Received: 31 July 2022 | Revised: 17 October 2022 | Accepted: 24 October 2022

DOI: 10.1002/psp4.12884

ARTICLE

Evaluating the performance of machine-learning regression models for pharmacokinetic drug–drug interactions

Jaidip Gill[1] | **Marie Moullet**[1] | **Anton Martinsson**[2] | **Filip Miljković**[2] | **Beth Williamson**[3] | **Rosalinda H. Arends**[4] | **Venkatesh Pilla Reddy**[1]

[1] Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, Biopharmaceuticals Research & Development, AstraZeneca, Cambridge, UK

[2] Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, Research & Development, AstraZeneca, Gothenburg, Sweden

[3] Oncology Drug Metabolism and Pharmacokinetics, Research & Development, AstraZeneca, Cambridge, UK

[4] Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, Biopharmaceuticals, Research & Development, AstraZeneca, Gaithersburg, Maryland, USA

Correspondence
Venkatesh Pilla Reddy, Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, Biopharmaceuticals Research & Development, Aaron Klug Building, Granta Park, Cambridge CB21 6GH, UK.
Email: venkatesh.reddy@astrazeneca.com

Present address
Beth Williamson, Drug Metabolism and Pharmacokinetics, Union Chimique Belge (UCB), Surrey, UK

Rosalinda H. Arends, Bioinformatics & Data Science, Exelixis, Alameda, CA, USA

Abstract

Combination therapy or concomitant drug administration can be associated with pharmacokinetic drug–drug interactions, increasing the risk of adverse drug events and reduced drug efficacy. Thus far, machine-learning models have been developed that can classify drug–drug interactions. However, to enable quantification of the pharmacokinetic effects of a drug–drug interaction, regression-based machine learning should be explored. Therefore, this study investigated the use of regression-based machine learning to predict changes in drug exposure caused by pharmacokinetic drug–drug interactions. Fold changes in exposure relative to substrate drug monotherapy were collected from 120 clinical drug–drug interaction studies extracted from the Washington Drug Interaction Database and SimCYP compound library files. Drug characteristics (features) were collected such as structure, physicochemical properties, in vitro pharmacokinetic properties, cytochrome P450 metabolic activity, and population characteristics. Three different regression-based supervised machine-learning models were then applied to the prediction task: random forest, elastic net, and support vector regressor. Model performance was evaluated using fivefold cross-validation. Strongest performance was observed with support vector regression, with 78% of predictions within twofold of the observed exposure changes. The results show that changes in drug exposure can be predicted with reasonable accuracy using regression-based machine-learning models trained on data available early in drug discovery. This has potential applications in enabling earlier drug–drug interaction risk assessment for new drug candidates.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Machine learning has been used to classify drug–drug interactions by type (pharmacokinetic or pharmacodynamic) and by severity using chemical structure feature sets or knowledge graphs rich in biomedical data.

WHAT QUESTION DID THIS STUDY ADDRESS?

The utility of machine-learning regression models to predict the area under the curve ratio of a drug administered alone versus the drug administered with another "perpetrator" drug. In addition, the utility of features available early in the drug-discovery process (e.g., cytochrome P450 [CYP450] activity data) was investigated.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

Regression modeling using a support vector regressor is effective, with the majority (78%) of points predicted correctly within twofold of actual area under the curve ratios. In addition, CYP450 activity and fraction metabolized data are effective as features even when used as a lone feature set, likely due to their mechanistic relation to drug–drug interactions.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

This study demonstrates the potential for regression machine-learning modeling to be used in early drug–drug interaction risk assessment using features available early in drug discovery.

INTRODUCTION

The use of polypharmacy, where patients are prescribed multiple drugs at the same time, is increasing, with 32% of elderly adults affected in Europe.[1] Polypharmacy comes with the risk of drug–drug interactions (DDIs). DDIs can result in under- or overexposure to the drugs administered, which can reduce efficacy or increase adverse drug events (ADEs). In DDI studies, the drug that exerts an effect on another drug is known as the "perpetrator," whereas the affected drug is known as the "substrate." Pharmacokinetic (PK) DDIs are a subset of DDIs where changes in drug activity are the result of changes in the absorption, distribution, metabolism, and excretion (ADME) characteristics of the drugs involved. This usually occurs through a change in the activity of transporters or metabolic enzymes that are involved in the ADME of both the perpetrator and the substrate.[2] The cytochrome P450 (CYP) enzyme family metabolizes ~45% of marketed drugs.[3] The fraction of a drug metabolized by a specific CYP is known as the fraction metabolized (f_m).[2] CYP-mediated PK DDIs can potentially occur as a result of two conditions. First, the perpetrator drug must inhibit or induce a CYP. Second, the f_m by that CYP for a coadministered substrate drug must be >25%.[4] These DDIs have the potential to be clinically significant, with 7%–44% of ADEs caused by DDIs.[5] Inhibition of CYPs can lead to reduced substrate metabolism, resulting in increased exposure, whereas induction would result in the opposite effect.[4] Physiologically based PK (PBPK) software, such as SimCYP,[6] can model and predict DDIs. However, these models require the collection of in vitro and clinical in vivo data to build models for each drug and take time to validate. Therefore, the application of machine learning (ML) has been explored to predict DDI risk for drugs sharing ADME features with established drugs without needing to build individual PBPK models.

Machine learning algorithms consist of a set of predictor variables (features) to make predictions for the dependent variable (label). A prevalent approach for applying ML in DDI prediction has been to compute similarity metrics between features of a drug to predict their interaction because drugs sharing similar features are more likely to share interaction targets and therefore affect each other's pathways.[7] An early approach was a heterogenous network-assisted inference framework to predict DDIs.[7] This involved generating a heterogenous set of similarity features between drug A and drug B describing chemical, genomic (based on target proteins), phenotypic (based on ADEs), and therapeutic similarities to build a binary classifier for whether a DDI will occur between the two drugs. The use of similarity measures was further developed by making predictions based on the structural similarity between a drug A and the drugs that interact with the proteins, enzymes, and transporters in the interaction network of drug B.[8] However, due to the binary nature of these classifiers, the predictions were limited in their descriptive granularity. The multitask classifier "DeepDDI" method helped to address this by classifying 86 types of interaction using structural similarity profiles. This enabled the prediction of possible mechanisms behind the predicted DDIs, enhancing model interpretability.[9] These predictions were made using features derived only from structural information.

Another approach to predicting DDIs uses features from knowledge graphs, which describe entities (e.g., drug properties) and the relationships between them. This has the benefit of being able to use all the information relating to the PK and pharmacodynamics (PD) of a drug, such as drugs, pathways, structures, and diseases. The complex embedding method[10] was used to generate features from a knowledge graph for high classification accuracy.[11] An alternative embedding technique improved the area under the precision-recall curve (AUCPR) score from 0.88 to 0.97 and also enabled prediction of the adverse effect from a DDI even when the adverse effect was rare.[12] A different approach also demonstrated that integrating structural information together with knowledge graphs rather than using structure or knowledge graphs alone improves performance.[13]

The performance of these more recent classifiers shows highly accurate classification scores. However, the approaches using knowledge graphs is limited to late-stage DDI risk assessment and cannot be applied in early discovery because to build these algorithms, large amounts of biomedical data to inform the knowledge graphs are required and need to include features such as affected pathways, target proteins, and adverse effects. Although structure-based models do not rely on PK data that only become available after in vitro assays are performed, not incorporating these data may limit the performance in predicting PK DDIs. Furthermore, models thus far have focused on classification tasks, whether binary or multitask. However, the development of regression-based models for PK DDI prediction would be expected to enable a more precise estimate of the extent of the PK changes in the substrate drug, which would enable better-informed decisions regarding the risk associated with the PK DDI that could warrant dose reduction or discontinuation of this drug candidate.

