) *ChEMBL* https://www.ebi.ac.uk/chembl/. (accessed 2021-01-15)
(13) *PubChem* https://pubchem.ncbi.nlm.nih.gov/. (accessed 2021-01-15)
(14) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *Aust. J. Chem.* **2011**, *3*, 33.
(15) *RDkit: Open-Source Cheminformatics software*. https://www.rdkit.org/. (accessed 2020-10-01)
(16) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Aust. J. Chem.* **2018**, *10*, 4.
(17) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A. jCompoundMapper: An Open Source Java Library and Command-Line Tool for Chemical Fingerprints. *Aust. J. Chem.* **2011**, *3*, 3.
(18) Leland, M.; John, H.; James, M. *Umap: Uniform Manifold Approximation and Projection for Dimension Reduction*. http://arxiv.org/abs/1802.03426.
(19) Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* **2017**, *45*, D353−D361.
(20) Kursa, M. B.; Rudnicki, W. R. Feature Selection with Theborutapackage. *J. Stat. Software* **2010**, *36* (), DOI: 10.18637/jss.v036.i11.
(21) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A next-generation hyperparameter optimization framework. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019; pp. 2623−2631.
(22) Hisaka, A.; Ohno, Y.; Yamamoto, T.; Suzuki, H. Theoretical Considerations on Quantitative Prediction of Drug-Drug Interactions. *Drug Metab. Pharmacokinet.* **2010**, *25*, 48−61.
(23) Rohatgi, A. *Webplotdigitizer*. https://automeris.io/WebPlotDigitizer. (accessed 2021-07-09)
(24) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* **2018**, *4* (), DOI: 10.1038/s41524-018-0081-z.
(25) Butler, D. Drug Firm to Share Raw Trial Data. *Nature* **2012**, *490*, 322.
(26) Kobayashi, S.; Kane, T. B.; Paton, C. The Privacy and Security Implications of Open Data in Healthcare. *Yearbook Med. Inform.* **2018**, *27*, 41−47.
(27) Paton, C.; Kobayashi, S. An Open Science Approach to Artificial Intelligence in Healthcare. *Yearbook Med. Inform.* **2019**, *28*, 47−51.
(28) Komura, H.; Watanabe, R.; Kawashima, H.; Ohashi, R.; Kuroda, M.; Sato, T.; Honma, T.; Mizuguchi, K. A Public-Private Partnership to Enrich the Development of in Silico Predictive Models for Pharmacokinetic and Cardiotoxic Properties. *Drug Discovery Today* **2021**, *26*, 1275−1283.
(29) Humbeck, L.; Morawietz, T.; Sturm, N.; Zalewski, A.; Harnqvist, S.; Heyndrickx, W.; Holmes, M.; Beck, B. Don't

Overweight Weights: Evaluation of Weighting Strategies for Multi-Task Bioactivity Classification Models. *Molecules* **2021**, *26*, 6959.

https://doi.org/10.1021/acs.molpharmaceut.2c00698
Mol. Pharmaceutics 2023, 20, 419−426