To explore a more quantitative predictive approach, regression-based ML models for PK DDI prediction were investigated based on the use of early PK data, such as the CYP activity profile and f_m for a given drug. The performance of the model using different feature sets was also assessed as well as whether the regression outputs could be converted to correct DDI classes based on US Food and Drug Administration (FDA) guidelines[14] by evaluating the classification performance of the model.

MATERIALS AND METHODS

Data collection

DDI data

Clinical DDI data were collected from the Washington Drug Interaction Database[15] and SimCYP compound files.[6] Studies were included based on the following criteria: drugs involved must be available in SimCYP (Version 20),[6] dosing regimen should begin with steady-state dosing of the perpetrator drug before a single substrate dose (to simulate clinical coadministration of the drugs), and the interaction must be mediated via time-dependent inhibition (because the CYP activity profiles used as features are time dependent). In total, data from 120 studies were collected. The format of the PK data used in the analyses was the observed area under the plasma drug concentration-time curve (AUC) ratio, defined as the ratio of the AUC of the substrate drug administered in monotherapy over-the AUC of the substrate drug with the perpetrator drug. Remaining data were used directly as features or used to generate other features using in silico models. Study design details were used to specify the simulation conditions used in each SimCYP simulation. CYP activity-time profiles and f_m were generated from these simulations. A total of 201 timepoints of CYP activity for each CYP in a study simulation were extracted and used as features to represent the CYP activity-time profile. The f_m data were collected for each CYP with and without the perpetrator drug for each DDI. System-specific features are described in more detail elsewhere (Table S1).

Drug-specific data

For each DDI, drug-specific data were also collected. This included physicochemical properties and in vitro ADME data, collected from SimCYP compound files. Simplified molecular-input line-entry system[16] representations of drug chemical structures were used as an input for in silico models to derive drug-specific data additional to the SimCYP features. Chemical descriptors composed of AstraZeneca's internal OESelma descriptors, which are 84 molecular descriptors of additional physicochemical properties,[17] and extended connectivity fingerprints of diameter 4 (ECFP4),[18] a set of topological circular fingerprints to describe molecular structures. PK descriptors were features derived from AstraZeneca's internal ML models for predicting human PK[19] and in vitro ADME parameters.[20,21] Drug-specific features are described in more detail in Table S2.

Feature engineering

All models (random forest, elastic net, and support vector regression [SVR]) were implemented in Python using the Scikit-learn library.[22] To enable compatibility between the feature data and the Scikit-learn package, categorical features were encoded into binary format using one-hot encoding. To control for different scales between features, values were converted from its original scale into standard deviation units as follows:

$$z = \frac{x - \mu}{\sigma}$$

where z is the new feature value, x is the original feature value, μ is the mean value for the feature, and σ is the standard deviation of the feature. To remove redundant features, feature selection was implemented by regenerating the models using the most important features via the "Select from model" function in Scikit-learn. This method of feature selection was not included for the SVR implementation due to software incompatibility. Therefore, to ensure differences in performance between SVR and other models were not due to feature selection, additional random forest and elastic net models were built for comparison without feature selection.

Logarithmic transformations of AUCs

AUC ratios showed a strong positive skew because induction observations ranged between 0 and 1, whereas inhibition observations had a range of 1 to ~30. To normalize the AUC distribution and make it more uniform (reducing the potential for prediction bias toward the ratios < 1), a $\log_{10}$ transformation was applied.

ML modeling

Collected data were used to train regression-based ML models. Model performance was evaluated using the nested fivefold cross-validation. This involved splitting the full dataset (120 samples) into 96 training (80%) and 24 test (20%) samples five times with no overlap in each step/fold. For each such trial, additional (nested) fivefold cross-validation was performed on the training data to optimize hyperparameters. Then, an optimized model was applied to estimate the test set performance. Performance values of five independent test sets were then averaged to provide final model accuracy. The candidate hyperparameter values are detailed in Table 1. All random states were set to zero. For each model, collected data for each DDI were used as features, whereas observed $\log_{10}$ AUC ratio data were used as labels. Deep-learning techniques were not implemented due to the small sample size and to maintain model interpretability. Random forests are an ensemble of decision trees, where the output prediction for a given input is the average of the predictions of all the decision trees.[23] Elastic net regression uses linear regression with regularization using a combination of the lasso and ridge penalties.[24] SVR uses subsets of the training data (support vectors), rather than all of the data, to determine the regression line to enable increased generalizabiltiy.[25] A negative control regressor model was also implemented to model random prediction by predicting the mean of the training label data. The ML models were expected to significantly outperform the negative control regressor model to be considered predictive. Feature importance was determined using Shaply Additive Explanations (SHAP) values.[26]

Classification criteria

A potential use case of the $\log_{10}$ AUC ratio predictions is to classify interactions according to established FDA DDI guidelines,[14] which are described in Table 2 (to two significant figures). These reference thresholds were therefore used to convert the regression output into categorical labels.

Statistical analysis

Regression performance

Regression performance was evaluated using the coefficient of determination (R^2), root mean square error

TABLE 1 Hyperparameter values used for optimization of regression models

Model	Hyperparameter	Candidate values
Select from model (feature selection step)	max_features	20, 40, 60, 80, 100, 120
Random forest	n_estimators	50, 100, 500, 1000
	min_samples_leaf	1, 5, 10, 50, 100, 200, 500
	max_features	"auto," "sqrt"
Elastic net	Alpha	0.0001, 0.001, 0.01, 0.1, 0, 1, 10, 100, 500, 1000
	l1_ratio	0.001, 0.01, 0.1, 0, 1
Support vector machine	C	0.01, 0.1, 1, 10, 100, 500
	Epsilon	0.01, 0.05, 0.1, 0.5, 1, 10
	Kernel	"linear," "poly," "rbf," "sigmoid," "precomputed"

TABLE 2 Thresholds for classification of predictions using $\log_{10}$ AUC ratio values

Magnitude	Inhibition	Induction
No interaction	$-0.10 < \log_{10}$ AUC ratio < 0.10	
Weak	$0.10 \le \log_{10}$ AUC ratio ≤ 0.30	$-0.10 \ge \log_{10}$ AUC ratio ≥ -0.30
Moderate	$0.30 < \log_{10}$ AUC ratio ≤ 0.70	$-0.30 > \log_{10}$ AUC ratio ≥ -0.70
Strong	Log_{10} AUC ratio > 0.70	Log_{10} AUC ratio < -0.70

Abbreviation: AUC, area under the curve.

(RMSE), and the twofold error score. All regression performance metrics were reported as the mean of the result (to two significant figures) after fivefold cross-validation for each model along with the 95% confidence intervals.

R^2 describes the proportion of variation of the dependent variable (predicted $\log_{10}$ AUC ratio) that can be explained by the independent variable (observed $\log_{10}$ AUC ratio). It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}$$

where n is the sample size and for the i-th sample, y is the observed value, $\hat{y}$ is the predicted value, and $\bar{y}$ is the mean observed value. To investigate whether the non-normally distributed predictions and observations were statistically significantly different, the Wilcoxon signed-rank test was used.

Root mean square error describes the magnitude of error for predictions, meaning lower scores are better, and is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}}$$

A twofold error margin was used as the range for acceptable predictions. The metric score was defined as the proportion of predicted $\log_{10}$ AUC ratios within twofold of the observed $\log_{10}$ AUC ratios. This wider margin was chosen because it is the industry standard for AUC prediction of drugs with a wide therapeutic window,[27] making it appropriate for this proof-of-concept analysis.

Classification performance

Classification performance was evaluated using macro F1 scores, reported as the mean (to two significant figures) across after fivefold cross-validation ±95% confidence limits. Macro scores were used to provide equal weighting to each class because the prediction of each class had equal importance in this model. Macro F1 scores were calculated using the precision and recall metrics. The macro precision describes the proportion of positive predictions that were true positives per fold, defined as

$$\text{Macro Precision} = \frac{\sum_{i=1}^{l}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}}{l}$$

where TP_i is the number of true positives for class i, FP_i is the number of false positives, and l is the number of classes. Recall describes the proportion of actual positives that were correctly predicted (sensitivity), defined as

$$\text{Macro Recall} = \frac{\sum_{i=1}^{l}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}}{l}$$

where FN_i is the number of false negatives for a given class. Because there is a trade-off between precision and recall performance,[28] F1 score (harmonic mean of precision and recall for each fold) was used to evaluate the overall performance of the model. The macro F1 score is defined as

$$\text{Macro F1 Score} = 2\frac{\text{Macro Precision} * \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}}$$

RESULTS

Observed $\log_{10}$ AUC ratios were unevenly distributed

Distributions of observed data in a training dataset can be a source of bias in a ML model, where less-represented data ranges show higher prediction error.[29] Therefore, the distribution of the label data in the dataset was investigated to identify potential poorly represented data ranges in the DDI dataset used to train this model. Figure 1 shows a histogram of the $\log_{10}$ AUC ratios of the collected DDI studies (n = 120). The distribution of $\log_{10}$ AUC ratios were determined to be significantly different from a uniform distribution using the Kolmogorov–Smirnov test ($p < 0.005$). There were a greater number of samples showing inhibition (16 weak, 38 moderate, 19 strong) compared with induction (9 weak, 22 moderate, 10 strong), whereas noninteraction samples were the least common (6). This uneven distribution of $\log_{10}$ AUC ratios in the dataset may have been a source of bias in the ML model.

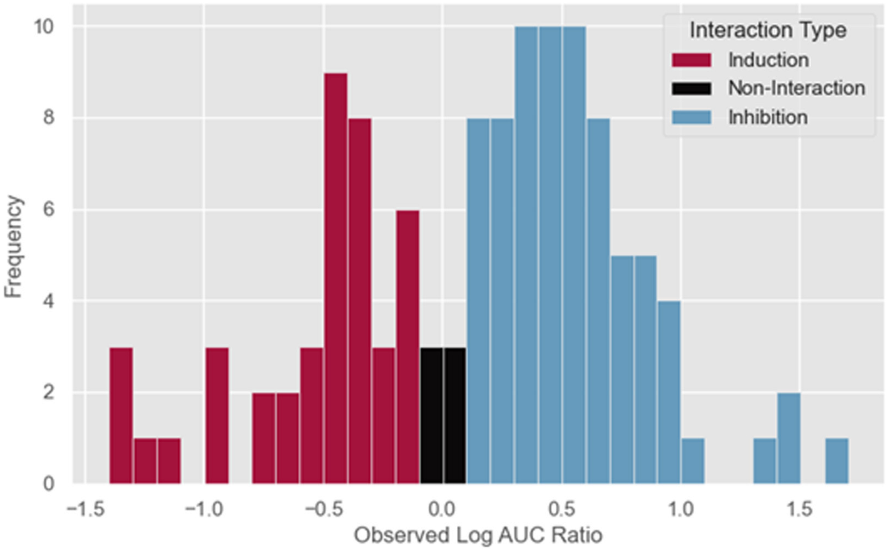


FIGURE 1 Frequency histogram of the distribution of collected $\log_{10}$ area under the curve (AUC) ratios. Observed $\log_{10}$ AUC ratios were extracted from 120 clinical drug–drug interaction studies. Of the interactions, 73 were inhibition, 41 were induction, and six were noninteractions. The ratios were not uniformly distributed (Kolmogorov–Smirnov test; $p < 0.005$).

SVR showed the strongest regression performance

The regression performance of each model was compared to determine which ML model was most appropriate for the DDI dataset. Each model was subject to fivefold cross-validation, and a mean predictor model was used as a negative control for predictive performance. Figure 2 shows that random forest, elastic net, and SVR all had significantly lower RMSE and higher R^2 and twofold error scores compared with the control model. Figure 2a shows that the random forest regressor achieved the lowest/weakest twofold error score (0.69), whereas SVR showed the highest (0.78). Figure 2b shows that the elastic net regressor demonstrated the lowest R^2 (0.67), whereas SVR showed the highest (0.73). Figure 2c shows that the highest RMSE score was achieved by the elastic net (0.33), whereas SVR achieved the lowest/best (0.30). SVR also outperformed the elastic net and random forest models built without feature selection (Figure S1). However, the differences between the ML models were small and statistically insignificant, indicating that SVR only marginally outperformed the elastic net and random forest methods. However, all ML models significantly outperformed the mean predictor control.

SVR showed the strongest classification performance after conversion to categorical labels

Classification performance was evaluated to determine whether the regression outputs would be correctly classified according to FDA guidelines.[14] Predicted $\log_{10}$ AUC ratios from the regression models were assigned to a class of DDI, and these classes were compared with the actual class (assigned based on the same thresholds) to determine the F1 score as the measure of performance. Figure 3a shows that all models showed significantly stronger mean performance than the control model. The elastic net showed the lowest macro F1 score of the noncontrol models (0.31), whereas the SVR showed the highest macro F1 score (0.40). However, intermodel performance differences were small and considered insignificant.

Across models, inhibition was generally better predicted than induction with average F1 scores of 0.42 versus 0.33, respectively. The prediction of moderate drug interactions (considering both inhibition and induction, average 0.54) were improved compared with strong (0.32) or weak (0.27) drug interactions. These two trends are concordant with the significantly higher F1 score for moderate inhibition compared with the average of all the classes in each model (0.66 vs. 0.32). In addition, there was no prediction of noninteractions, and therefore there was an F1 score of zero for noninteractions in every model. Figure 3b–d also shows that moderate drug interactions, especially moderate inhibition, had the most true positives and false positives. The random forest model showed the greatest tendency to overpredict the strength of weak interactions and correctly classified strong induction less often than the SVR. However, it correctly classified moderate induction more often than the SVR model. The raw classification matrix values are shown in Figure S2. Overall, SVR showed a stronger classification performance compared with the other two methods/models, with moderate inhibition being the most accurately predicted class across the three methods/models.

Strong correlation between predicted and observed $\log_{10}$ AUC ratios

Next, the performance of the regressor was investigated visually. Linear regression supported a strong correlation

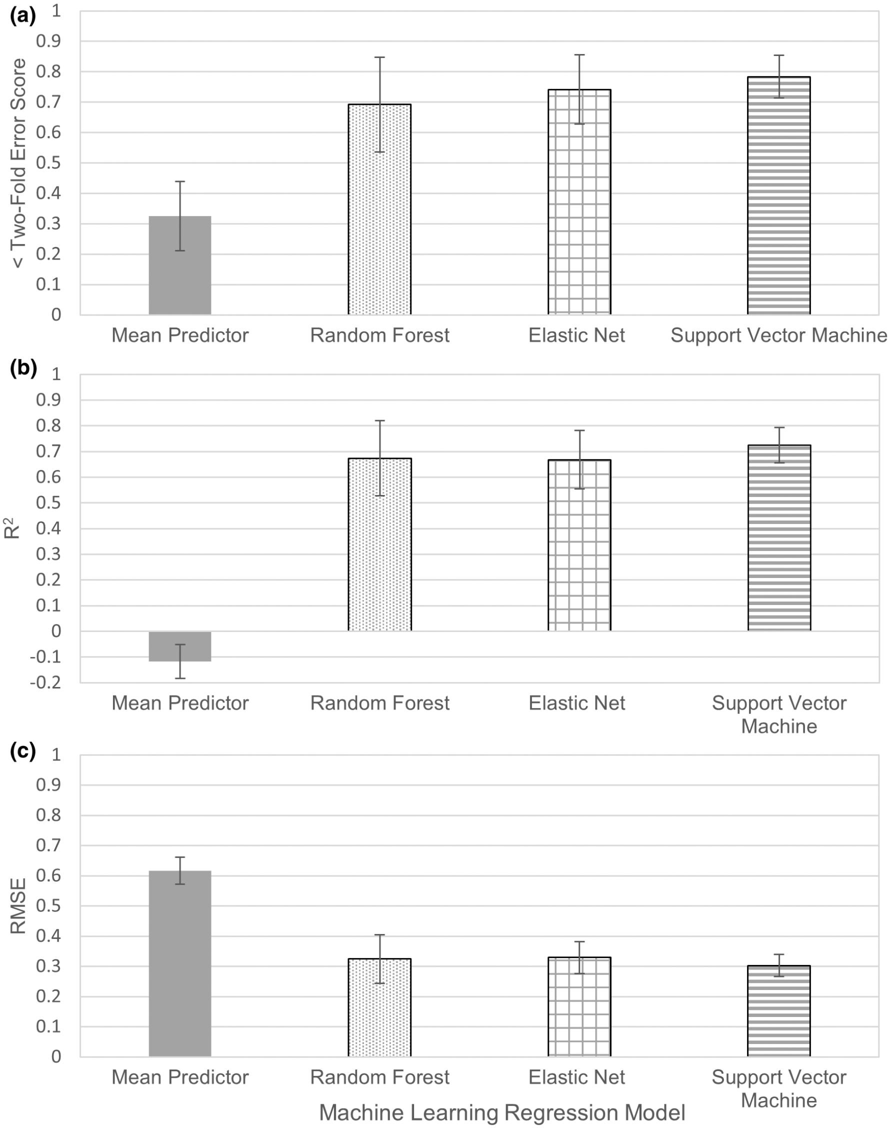
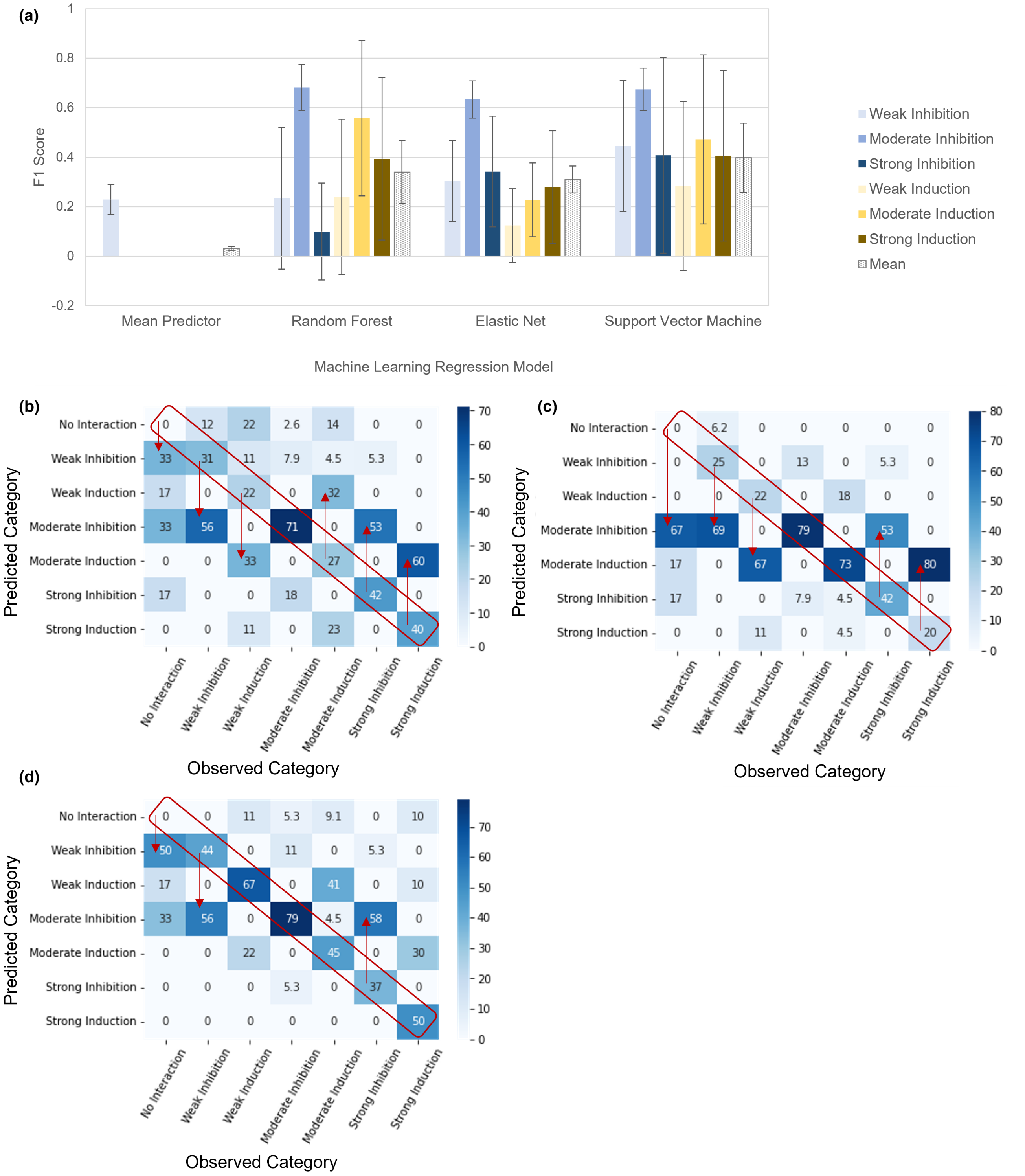


FIGURE 2 Regression performance of different machine-learning models on the drug–drug interaction (DDI) dataset. Fivefold cross-validation results (n = 5) are provided for different machine-learning models applied to the DDI dataset. Random forest (dotted), elastic net (grid), and support vector machine regressor (striped) were tested, and the mean predictor was a negative control for predictive capability, which always predicted the mean observed $\log_{10}$ area under the curve (AUC) ratio of the training fold. (a) "<Two-fold error score" reflects the proportion of predicted $\log_{10}$ AUC ratios that were within twofold of their corresponding observed $\log_{10}$ AUC ratios. (b) R^2 represents the coefficient of determination. (c) RMSE represents the root mean square error (lower is better). Data are represented as the mean metric of five test folds, and the gray lines topping each bar indicate the 95% confidence intervals.

between the observed and predicted $\log_{10}$ AUC ratios (R^2 =0.73; Figure 4a). The result of the Wilcoxon signed-rank test indicated that the distributions of predictions and observations were not statistically significantly different (p = 0.51). The model predicted $\log_{10}$ AUC ratios between −1.5 and 1 effectively, but ratios >1 were poorly predicted because the predictions were outside the twofold error margin. Figure 4b shows the plot of residuals against observed $\log_{10}$ AUC ratios. There is a bias in the prediction in that both induction and inhibition effects are underpredicted.

CYP450 and f_m data enabled the highest model performance

Because the SVR model overall performed the best (Figure 4), we assessed which features were most influential for driving the AUC ratio prediction. Subsets of the original features were grouped by the type of information the feature represented. The chemical features represented ECFP4 fingerprints and OESelma descriptors. The in vitro ADME features represented the non-CYP descriptors of ADME derived from early in vitro assays, and the other feature set contained the CYP and f_m descriptors. Each of the four feature sets was evaluated using fivefold cross-validation with the SVR model. Individual feature importance was then evaluated using SHAP methodology, measured by the average impact on model output. Figure 5a shows that the RMSE scores were not significantly different across the different models. However, twofold error scores were significantly lower for the chemical (0.63) and in vitro ADME (0.66) models compared with the full model (0.78). In addition, the chemical model showed a significantly reduced R^2 (0.60) compared with the full model (0.73). Only the CYP and f_m model showed no significant difference compared with the full model. Figure 5b shows the 20 features with the



(b)

	No Interaction	Weak Inhibition	Weak Induction	Moderate Inhibition	Moderate Induction	Strong Inhibition	Strong Induction
No Interaction	0	12	22	2.6	14	0	0
Weak Inhibition	33	31	11	7.9	4.5	5.3	0
Weak Induction	17	0	22	0	32	0	0
Moderate Inhibition	33	56	0	71	0	53	0
Moderate Induction	0	0	33	0	27	0	60
Strong Inhibition	17	0	0	18	0	42	0
Strong Induction	0	0	11	0	23	0	40

(c)

	No Interaction	Weak Inhibition	Weak Induction	Moderate Inhibition	Moderate Induction	Strong Inhibition	Strong Induction
No Interaction	0	6.2	0	0	0	0	0
Weak Inhibition	0	25	0	13	0	5.3	0
Weak Induction	0	0	22	0	18	0	0
Moderate Inhibition	67	69	0	79	0	53	0
Moderate Induction	17	0	67	0	73	0	80
Strong Inhibition	17	0	0	7.9	4.5	42	0
Strong Induction	0	0	11	0	4.5	0	20

(d)

	No Interaction	Weak Inhibition	Weak Induction	Moderate Inhibition	Moderate Induction	Strong Inhibition	Strong Induction
No Interaction	0	0	11	5.3	9.1	0	10
Weak Inhibition	50	44	0	11	0	5.3	0
Weak Induction	17	0	67	0	41	0	10
Moderate Inhibition	33	56	0	79	4.5	58	0
Moderate Induction	0	0	22	0	45	0	30
Strong Inhibition	0	0	0	5.3	0	37	0
Strong Induction	0	0	0	0	0	0	50

FIGURE 3 Classification performance of machine-learning regressor outputs after conversion to categorical labels. Predicted $\log_{10}$ area under the curve (AUC) ratios from different machine-learning models were used to classify interactions according to US Food and Drug Administration guidelines. The possible classes were weak, moderate, and strong inhibition (light, medium, or dark blue) or induction (light, medium, or dark orange). (a) The mean performance of the model across the different classes and folds was also compared (dotted). Data are presented as the mean macro F1 score (harmonic mean of precision and recall) across the five folds ($n = 5$ for each model), along with lines indicating the 95% confidence intervals. Classification matrices of (b) elastic net, (c) random forest regressor, and (d) support vector regressor outputs are shown. Numbers correspond to the percentage of interactions of a given observed class that were predicted by the model to be a certain class. The red diagonal box indicates the ideal cases where predicted classes matched the observed class, whereas the red arrows indicate cases where observed classes were predicted to be a different class more often than the correct class.

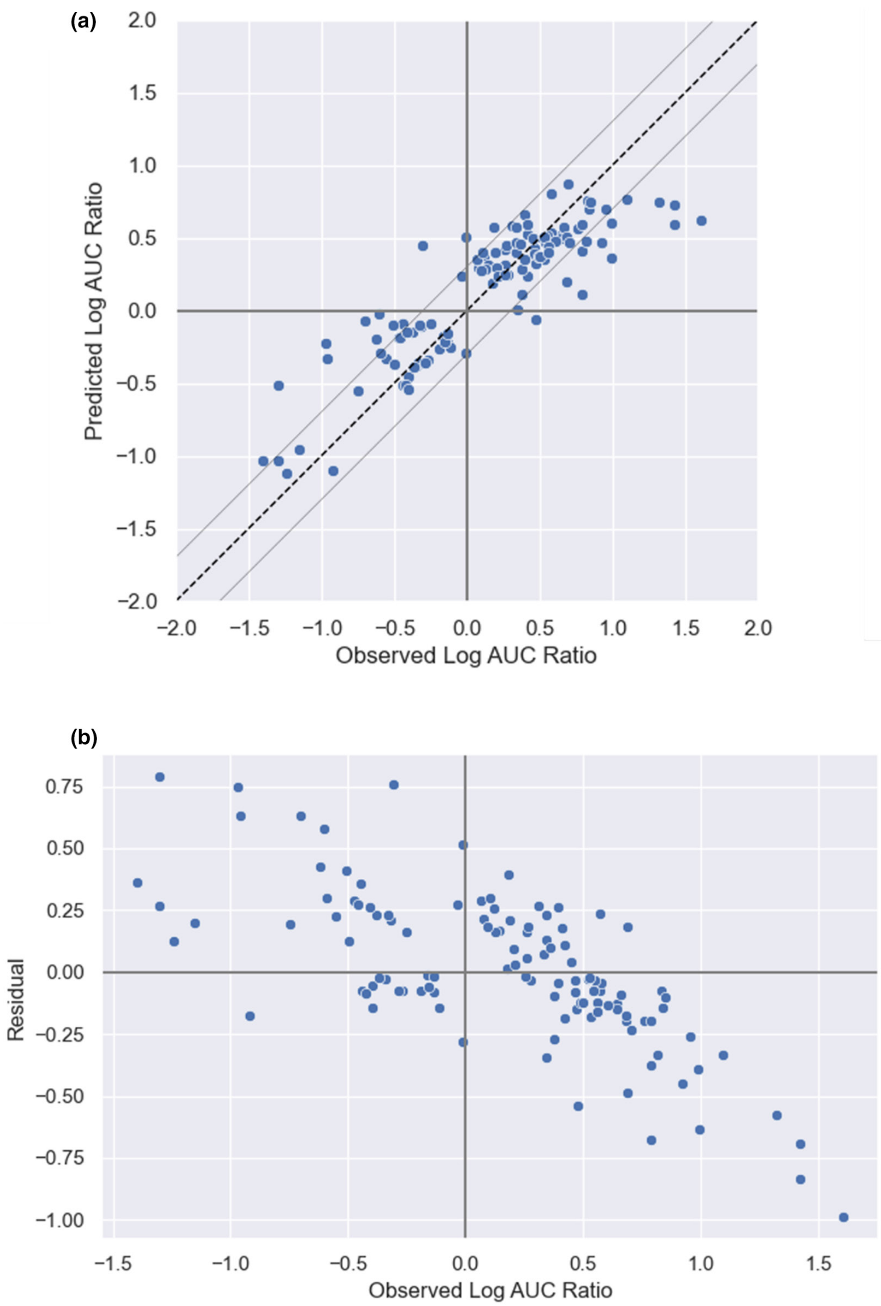


FIGURE 4 Regression analysis of observed and support vector regressor predicted $\log_{10}$ area under the curve (AUC) ratios. (a) Predicted $\log_{10}$ AUC ratios after fivefold cross-validation of a support vector regressor were plotted against the observed $\log_{10}$ AUC ratios of the corresponding drug–drug interaction studies. The dashed diagonal line indicates the perfect prediction line where predictions equal observations. The two gray lines on either side indicate the twofold error margins. Predictive performance was generally good (coefficient of determination =0.73, root mean square error = 0.30, proportion of predictions less than twofold error = 0.78). (b) A plot is shown of the resulting residuals against the observed $\log_{10}$ AUC ratios.

highest impact on model outputs using SHAP methodology. Notably, eight of these most important features were related to CYP and f_m data (these were CYP activity at different timepoints and CYP3A5 f_m with the perpetrator). In summary, the feature importance investigation demonstrated that CYP and f_m data were the most influential for the predictive performance of the model.

DISCUSSION

This study focused on predicting CYP-mediated DDIs as a first step to predicting PK DDIs because CYPs mediate a significant proportion of PK DDIs.[4] This analysis described whether changes in drug exposure (specifically changes in CYP and f_m data) attributed to PK DDIs can be predicted based on drug features. The findings of this study highlight that clinically significant exposure changes attributed to drug induction or inhibition can be predicted early in drug discovery using a regression-based ML model. The ML models demonstrated high R^2, low RMSE, and a high proportion of predictions within twofold of the observed AUC ratios for DDIs of $\log_{10}$ AUC ratio <1 (all but the strongest cases of inhibition). The SVR model performed marginally better than the other approaches. Furthermore, predictive performance was retained using just CYP and f_m data for PK DDIs mediated via time-dependent inhibition or induction. This was

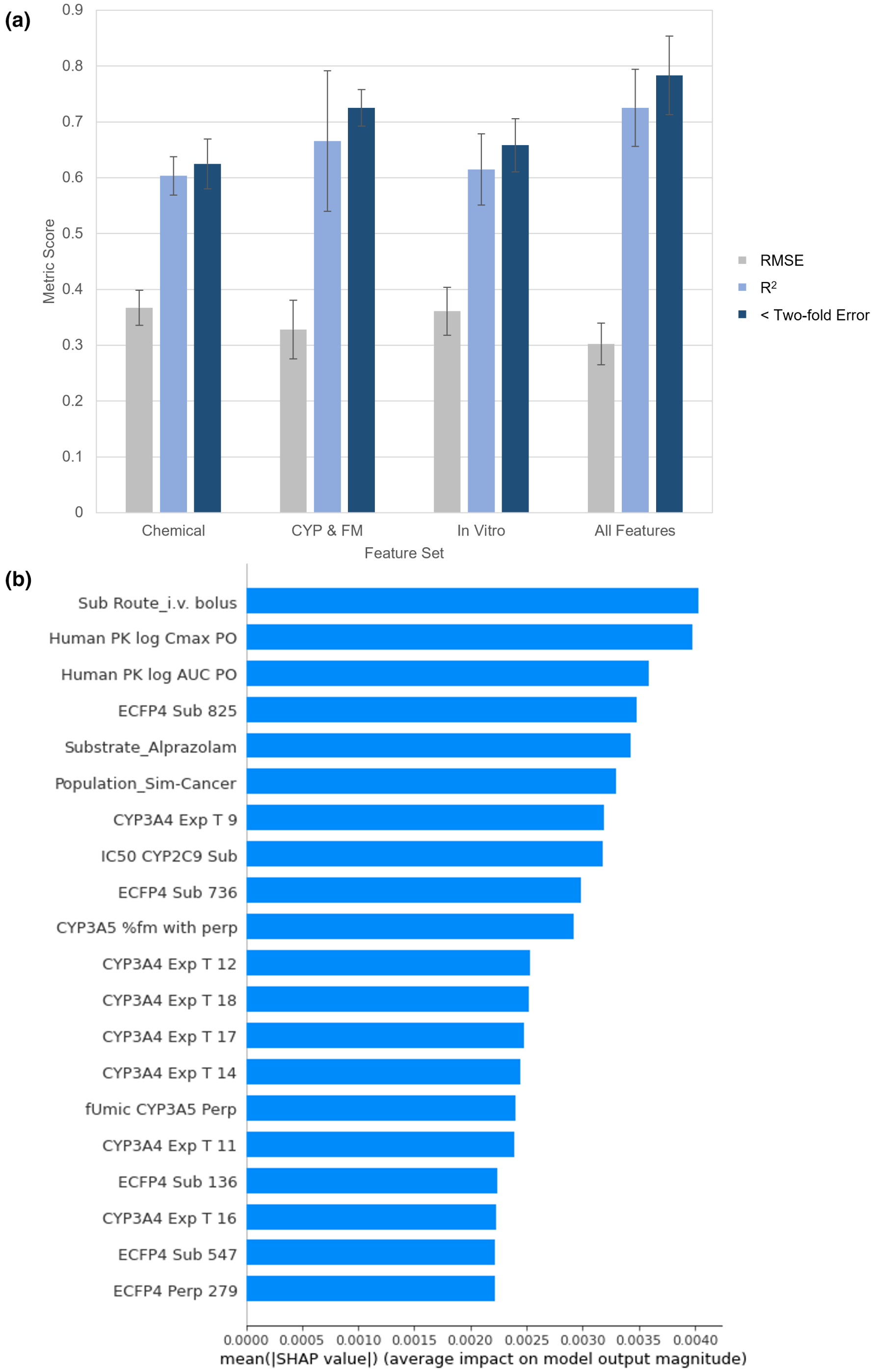


FIGURE 5 Regression performance of support vector regressor using different feature sets. (a) Support vector regressor model performance is shown using either chemical; in vitro absorption, distribution, metabolism, and excretion; or CYP and f_m/FM feature sets. Regression performance was compared using fivefold cross-validation on the drug–drug interaction dataset. The performance metrics used were the RMSE (gray), R^2 (coefficient of determination; light blue), and twofold error (proportion of predicted $\log_{10}$ AUC ratios within twofold of the observed value; dark blue). Metric scores were reported as the mean score of the five folds along with lines indicating 95% confidence intervals. (b) SHAP values of the 20 most important features in the support vector regressor model are shown. CYP and FM indicate the cytochrome P450 enzyme activity-time profiles and the fraction of substrate drug metabolized by each CYP enzyme with and without the perpetrator drug. AUC, area under the curve; $C_{\max}$, maximum concentration; CYP2C9, cytochrome P450 2C9; CYP3A4, cytochrome P450 3A4; CYP3A5, cytochrome P450 3A5; ECFP4, extended connectivity fingerprint with diameter 4; EM, extensive metabolizer; Exp, expected activity (simulation output); f_m, fraction metabolized; fUmic, fraction unbound in microsomes IC50, half maximal inhibitory concentration; I.V., intravenous administration; Perp, perpetrator; PK, pharmacokinetic; PM, poor metabolizer; PO, oral administration; Pop, population type; RMSE, root mean square error; SHAP, Shaply Additive Explanations; Sub, substrate, T, timepoint

demonstrated by the insignificant decrease when using CYP and f_m features only compared with the full feature set for the R^2 (0.06 decrease), RMSE (0.03 increase), and twofold error (0.05 decrease). Therefore, the development of a model using this refined feature set may be prioritized because it is simpler and has reduced data collection requirements without the significant loss of performance associated with the other feature sets. However, the use of chemical and in vitro feature set models may also have utility when available data are limited because RMSE scores did not differ significantly between any of the feature subset models.

Of the 20 most important features in the model, eight described CYP activity at timepoints soon after administration (generally between 24 and 72 h after the first dose of the perpetrator, although the exact times represented by each timepoint depend on the specific conditions used for the SimCYP simulation). This indicates that CYP activity may not need to be measured for the entire duration of administration of a perpetrator to predict the effect on AUC, providing the possibility to explore using more limited time ranges for prediction to reduce data collection requirements. Two of the features also described the predicted maximum concentration and AUC of the substrate—this demonstrates the utility of using outputs of other ML models in feature generation.

The model's underprediction of stronger cases of inhibition could be linked to the absence of other features describing mechanisms driving stronger inhibition. An important mechanism not considered in this model is transporter-mediated DDIs. For example, the hepatic uptake transporter Organic anion-transporting polypeptide 1B1 and 1B3 has been found to mediate more than half of the observed strong clinical DDIs highlighted by the FDA.[30] In addition, drugs can inhibit transporters expressed on the basolateral surface of proximal tubular epithelial cells.[31] Because these transporters enable the hepatic and renal clearance of other drugs, their inhibition can lead to an increased AUC of substrate drugs.[32] Similarly, the underprediction of induction could be explained by unmodeled transporter effects. Some perpetrator drugs included in this dataset, such as rifampicin and carbamazepine, induce the gene expression of P-glycoprotein (P-gp), which leads to a decreased AUC for P-gp substrates.[33,34] P-gp is particularly relevant because 75% of CYP3A4/5 substrates are also substrates of P-gp.[35] CYP3A4/5 metabolized the majority of interactions in the dataset used in this study and is the most common enzyme through which DDIs are mediated,[30,35] meaning P-gp is likely involved in many DDIs in this study. Regardless, the model predicted the majority (~80%) of $\log_{10}$ AUC ratios below 1 within twofold, indicating that the current model is capable of effective predictions for cases that are not strong inhibition.

Imbalance in the distribution of samples may have led to poorer performance in some data ranges. F1 scores for the inhibitory and/or moderate DDIs were higher than induction and strong/weak DDIs, and there was no prediction of noninteractions. There were more inhibition DDIs than induction, more moderate than strong/weak, and noninteractions were the least common. Therefore, F1 scores for each class appear to be dependent on the sample size for each class. Previous work has shown that regression and classification ML model predictions are biased toward predicting values that are in the ranges most represented in the training data.[29] This may have contributed to the bias toward moderate predictions and the consequent underprediction of strong inhibition and induction. Therefore, the use of techniques that correct for class imbalance biases, such as synthetic minority oversampling technique,[36] may lead to improved performance in these other data ranges for both classification and regression. The use of deep-learning algorithms could also enable improved regression performance using the set of features discussed in this study because deep learning improves DDI prediction.[9,37]

There were limitations to this study. Although the potential for overfitting (and thus reduced generalizability) caused by the small sample size to feature ratio is controlled by the use of the nested cross-validation,[38] the model needs to be tested on a dataset from a different source to confirm generalizability. Generalizability to clinical applications of DDI risk assessment may be limited by the feature space of the training DDI dataset. For example, ML models showed poor performance when making DDI predictions using chemical structures that the models were not trained on.[39]

Future work could explore the application of PK DDI modeling in other stages of drug development. This study demonstrated the predictive power based on just CYP and f_m features, and collecting these data for new drug candidates currently requires information from in vitro assays. Recent ML models can predict whether a drug will be metabolized by CYP3A4 using chemical features,[40] and deep learning has been used to classify which CYPs a drug may inhibit.[41] If this concept is developed to predict continuous values for f_m and CYP activity-time profiles, the necessary CYP and f_m inputs for this model could be collected from in silico models rather than in vitro assays. This would enable the model to be used by project teams at the point of design. Conversely, ML modeling could also be used for regression-based DDI modeling of already established drugs where clinical ADE data are already available because ADE data improve DDI prediction.[7] Also, the AUC label data used to train the model has uncertainty estimates from the SimCYP software used to generate it. The traditional ML architectures implemented in this study do not account for such uncertainties in the labels. However, future developments of this model could account for these by describing the label data as a probability distribution rather than as a single value. An example of this is the probabilistic random forest model, which accounts for uncertainty in feature or label data and outperformed the traditional random forest 30% of the time when used with noisy label data.[42]

In conclusion, this study has shown that regression-based ML can be used to predict changes in substrate AUC due to PK DDIs caused by time-dependent inhibition/induction using features available early in the drug-discovery process. Predictive algorithms as described here can inform early DDI risk for drug candidates.

AUTHOR CONTRIBUTIONS

J.G., M.M., R.A., and V.P.R. wrote the manuscript. A.M., B.W., F.M., J.G., M.M., and V.P.R. designed the research. J.G. and V.P.R. performed the research and analyzed the data.

FUNDING INFORMATION

No funding was received for this work.

CONFLICT OF INTEREST

Venkatesh Pilla Reddy, Marie Moullet, Anton Martinsson, and Filip Miljković are full-time employees of and hold shares of AstraZeneca. Beth Williamson and Rosalinda H. Arends were employees of AstraZeneca at the time this research work was carried out.

ORCID

Anton Martinsson https://orcid.org/0000-0003-3963-6105
Filip Miljković https://orcid.org/0000-0001-5365-505X
Venkatesh Pilla Reddy https://orcid.org/0000-0002-7786-4371

REFERENCES

1. Midão L, Giardini A, Menditto E, Kardas P, Costa E. Polypharmacy prevalence among older adults based on the survey of health, ageing and retirement in Europe. *Arch Gerontol Geriatr*. 2018;78:213-220.
2. Peng Y, Cheng Z, Xie F. Evaluation of pharmacokinetic drug-drug interactions: a review of the mechanisms, in vitro and in silico approaches. *Metabolites*. 2021;11(2):1-16.
3. Manikandan P, Nagini S. Cytochrome P450 structure, function and clinical significance: a review. *Curr Drug Targets*. 2018;19(1):38-54.
4. Center for Drug Evaluation and Research. *In Vitro Drug Interaction Studies — Cytochrome P450 Enzyme- and Transporter-Mediated Drug Interactions*. Federal Information & News Dispatch, LLC. 2020. Accessed February 12, 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/vitro-drug-interaction-studies-cytochrome-p450-enzyme-and-transporter-mediated-drug-interactions
5. Fitzmaurice MG, Wong A, Akerberg H, et al. Evaluation of potential drug-drug interactions in adults in the intensive care unit: a systematic review and meta-analysis. *Drug Saf*. 2019;42(9):1035-1044.
6. Jamei M, Marciniak S, Feng K, Barnett A, Tucker G, Rostami-Hodjegan A. The Simcyp population-based ADME simulator. *Expert Opin Drug Metab Toxicol*. 2009;5(2):211-223.
7. Cheng F, Zhao Z. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc*. 2014;21(e2):e278-e286.
8. Takeda T, Hao M, Cheng T, Bryant SH, Wang Y. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *J Chem*. 2017;9(1):16.
9. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc Natl Acad Sci USA*. 2018;115(18):E4304-E4311.
10. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. In: Maria Florina B, Kilian QW, eds. *Proceedings of the 33rd International Conference on Machine Learning*. Proceedings of Machine Learning Research: PMLR; 2016:2071-2080.
11. Karim MR, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Niagara Falls, NY: Association for Computing Machinery; 2019:113–123.
12. Yao J, Sun W, Jian Z, Wu Q, Wang X. Effective knowledge graph embeddings based on multidirectional semantics relations for polypharmacy side effects prediction. *Bioinformatics*. 2022;38(8):2315-2322.
13. Chen Y, Ma T, Yang X, Wang J, Song B, Zeng X. MUFFIN: multi-scale feature fusion for drug–drug interaction prediction. *Bioinformatics*. 2021;37(17):2651-2658.
14. Center for Drug Evaluation and Research. *Clinical Drug Interaction Studies — Cytochrome P450 Enzyme- and Transporter-Mediated Drug Interactions Guidance for Industry*. Federal Information & News Dispatch, LLC. 2020. Accessed February 12, 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-drug-interaction-studies-cytochrome-p450-enzyme-and-transporter-mediated-drug-interactions.
15. Hachad H, Ragueneau-Majlessi I, Levy RH. A useful tool for drug interaction evaluation: the University of Washington Metabolism and transport drug interaction database. *Hum Genomics*. 2010;5(1):61-72.
16. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28(1):31-36.
17. Olsson T, Sherbukhin V. *SELMA, Synthesis and Structure Administration (SaSA)*. AstraZeneca R&D Mölndal; 2002.
18. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742-754.

19. Miljković F, Martinsson A, Obrezanova O, et al. Machine learning models for human in vivo pharmacokinetic parameters with in-house validation. *Mol Pharm*. 2021;18(12):4520-4530.
20. Oprisiu I, Winiwarter S. In silico ADME modeling. *Systems Medicine: Integrative, Qualitative and Computational Approaches*. Vol 2; Academic Press; 2020:208-222.
21. *ClogP*. (Version 4.3). 2022; Pomona College and BioByte Inc.
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *CoRR*. 2012;12:2825-2830.
23. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc, B: Stat Methodol*. 2005;67(2):301-320.
25. Drucker H, Burges C, Kaufman L, Smola A, Vapnik V. Support vector regression machines. *Adv Neural Inform Process Syst*. 1997;28:779-784.
26. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*. MIT Press; 2017.
27. Jones H, Chen Y, Gibson C, et al. Physiologically based pharmacokinetic modeling in drug discovery and development: a pharmaceutical industry perspective. *Clin Pharmacol Ther*. 2015;97(3):247-262.
28. Buckland M, Gey F. The relationship between recall and precision. *J Am Soc Inform Sci*. 1994;45(1):12-19.
29. Mac Namee B, Cunningham P, Byrne S, Corrigan OI. The problem of bias in training data in regression problems in medical decision support. *Artif Intell Med*. 2002;24(1):51-70.
30. Yu J, Petrie ID, Levy RH, Ragueneau-Majlessi I. Mechanisms and clinical significance of pharmacokinetic-based drug-drug interactions with drugs approved by the U.S. Food and Drug Administration in 2017. *Drug Metab Dispos*. 2019;47(2):135-144.
31. Ivanyuk A, Livio F, Biollaz J, Buclin T. Renal drug transporters and drug interactions. *Clin Pharmacokinet*. 2017;56(8):825-892.
32. Kusuhara H, Ito S, Kumagai Y, et al. Effects of a MATE protein inhibitor, pyrimethamine, on the renal elimination of metformin at oral microdose and at therapeutic dose in healthy subjects. *Clin Pharmacol Ther*. 2011;89(6):837-844.
33. Yamada S, Yasui-Furukori N, Akamine Y, Kaneko S, Uno T. Effects of the P-glycoprotein inducer carbamazepine on fexofenadine pharmacokinetics. *Ther Drug Monit*. 2009;31(6):764-768.
34. Greiner B, Eichelbaum M, Fritz P, et al. The role of intestinal P-glycoprotein in the interaction of digoxin and rifampin. *J Clin Invest*. 1999;104(2):147-153.
35. Yu J, Zhou Z, Tay-Sontheimer J, Levy RH, Ragueneau-Majlessi I. Risk of clinically relevant pharmacokinetic-based drug-drug interactions with drugs approved by the U.S. Food and Drug Administration between 2013 and 2016. *Drug Metab Dispos*. 2018;46(6):835-845.
36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
37. Jin B, Yang H, Xiao C, Zhang P, Wei X, Wang F. Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction. *Proc AAAI Conf Artif Intell*. 2017;31(1):1367-1373.
38. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365.
39. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*. 2018;34(9):1538-1546.
40. Hu B, Zhou X, Mohutsky MA, Desai PV. Structure–property relationships and machine learning models for addressing CYP3A4-mediated victim drug–drug interaction risk in drug discovery. *Mol Pharm*. 2020;17(9):3600-3608.
41. Li X, Xu Y, Lai L, Pei J. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Mol Pharm*. 2018;15(10):4336-4345.
42. Reis I, Baron D, Shahaf S. Probabilistic Random Forest: a machine learning algorithm for Noisy data sets. *Astron J*. 2018;157(1):16.

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: Gill J, Moullet M, Martinsson A, et al. Evaluating the performance of machine-learning regression models for pharmacokinetic drug–drug interactions. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:122-134. doi: 10.1002/psp4.12884